Mining the red sea metagenomics libraries for betaine pathways

Sherouk Allam

Follow this and additional works at: https://fount.aucegypt.edu/etds

Recommended Citation

APA Citation

MLA Citation

This Master's Thesis is brought to you for free and open access by the Student Research at AUC Knowledge Fountain. It has been accepted for inclusion in Theses and Dissertations by an authorized administrator of AUC Knowledge Fountain. For more information, please contact thesisadmin@aucegypt.edu.
School of Sciences and Engineering

Mining the Red Sea Metagenomics Libraries for Betaine Pathways

A Thesis Submitted to

The Biotechnology Master’s Program

In partial fulfillment of the requirements for

the degree of Master of Science

By: Sherouk Abou Allam

Under the supervision of:

Dr. Walid Fouad

2018
The American University in Cairo

Mining the Red Sea Metagenomics Libraries for Betaine Pathways

A Thesis Submitted by

Sherouk Abou Allam

To the Biotechnology Graduate Program

2018

In partial fulfillment of the requirements for

The degree of Master of Science

Has been approved by

Thesis Committee Supervisor/Chair ____________________________________________

Affiliation ____________________________________________

Thesis Committee Reader/Examiner ____________________________________________

Affiliation ____________________________________________

Thesis Committee Reader/Examiner ____________________________________________

Affiliation ____________________________________________

Thesis Committee Reader/External Examiner _____________________________________

Affiliation ____________________________________________

________________                __________________                 __________________   ________________
Dept. Chair/Director              Date                        Dean                             Date
DEDICATION

To my beloved parents who taught me to love science and never lose hope.

To Mokhtar, my husband, who has been a constant source of support and encouragement during the challenges of graduation.

To Arwa and Salman, my children, who brought smiles to my life and gave me the reason to live.

To Nissma, my sister, the best friend and best guidance who taught me faith.
ACKNOWLEDGEMENTS

This work could have never been done without the help, support and guidance of Dr. Walid. I sincerely thank him for everything.

Special thanks to Dr. Rania Siam, without the supportive words of her, I would not have continued my studies.

Special thanks to Dr. Ahmed Moustafa, who taught me lots techniques of bioinformatics.

I am very grateful to Mustafa Adel and Nahla Hussein for their help through the years of my graduate study.
Glycine Betaine (GB) is a fundamental solute for salinity tolerance in many organisms. Our aim in this study was to identify the GB pathway in highly complex natural community inhabiting brine pool and Red Sea column in Atlantis II Red Sea. Environmental genomics is a powerful tool for collecting information on microbial communities. Knowledge can be collected about highly complex natural communities. Microbial community living in extreme conditions evolved diverse molecular mechanisms to withstand these conditions. Metagenomic libraries from Atlantic II samples were used to study betaine biosynthetic pathway that is more abundant. Betaine is synthesized independently from two different substrates, Choline and Glycine, through two oxidation and three methylation enzymatic steps, respectively. A comparative sequence based analysis was made for the enzymes of the two branches in GB pathway. Assembled sequences were blasted, and comparisons were done on number of occurrence. Some more procedures were accomplished, to ensure that data encountered are valid, and to eliminate false hits. Conserved domain was checked in blast results. The existence of the active sites indicates enzymes activity. Species were mapped to ensure that the enzymes in each pathway existed in the same species.

After conducting analysis, it was found that Choline Pathway enzymes were encountered in abundance in targeted sites. The conserved domains of the enzymes, CDH and BADH, were found in vast amounts as well. In a comparative analysis between Choline pathway and Glycine pathway, data was normalized to show number of hits with consideration to sample size. It is found that in Salinity 4% in depth 200m, Choline Pathway recorded 0.0115, and Glycine pathway had 0.029. Whereas, in Salinity 4% in depth
700m, Choline Pathway had 0.53 hits, and Glycine Pathway had a record of 0.02. Salinity 8% in depth 1500m, Choline Pathway had 1.6, and Glycine Pathway had 0.46. In Salinity 9-16% in UCL, Choline Pathway was 1.2 and Glycine Pathway had 0.11 records. In Salinity 26% in LCL, Choline Pathway was recorded into 1.02, where as, Glycine Pathway had 0.13 records. The records indicates that adaptation encountered by Choline pathway tends to be preferable pathway for the targeted community. Glycine pathway, on the other hand, was found in less abundance. The environment in most depths showed less bias to Glycine PW, but had consistent increase with the increase of salinity, the more salinity the more the enzymes were available in full length. Glycine pathway enzymes and their active domains that appeared through depths in accordance with salinity could be more conserved and that is the reason for their conservative abundance.
# Table of Content

DEDICATION ................................................. III

ACKNOWLEDGEMENTS ...................................... IV

ABSTRACT .................................................... V

The American University in Cairo ........................................ V

Table of Content ................................................ VII

Table of Figures .............................................. Error! Bookmark not defined.

List of Tables ................................................ X

LIST OF ABBREVIATIONS ........................................ XI

CHAPTER 1: REVIEW OF LITERATURE AND STUDY OBJECTIVES 1

Introduction ................................................................. 1

1.1 Stressful environment definition & Stressful environment geographical sites ....................... 2

1.2 Effect of high salinity & Adaptation .............................................................................. 2

1.3 Glycine Betaine Pathway .............................................................................................. 5

1.4 Further studies that focused on natural adaptation to high salinity in different organism ........ 7

1.5 Red Sea Atlantis II Brine Pool ....................................................................................... 10

1.5.1 Studies Conducted in Atlantis II Brine Pool in the Red Sea ...................................... 11

1.6 Bioinformatics ............................................................................................................. 11

1.6.1 Assembly ................................................................................................................. 11

1.6.2 Blast ........................................................................................................................ 13

1.6.3 Conserved domains ................................................................................................. 13

1.7 Metagenomic approaches in the search of novel genes: ................................................. 14

1.8 Study objectives and experimental design ...................................................................... 14

CHAPTER 2: MATERIALS AND METHODS ........................................... 22
2.1 Data Processing Conducted: .................................................................22
  2.1.1 Reads Assembly 22
  2.1.2 Sequence Annotation by BlastX 23
  2.1.3 Details Collection for the hits 23

2.2 Data Analysis ..................................................................................24

CHAPTER 3: RESULTS AND DISCUSSION ............................................28
  3.1 Screening of Glycine Betaine biosynthetic pathways 28
     3.1.1 Analysis of the Assembly results 28
     3.1.2 Glycine Betaine pathway enzymes identification 29
     3.1.3 Distribution of Glycine Betaine biosynthetic pathways 29
     3.1.4 Domain analysis for identified enzymes 33
     3.1.5 Taxonomical Screening and Organism Counts 34
     3.1.6 Glycine Betaine Pathway Data Normalization 35
  3.2 Prevailed Species 36

CHAPTER 4: CONCLUSIONS AND RECOMMENDATIONS ..................47
  4.1 Conclusion ....................................................................................47
  4.2 Recommendation for Further Research ........................................47
Table of Figures

Figure 1: Sites of brine pools in Red Sea. It is clear that Atlantis II is the largest one. (Zierenberg 1988) 17

Figure 2: Atlantis II Deep region of the Red Sea 18

Figure 3: The figure illustrates the comparison between Plastx, Blastx, and Diamonds in time consumed. Blastx takes longer performance time. 19

Figure 4: Blastx vs other programs: n performance Blastx still is the highest accuracy 20

Figure 5: Glycine Betaine Biosynthesis Pathways 22

Figure 8: Work Flow: Steps of the process made, assembled the sequences, blasting the contigs against subgroup of nr, Collecting Detailed Description for the hits from NCBI, uniprot, and Kegg, to describe each reference ontology, and species blast, taxonomy 28

Figure 7: Data Analysis workflow. Data analysis done to show the preferred pathway for organism lived in the selected depths 29

Figure 8: Choline Pathway abundance. BADH shows the most abundance in AT1500, UCL, LCL and less abundance in AT700 and AT200 44

Figure 9: Glycine Pathway abundance. Here the figure shows all five sites together and how each enzyme of the Glycine pathway is distributed 45

Figure 10: Each circle resembles an enzyme. The area represents the number of hits produced. The scale was plotted proportionately to compare the numerical quantity of each Enzyme. In LCL, UCL, 1500m, 700m, 200m comparing the six enzymes of the two branches of Glycine Betaine Pathway 46

Figure 11: Shows the intersection and disjointed between the two proteins in Choline pathways in species' occurrence. The intersection area shows that both enzymes exist in the same species 47

Figure 12: Comparison done in abundance between the two pathways Choline and Glycine. Glycine pathway shows consistent increase with the increase of salinity 48

Figure 13: Taxonomy of Betaine Aldehyde Dehydrogenase in LCL 49
List of Tables

TABLE 1: PHYSICAL CHARACTERISTICS OF THE ATLANTIS II BRINE POOL, THICKNESS OF LAYER, SALINITY, TEMPERATURE IN EACH LAYER, AND PERCENTAGE OF OXYGEN IN WATER. LCL HAS THE HIGHEST DEGREE IN TEMPERATURE AND SALINITY, BUT OXYGEN IS ALMOST 0.  

TABLE 2: TABULATING CONTIGS IN A USER FRIENDLY EXCEL SHEET.  

TABLE 3: TOTAL READS AND BASES. THIS TABLE COMBINES THE FILES OF THE SAMPLE, SHOWING THE SIZE OF EACH FILE IN NUMBER OF READS AND NUMBER OF BASES. AT1500 SHOWS THE LARGEST NUMBER OF READS. AT0200 AND 0700M SHOW A SMALLER SAMPLE, WHICH INDICATES SMALLER COMMUNITY.  

TABLE 4: SUMMERY OF ASSEMBLY. MAIN PROPERTIES OF SAMPLE ASSEMBLY ARE ILLUSTRATED; CONSENSUS RESULTS ARE DISPLAYED IN THIS INTEGRATED TABLE AGAINST AT II 200, 700, 1500, AND LCL AND UCL.  

TABLE 5: QUANTITATIVE HITS FOR GLYCINE PATHWAY. HIT ARE DISPLAYED, DEPTH LCL AND UCL HAS THE LARGEST QUANTITIES.  

TABLE 6: HITS FOR CHOLINE PATHWAY. BETAIENADH HAS THE LARGEST HITS IN THE LCL, UCL, AND 1500M  

TABLE 7: DISTRIBUTION OF GLYCINE BETAINE PATHWAY QUANTITATIVE  

TABLE 8: NORMALIZATION INTEGRATED DATA FOR GLYCINE BETAINE PATHWAYS IN 5 DEPTHS  

TABLE 9: CONSERVED DOMAINS IN PROTEINS OF CHOLINE PATHWAYS, (1500N, 700M, UCL, LCL)  

TABLE 10: CONSERVED DOMAINS IN PROTEINS INVOLVED IN GLYCINE PATHWAY OF DEPTH 1500M IN ROW, 700M IN ROW 2, UCL IN ROW 3.  

X
LIST OF ABBREVIATIONS

ABL: class A beta-lactamase
ATII: Atlantis II Brine Pool
BADH: Betaine Aldehyde Dehydrogenase
BLAST: Basic Local Alignment Search Tool
DBG: de Bruijn Graph
DHALS: deep sea hypersaline anoxic lakes
DNA: Deoxyribonucleic Acid
GB: Glycine Betaine
GbsA: Glycine Betaine-aldehyde dehydrogenase, glycine betaine synthesis
GbsB: Glycine Betaine biosynthetic enzymes
GbsR: Regulatory Protein from
KAUST: King Abdullah University of Science and Technology
LCL: Lower Convective Layer
LT: Lake Tyrrell
NaCl: Sodium chloride
NAD: Nicotinamide Adenine Dinucleotide
NCBI: National Center for Biotechnology Information
NGS: Next Generation Sequencing
OLC: Overlap/Layout/Consensus
OpuB ABC-type transporters
OpuC: ABC-type transporters
ORF: Open Reading Frame
PCR: Polymerase Chain Reaction
PGDH: 3-phosphoglycerate dehydrogenase
QTP: Qinghai-Tibetan Plateau
TetR: Tet Repressor Protein
UCL: Upper Convective Layer
WGS: Whole genome shotgun
CHAPTER 1: REVIEW OF LITERATURE AND STUDY OBJECTIVES

Introduction

The Red Sea is characterized by the presence of deep-sea, hypersaline anoxic basins called brine pools, which are large bodies of water at the bottom of the ocean that are characterized by high temperature and salinity (Qian, et al., 2010). The Red Sea Atlantis II brine pool is a harsh environment that addresses multiple extreme conditions such as, high salinity, high temperature and high concentrations of multiple toxic heavy metals (Mohamed et al, 2013).

Living organisms adapted to saline environment evolved a mechanism of accumulating various molecules in their cell in order to regulate their cellular osmotic pressure and retain water inside their cells (Flowers and Colmer 2008; Salma et al., 2015). There are varieties of molecules that serve this function such as nonorganic molecules (Na+, K+, Ca++) and organic molecules (alcohol sugars, amino acids and quaternary amino compounds). Glycine Betaine (Betaine) is among the organic molecules that accumulate in the cytoplasm for regulating the cellular osmotic pressure. Betaine was found to be widely distributed across many saline adapted organisms (Bedford et al 1998; Hanson et al., 1994). There are two known pathways for betaine biosynthesis; the first is from Glycine through three sequential methylation of its amino group. The second pathway involves two steps, oxidation of choline followed by reduction of betaine aldehyde. It has been proposed that the choline branch of betaine biosynthesis is not preferred under anoxia conditions due to the required oxygen in the first reaction (Hanson et al., 1994). All enzymes catalyzing the reactions in both branches had been cloned and characterized form various species (Rathinasabapathi et al., 1997, Takata et al 2003; Jo et al 2008; Lai and Lai 2011). However, the presence of the two branches of betaine biosynthesis has not been reported in the unique microbial community of the red sea.

Therefore, the main objective of this research is to understand the distribution of betaine biosynthetic pathways in the microbial community of the red sea. We propose to analyze the metagenomics libraries available from various micro environments in the Red Sea for the presence of all enzymes involved in the two branches of betaine biosynthesis.
1.1 Stressful environment definition & Stressful environment geographical sites

Abiotic factors vary from one area to another, from one environment to another. Some areas are very harsh for living organisms. Some have mild conditions. Some are suitable for diversity of living creatures. Some need adaptations to survive. Even the same area can have variety of conditions depending on the site. Harsh factors could exist all over the year, or in specific conditions or seasons. An extreme environment is having conditions that make survival hard for most known life forms. These conditions may be high or low content of oxygen, or carbon dioxide in the atmosphere; high levels of radiation, acidity, or alkalinity, or extremely high or low temperature or pressure, or presence of sulphur, petroleum, and other toxic substances, or absence of water, or water containing a high concentration of salt or sugar. Also, the accumulation of metals in sediments.

Examples of extreme environments exist in many places on Earth. It is in very dry deserts, the poles, also in volcanoes, upper atmosphere, Mountain Everest, and Oceanic trenches, which is topography that has sort of depression of the sea floor. It is narrow in width, but very deep. Every area of mentioned here, has in itself different sites that could differ in the stress from harsh to very harsh. For example, in brine pools, depth 500m for example, is less stressful than the sea floor and sediments, such as Atlantis II deep in Red sea that has metal accumulation, Fe-Mn-oxides, Sulfide, Sulfates, Silicate and Heat of magma increases the temperature of the water (Zierenberg 1988).

1.2 Effect of high salinity & Adaptation

High salinity is one of the stresses that affect the cell and can cause death. The concept of osmosis can show the reason for this. Osmosis mainly means the water movement across a membrane. Osmosis is triggered by solute, such as salt. It attracts water, across the membrane. When you add water to a solute, it diffuses, decreasing the concentration of salt, forming a solution. It is called an isotonic solution, when inside concentration of salt is the same as the
salt concentration outside the cell, the level of water will stay the same, and cell will be in equilibrium.

Water moves to the highest concentration of salt. If the cell is in a place of high salinity outside its membrane, the water will leave the cell and get attracted to the high salt concentration. This movement of water outside the cell membrane causes. A person may die from dehydration if he drinks enough sea water. (Chester 2012)

One of the most harsh environments to organisms are Hypersaline environments. Organisms living in such conditions must develop many molecular mechanisms to adapt the conditions. A study was made at moderate-salinity rhizosphere and brine from the Es Trenc saltern (Mallorca, Spain), and an interesting discovery of E coli strain (MKH13) that has salt resistance. Eleven genes that addressed salt adaptation were identified, some encoding for previously known proteins related to osmoadaptation (Mirete 2015).

Some environments have extreme conditions. Osmatic pressure in extreme cases challenges the life of vast number of microorganisms. Hypersaline habitats are diverse. Examples are sediments and brine pools in marine ecosystems, salt lakes, and salt ponds. Salinity gradient can varies from low salinity to high level of saturation (Oren, 2002). These habitats, salt-enriched, can reach to amore than 30% (w/v) total salts, and native. Researches are concerned with the study of the molecular mechanisms to adapt the elevation of NaCl. The native microbial that inhabit these hypersaline environments can grow in the presence of (Rodriguez-Valera et al., 1985; Antón et al., 2000). Even” halophilic Archaea such as the members of the family Halobacteriaceae, representatives of Bacteria and Eukarya can also thrive under these harsh conditions” (Oren, 2008).

In the studies generally halophiles adjust to high salinity by retaining mainly one of two methods to sustain the osmotic stability between the cytoplasm and the adjacent medium: “the “salt-in-cytoplasm” strategy and the compatible solute strategy” (Galinski, 1995; Sleator and Hill, 2001; Oren, 2008). The ‘salt-in’ strategy has main feature, that the organism increase the salt inside its body. That causes modifications in the enzymatic process. Modifications include the high proportion of high acidic amino acids such as aspartate (Asp), and a low proportion of hydrophobic residues, in order to form coil regions “instead of helical structures when compared to non-halophile proteins” (Paul et al., 2008; Rhodes et al., 2010). This strategy is used by the bacterium Salinibacter ruber and also extremely halophilic Archaea such as
**Halobacterium** sp. Their proteins are very acidic (Oren, 2008). The solute strategy and its compatible strategies are more spread than the “salt-in” strategy phylogenetically. The use of osmoprotectants and similar solutes do not interfere with the metabolism of the cell. In microorganisms using this strategy, high osmolarity conditions can cause accumulation of K+ ions in the cytoplasm that can eventually leads to salt tolerance as they can serve as intracellular osmoprotectants (Csonka, 1989; Sleator and Hill, 2001). “In a secondary response, compatible solutes can act as organic osmoprotectants that are biosynthesized and/or accumulated inside the cell to restore the cell volume and turgor pressure lost during the osmotic stress” (Csonka, 1989; Sleator and Hill, 2001). There is a vast diversity of organic solutes that can act as osmoprotectants, including glycine betaine and glycerol (Oren, 2008).

The great majority of the mechanisms of elevated salt resistance and osmoprotection are derived from the knowledge of cultivated microorganisms and their sequenced genomes, thus this information may be biased and may overlook specific strategies of adaptation (Wu et al., 2009). In fact, previous studies using metagenomic sequencing approaches in well-characterized hypersaline environments have revealed novel lineages and genomes from diverse microorganisms without previously cultured representatives (Narasingarao et al., 2012; López-López et al., 2013). Moreover, recent genomic studies on the genus **Halorhodospira** have revealed a combined use of both strategies of salt adaptation (Deole et al., 2013) and through metagenomic analysis an acid-shifted proteome has been described in a hypersaline mat from Guerrero Negro (Kunin et al., 2008). On the basis of these findings, the notion of a correlation between phylogenetic affiliation and the strategy of osmotic adaptation should be revised (Oren, 2013).

Living organisms adapted to saline environment evolved a mechanism of accumulating various molecules in their cell in order to regulate their cellular osmotic pressure and retain water inside their cells (Flowers and Colmer 2008; Salma et al., 2015). There are varieties of molecules that serve this function such as nonorganic molecules (Na+, K+, Ca++) and organic molecules (alcohol sugars, amino acids and quaternary amino compounds). Glycine Betaine (GB) is among the organic molecules that accumulate in the cytoplasm for regulating the cellular osmotic pressure. Betaine was found to be widely distributed across many saline adapted organisms (Bedford et al 1998; Hanson et al., 1994). There are two known pathways for betaine biosynthesis; the first is from Glycine through three sequential methylation of its amino group.
The second pathway involves two steps, oxidation of choline followed by reduction of betaine aldehyde. It has been proposed that the choline branch of betaine biosynthesis is not preferred under anoxia conditions due to the required oxygen in the first reaction (Hanson et al., 1994). All enzymes catalyzing the reactions in both branches had been cloned and characterized from various species (Rathinasabapathi et al., 1997, Takata et al 2003; Jo et al 2008; Lai and Lai 2011).

Any organisms living in these conditions are often very well adapted to their living circumstances, which is usually a result of long-term evolution. Physiologists well declared that organisms living in extreme environments are exposed to high evolutionary adaptation out of intense selection. Adaptation has many forms. The one we are discussing in this study, is genetic adaptation that alter enzymes pathway to serve the molecular need of the cell. (Garland 1994)

Environmental stresses cause many crops to wither. Stresses such as drought and salinity are the most destructive. Biotechnology is becoming the tool to solve this problem by increasing the organism ability in osmotic adjustment. Increasing number of researches were dedicated to study adaptive mechanism of stresses in organisms. Conclusive evidence of many studies that from a series of in vivo and in vitro studies of the physiology, biochemistry, genetics, and molecular biology of plants suggest strongly that Glycine Betaine (GB) performs an important function in plants subjected to environmental stresses. Biotechnology developed a strategy for exogenic application and transgenic of organism, in order to simulate other organism that accumulates Gglycine Betaine. Many successful studies genetically engineered biosynthesis of GB are done to help abiotic stress tolerance (Wani 2013).

1.3 Glycine Betaine Pathway

Glycine Betaine (GlyBet) is a compound that is quaternary ammonium. It works as osmoprotectant. Glycine Betaine synthesis works through two-step of choline oxidation, first step is catalyzed by Choline Oxidase (ferredoxine-dependent) and second step is catalyzed by Betaine Aldehyde dehydrogenase (NAD +/- dependent). Choline → Betaine Aldehyde → Glycine Betaine (Sakamoto 2002).
In extreme salinity, bacteria operate its survival mechanism. Glycine Betaine accumulates in large quantities in the cytoplasm without interfering with the regular cell functions. This accumulation helps the cell to maintain osmotic balance of cytoplasm and prevent dehydration. In extreme salinity, the cell needs external supply of Glycine Betaine or its precursors Choline and Betaine aldehyde (Lamark et al. 1991). In E. coli, two dehydrogenase is expressed which oxidise choline to Betaine aldehyde and Betaine aldehyde to glycine Betaine at the same rate. Also, E. coli has NAD-dependent Betaine aldehyde dehydrogenase soluble, which is very specific for Betaine aldehyde, in the choline pathway shown in Figure 6. The bet genes encoding all of these enzymes are triggered by osmotic stress. Glycine Betaine and its precursors choline and glycine Betaine aldehyde are main players in balancing high level of osmotic tolerance in high salinity environment and when added to Escherichia coli culture in high osmotic media. Choline shows an osmoprotectant in aerobic conditions. Glycine betaine aldehyde and glycine Betaine work in both aerobic and anaerobic conditions. In a study, cells grown in osmotically stress with no osmoprotectant or the precursors did not accumulate glycine betaine. O2-dependent, a transmembrane protein, and “electron transfer-linked dehydrogenase was found which oxidized choline to glycine Betaine aldehyde and aldehyde to glycine Betaine” at almost same rate. NAD-dependent dehydrogenase oxidized glycine Betaine aldehyde. In the same study it was addressed that choline-glycine Betaine pathway showed regulation with full enzymatic activates in aerobically osmatically stressed environment, in choline-containing medium. (Landfald 1986).

A study on *Bacillus subtilis*, investigated the choline oxidation pathway in high salt. Two enzymes of the choline oxidation pathway (GbsA/GbsB) are highly expressed in extreme salinity environment in *B. subtilis*. (Nau-Wagner et al., 2012). Here, the expression of TetR (BetI in *A. baylyi*) is controlled negatively by of the MarA family, labeled GbsR. The *gbsR* gene is not co-transcribed with the operon genes. GbsR was found to respond to choline levels. Moreover, GbsR is not associated in salt-dependent induction (Nau-Wagner et al., 2012). In addition to the choline oxidation pathway, GbsR controls the expression of two ABC transporters OpuC and OpuB facilitating choline uptake (Nau-Wagner et al., 2012).
1.4 Further studies that focused on natural adaptation to high salinity in different organism

Several studies focused on life in salt environment, adaptation and diversity of organisms. One study is about lake surface sediments of the Qinghai-Tibetan Plateau (QTP). In this study, microbial abundance and community structure in the surface sediments of nine lakes on the QTP were examined using the Illumina Miseq sequencing technique. The sample data produced were statistically analyzed put into consideration environmental variables. The results showed that lake salinity is the major factor for microbial structure and diversity and more important to life than geographic distance. After assembly and analysis, the study concluded that the abundant and rare taxa were significantly (P < 0.05) correlated (r = 0.427 and 0.783, respectively) with salinity. The study has suggested that rare taxa might be more sensitive to salinity than their abundant ones. (Yang 2016).

Naturally, environmental microbial community could be grouped into abundant and rare taxa with respect to their contribution to biomass and biodiversity. Contribution to major biomass and minor diversity depend on abundance of taxa, whereas rare taxa contribute minor biomass but major biodiversity to the ecosystem (Yang 2016).

Accumulation of many osmoprotectants in plants, improves their tolerance to abiotic harsh conditions. Betaine is one of the osmoprotectants that plays an important role in resisting stresses. Synthesiss of Betaine is by choline oxidation or glycine methylation. A study of osmoprotectants showed that exogenous supply of serine or glycine to a halotolerant cyanobacterium *Aphanotoche halophytica*, which synthesizes betaine from glycine by a three-step methylation, elevated intracellular accumulation of betaine under salt stress. The gene encoding 3-phosphoglycerate dehydrogenase (PGDH), which catalyzes the first step of the phosphorylated pathway of serine biosynthesis, was isolated from *A. halophytica*. Expression of the *Aphanotoche* PGDH gene in *Escherichia coli* caused an increase in levels of betaine as well as glycine and serine. Expression of the *Aphanotoche* PGDH gene in *Arabidopsis* plants, in which the betaine synthetic pathway was introduced via glycine methylation, further increased betaine levels and improved the stress tolerance. These results demonstrate that PGDH enhances the levels of betaine by providing the precursor serine for both choline oxidation and glycine methylation pathways.
In a current study, it was proofed that halotolerant cyanobacterium, *Aphanothece halophytica* (*A. halophysica*), synthesizes betaine from glycine by a three-step methylation, which is catalyzed by two *N*-methyltransferases (*ApGSMT* and *ApDMT*). *ApGSMT* is responsible for the two-step methylation reactions of glycine to sarcosine and then sarcosine to dimethylglycine. *ApDMT* is responsible for the specific methylation of dimethylglycine to betaine. Co-expression of *ApGSMT* and *ApDMT* in fresh water *Synechococcus* cells accumulated significant amounts of betaine and conferred sufficient salt tolerance so the cells were capable of growth in sea-water. *Arabidopsis* plants that were transformed with *ApGSMT* and *ApDMT* accumulated large amounts of betaine and increased tolerance to salt stress. However, in this case, the exogenous supply of glycine enhanced the accumulation level of betaine.

In the study, the *Aphanothece* PGDH gene (*ApPGDH*) was isolated and functionally characterized. *A. halophysica* was originally isolated from the Dead Sea and is known to accumulate significant amounts of betaine at high salinity. The *ApPGDH* gene was transferred into *Escherichia coli*, which has a betaine synthetic pathway via choline oxidation; it was also transferred into *Arabidopsis* plants, in which the betaine synthetic pathway was introduced via glycine methylation. The results manifested that the heterogeneous expression of *ApPGDH* significantly enhanced betaine levels in both betaine accumulating and non-accumulating organisms. (Waditee 2007)

Halophiles are found within the Bacteria phyla *Cyanobacteria*, *Proteobacteria*, *Firmicutes*, *Actinobacteria*, *Spirochaetes*, and *Bacteroidetes*. Also within the Archaea most are found in the class *Halobacteria*. *Halobacterium* and most of its relatives require over 100–150 g/l salt for growth and structural stability. Also within the order *Methanococci* we encounter halophilic species. Halophiles and non-halophilic relatives are often found together in the phylogenetic tree, and many genera, families and orders have representatives with greatly different salt requirement and tolerance. A few phylogenetically coherent groups consist of halophiles only: the order *Halobacteriales*, family *Halobacteriaceae* (*Euryarchaeota*) and the anaerobic fermentative bacteria of the order *Halanaerobiales* (*Firmicutes*). In *Halomonadaceae* (*Gammaproteobacteria*), it is exclusively contains halophiles. Halophilic microorganisms use two methodologies for cytoplasm equilibrium, osmotically with their medium. The first involves accumulation of molar concentrations of KCl. This strategy requires adaptation of enzymes of
the intracellular machinery, for example, maintenance of proper conformation of proteins, also protein activity at saturating salt concentrations. The proteome of such organisms is highly acidic, and in low salt environment, proteins cannot function. Such microorganisms generally need high salt to survive. The second strategy is to expel salt from the cell and to synthesize and/or accumulate organic 'compatible' solutes that do not interfere with enzymatic activity. Not much adaptations of the cells' proteome are needed, and organisms using the 'organic-solutes-in strategy' often adapt to a surprisingly broad salt concentration range. Most halophilic Bacteria, but also the halophilic methanogenic Archaea use such organic solutes. Many are solutes known, including glycine betaine, ectoine and other amino acid derivatives, sugars and sugar alcohols. The 'high-salt-in strategy' is not limited to the Halobactereae. The Halanaerobiales (Firmicutes) also accumulate salt rather than organic solutes. A third, phylogenetically unrelated organism accumulates KCl: the red extremely halophilic Salinibacter (Bacteroidetes), recently isolated from saltern crystallizer brines. Analysis of its genome showed many points of resemblance with the Halobacteriaceae, probably resulting from extensive horizontal gene transfer. Salinibacter shows that more unusual types of halophiles may be waiting to be discovered (Oren 2008).

The ability to adapt to changes in the osmolarity of the external environment is therefore of fundamental importance for growth and survival, and as such, prokaryotic cells have evolved a number of osmoadaptive strategies to cope with fluctuations in this important environmental parameter. Hypersaline environments are extremely diverse, and so are the microorganisms that inhabit them (Oren 2008).

Compared to other species, bacteria are the most versatile of all living organisms, surviving in most environments. This successful survival is due to the high management of adaptation and stress control strategies. One such environmental parameter is the osmolarity of the extracellular medium. Bacterial cells maintain an intracellular osmotic pressure greater than the surrounding medium in order to generate turgidity, which is considered the driving force for cell extension, growth and division. The capability of adaptations in osmolarity of the external environment is the key for survival, and as such, prokaryotes have evolved a number of osmoadaptive strategies to cope with variations in this important environmental parameter.
1.5 Red Sea Atlantis II Brine Pool

The theory of the history of the Red Sea development was addressed to have started 25 million years ago through the continuous separation of both the African and Arabian tectonic plates (Osama 2013). The structure of the bottom of Red sea has an axial rift that appears from the southern area to the northern area. In the rift of the Red Sea, twenty-five brine pools were discovered. Due to the high volcanic activity, five millions years ago, the deep basin in the rift were filled brines, which is water with high concentration of minerals and salt, with high temperature and lack of oxygen.

The Atlantis II deep is the largest brine pool in the Red Sea and is located at $21^\circ21'$N on the axial rift, illustrated in Figure 1. Atlantis II deep is considered to be hydrothermally active. There is an observation of for long-term continuous increase in the temperature from 55.9 to 68.2°C over thirty years period (Swift 2010). Atlantis II deep is divided into three main distinctive layers, upper convective layer (UCL) that is 46 meters deep, lower convective layer (LCL) of 135 meters thickness and a seawater-brine interface layer, which is about 14 meters (Anschutz 2000). Atlantis II deep has a maximum depth of about 2194 meters with a sharp increase in both temperature of a maximum 68.2°C and salinity of a maximum 25.7% at the lower convective layer (Antunes 2011). Moreover, the brine is considered anoxia. There is a decrease in Oxygen content to reach 0 mM at LCL layer (Table 1) (Figure 2).

Harsh conditions stated above and adding the acidic pH of Atlantis II deep brines, have attracted recent scientific researchers such as microbiologists to explore this extreme environment looking for novel microbial species, special functional genes and identifying new taxonomic groups (Antunes 2008).

Expenditures were accomplished for Red Sea investigation using sampling technological tools to collect data sampling from different sites and depth. Culturing and Data Analysis were conducted to investigate in the biodiversity microbial communities inhabiting Red Sea (Osama 2013).
1.5.1 Studies Conducted in Atlantis II Brine Pool in the Red Sea

The isolated is part of a highly conserved operon in different bacterial phyla with indiscernible function. In this study, isolation and cloning of 338 amino-acid nitrilase (NitraS-ATII) was done and the enzyme was expressed and purified. The purified NitraS-ATII showed selectivite characterization towards dinitriles, which can be possibley used in industrial application in the synthesis of cyanocarboxylic acids. Moreover, NitraS-ATII showed higher stability in thermal conditions compared to a closely related nitrilase. In addition, it has high tolerance towards to high concentrations of selected heavy metals. This enzyme has demonstrated adaptation of microbes in the Atlantis II Deep LCL to the diverse extreme environment and could be valuable in Process of bioremediations. (Sonbol Et al, 2016)

1.6 Bioinformatics

1.6.1 Assembly

Assemblers were designed to assemble genomes of individual organisms, however, they were not designed to assemble chunks of sequences from different species. The performance of different assembler varies when assembling metagenomics. In general, assembling genomic data structure of whole genome shotgun (WGS) sequences of the sample data, and sorted then arranged in a hierarchal so that all reads that has identity percentage of overlaps are aligned together and form a common contiguous sequence called contig. Next contigs are grouped again if the structures show more similar sequences or overlaps, and a larger structure called scaffolds are formed, taking into consideration the order and the size of the gaps that may separate any two successive contigs to form a complete genome sequence (Jason R. Miller 2010). The accuracy of the output is not assured but a parameter called N50 is used to indicate the quality or the accuracy of the assembly. N50 is defined as the median contig size. It is calculated by sorting the contigs size in descending order then N50 will be the number reached that corresponds to the percentage of the total number of bases contributing in the assembly equals to 50% (Jason R. Miller 2010).
The progress in algorithm of Assembler is tightly related to the developments of sequencing technologies. For the next generation sequencing (NGS), The Overlap/Layout/Consensus (OLC) and de Bruijn Graph (DBG) are broadly used algorithms in assemblers Miller 2010. Contig construction, the output of the assemblers, is the main aim of the algorithms. The difference between the algorithms is how they build the contig. OLC algorithm uses pairwise alignment between all reads and build a graph that contains node of aligned reads and edges that resemble overlaps. Then interpret of the graph into contigs by calling consensus sequences from multiple sequence alignment of the reads. OLC graph construction needs more efficient computational resources, memory and processor, because the size of the dataset becomes larger. Accordingly, OLC algorithm is the favored solution for lower-coverage and long reads (100-800bp), such as Roche/454 sequencing platform in Metagenomics.

Newbler is example of assembler that uses OLC algorithm (Zhenyu 2011) which is the official Roche assembler and distributed by 454 life sciences and Celera (Margulies 2005). In metagenomic datasets studies, the aims is either recreating of genomes from environmental samples or constructing longer pieces of DNA sequences for clear characterization. Contigs then are not the aim of but they are used for further understanding of the structure and function of the microbial community (Desai 2012).

Assemblers use a reference-based or De novo assembly. If closely related reference genome to the metagenomic dataset is found, then it is reference. However, it is called De novo method of metagenomic reads when it is build from scratch without any reference sequence that is due to the complexity of sequences.

In a previous study of hypersaline Lake Tyrrell (LT) in Australia, they deeply-sequenced libraries with both Sanger and 454 pyrosequencing technologies and were assembled independently by using different combinations. The phylogenetic analysis of the assembled contigs resulted in the discovery of two new halophilic archaeal lineages that are highly abundant in the surface water of LT. The reconstruction of these two novel uncultured genomes proved the promising capabilities of De novo metagenomic assembly (Narasingarao 2012).

In another recent study accomplished creating catalogue of the human gut microbial genes by using De novo illumina-based metagenomic sequencing, assembly and characterization of 3.3 million non-redundant human intestinal microbial genes.
1.6.2 Blast

**Aligning** sequence of reads from an environmental sample against referenced database of proteins is highly needed in vast number of studies. Blast is type of alignment that is a bottleneck in bioinformatics. Recent algorithms improved the performance. Blastx is a common tool for the type of blast where the research need to align reads to protein database. In application of metagenomics, where the task is to compare billions of short DNA reads against tens of millions of protein references, BlastX although it is slow in performance (Figure 3) but it is more sensitive and produces more hits (2-BitBio 2017).

BLASTx is “the gold standard for sensitivity” and if long alignment time is not a considerable criteria. Then blastx should be considered for alignment and annotating a transcriptome (Figure 4).

1.6.3 Conserved domains

A domain is a part in the protein that is conserved in a given protein sequence. The protein domain (tertiary) structure can evolve independently and its function, exist of the rest of the protein chain. Domain can form a compact three-dimensional structure and often can be folded indendetly sTable.

The definition of protein domains tends to overlap remarkably with what has appear from systematic analyses of sequence data—the characterization of protein domains as units of molecular evolution. Protein domains were initially described as sTable or autonomously folding units of protein structure, inspired by first results from the experimental characterization of protein three-dimensional (3D) structure. The contents of what has been incorporated into Conserved Domain Database (CDD) is the majority of protein domain models collected in databases such as Pfam (1) and SMART (2), stem from the results of such sequence analyses, and the CDD in-house curation effort has adopted a similar view of protein domains.

A domain model would not be split into smaller parts, even if the analysis of 3D structure suggests the presence of two or more structurally autonomous units, or ancient rearrangements at the gene level have resulted in tandemly repeated units, unless the analysis of sequence and structure databases strongly suggests that fragments homologous to one or more such smaller parts exist in different contexts.
1.7 Metagenomic approaches in the search of novel genes:

The tendency for studying the profound world of uncultured microorganisms is promising. Metagenomic researches have provided us with excess of microbial genomic sequences that can be a great source of novel enzymes, with potential in many industries. Microbial genomes that are extracted from different environments enlarge the possibility of identifying enzymes that can adjust to these environments.

The studies of microbial metagenomic are founded on analyzing the genomes of the microbial uncultured genome by the DNA direct extracting from the community, sidestepping the isolation and cultivation of individuals. Mechanical trimming or DNA breaking down using enzymes constructs libraries of Metagenomic environmental DNA sequences. Gene recognition in the libraries is done using two different approaches, either functional-based or sequencing. In the first approach, the library clones are screened for specific enzyme activity or protein function (Gong J-S 2013). In the second methodology, gene screening is based on comparisons between sequence homology with available sequences in the existing databases. This strategy combines bioinformatics, sequence analysis, sequencing techniques, PCR amplification and/or DNA synthesis (Gong J-S 2013). Different enzymes with different functions, retrieved from accessible databases, then annotated based on homologous sequence similarities. Available softwares such as Pfam can determine known patterns in gene sequences and thus assign these sequences to groups with major functions. However, Experimental testing determines the enzyme specific functions and substrate specificity (Seffernick et al 2009). In many studies done, several enzymes were recognized using sequence homology technique followed by expression of the targeted sequences in suitable bacterial hosts (Seffernick et al 2009).

1.8 Study objectives and experimental design

The main objective of this research is to understand the distribution of Betaine biosynthetic pathways in the microbial community of the red sea. To achieve the objective sequential analysis and computations are needed, to be processed on the metagenomics libraries available from various micro environments in the Red Sea for the presence of all enzymes involved in the two branches of Betaine biosynthesis.
The enzymes targeted in this research are grouped into two branches in the Betaine biosynthetic pathways. The first branch, the glycine pathway, there’re are four enzymes named; glycine N-methyltransferase, glycine/sarcosine N-methyltransferase, sarcosine/dimethylglycine N-methyltransferase and dimethylglycine N-methyltransferase. The second branch, the choline pathway, include two enzymes, choline oxidase and betaine-aldehyde dehydrogenase.

Assemble the matching sequences to develop full-length assembled sequences with 50% coverage of each enzyme.

The protein sequences of the enzymes listed above will be utilized independently for blasting the available metagenomics libraries for microbial community of the Red Sea.

Identify the conserved domains for each enzyme, and add the taxonomy for each hit.

Conducting multiple alignment, homology analysis and phylogeny for the assembled DNA sequences.

Classify the richness of the each pathway in libraries generated from different depth of the brine pools and water column.

Generating reports using tool of statistics. And visualizing the results for summarized output.

The objective of this study was to identify the enzymes from the several depths of Atlantis II Deep brine pool and column in the Red Sea, using a sequence-based metagenomic approach. To achieve this goal, extensive bioinformatics and computational work was done on the microbial sequences database.

The second objective was to recognize the abundance of each enzyme vs the depth, considering the amount of the sample available in that depth, and the properties of each site of Atlantis II.

The third objective was concluding the preferable pathway for salt adaptation mechanism in the cell.

The fourth objective was determination of the numbers and types of bacteria that has the abundant number in each depth. Also detect the numbers and types of bacteria the identified pathway.
Figure 1: Sites of brine pools in Red Sea. It is clear that Atlantis II is the largest one. (Zierenberg 1988)
Figure 2: Atlantis II Deep region of the Red Sea.

Showing internal layers, depth in meters, and temperature in Celsius. Adjacent is the halo kinetic mass body., Blank 1995
Alignment time per program

**Figure 3:** The figure illustrates the comparison between PLASTx, BLASTx, and DIAMONDS in time consumed. BLASTx takes longer performance time. (BitBio 2017)
**Figure 4**: BLASTX vs other programs: N PERFORMANCE BLASTX STILL IS THE HIGHEST ACCURACY. (BitBio 2017)
Table 1: Physical characteristics of the Atlantis II Brine Pool, thickness of layer, salinity, temperature in each layer, and percentage of oxygen in water. LCL has the highest degree in Temperature and salinity, but oxygen is almost 0.

<table>
<thead>
<tr>
<th>Depth (m)</th>
<th>Description</th>
<th>Thickness (m)</th>
<th>Salinity (%)</th>
<th>Temperature(C)</th>
<th>Oxygen (mM)</th>
</tr>
</thead>
<tbody>
<tr>
<td>200</td>
<td>Column</td>
<td>4</td>
<td>22</td>
<td>100</td>
<td></td>
</tr>
<tr>
<td>700</td>
<td>Column</td>
<td>4</td>
<td>22</td>
<td>100</td>
<td></td>
</tr>
<tr>
<td>1500</td>
<td>Column</td>
<td>4</td>
<td>22</td>
<td>100</td>
<td></td>
</tr>
<tr>
<td>2065</td>
<td>UCL</td>
<td>46</td>
<td>9-16</td>
<td>43-57</td>
<td>15</td>
</tr>
<tr>
<td>2200</td>
<td>LCL</td>
<td>135</td>
<td>26</td>
<td>68</td>
<td>0</td>
</tr>
</tbody>
</table>
**Figure 5: Glycine Betaine Biosynthesis Pathway**
CHAPTER 2: MATERIALS AND METHODS

2.1 Data Processing Conducted:

A sequence-based metagenomic approach was used for identifying the genes of interest in the 200m, 700m, 1500m, UCL, and LCL databases. The following processing was conducted for the five randomly chosen files for five depths in the Atlantis II. The Libraries were randomly chosen, B, C, and A from depth 200m from library C, 700m from library C, and 1500m from library C, and Brine Pool UCL library A and LCL from library B. All calculations were done on Red Hat Linux server. Data processing steps were performed, which is the first part of the research made. Steps are as follows, assembling the sequences, BLASTingX the consensus sequence of the contig, with the greatest number of reads, against subgroup of against the non redundant protein database in the National Center for Biotechnology Information (NCBI), Domain blast, taxonomy, data filtration, and finally analyzing the data, Figure 5 document the process.

2.1.1 Reads Assembly

Official Roche assembler, Newbler Java GUI (gsAssembler 3.5), was used to assemble files that contain reads sequences of 454 sequencing data resulted from three independent sequencing runs. This was achieved by stringency of the assembly parameters identity percentage and overlap length 90% / 40bp (default values). All other parameters used are the default. The computations were performed on Red Hat Linux server.

In Table 3 it is clear that the input data of AT1500 has the largest number of reads and AT0700 is the smallest. The environment starting from 1500m deep is rich with large community, baring in mind that the season, and time are controlled in sampling. Data comparisons show that the reads of the files of depth 200 and 700 are smaller than the three other files, therefore, the number of bases. Depth 200m and 700m are almost half the size of depth 1500 and the brine. The obtained sequences were assembled resulting in contigs. All contigs were tabulated as in Table 2.
2.1.2 Sequence Annotation by BlastX

The resulted assembled contigs were used as the input queries fed to Align Sequences Translated BLAST: Blastx. Blastx tool aligned ORFs against NCBI protein database and create best hit file using E-value of 1e-5. All ORFs from the assembled contigs were compared to NCBI nr database (version date December-2015) using BLASTX tool. Default parameters for BLASTX were applied and the number of best hits retrieved. Blastx were performed against nr database searching for 6 targeted enzymes

- Betaine-aldehyde dehydrogenase
- Choline dehydrogenase
- Sarcosine/dimethylglycine N-methyltransferase
- Glycine/sarcosine N-methyltransferase
- dimethylglycine N-methyltransferase
- Glycine N-methyltransferase

BlastX was repeated for the five group of contigs obtained for all depth used in the study. All data were tabulated in excel sheets for better visualization and to prepare them for the analysis and filtration.

2.1.3 Details Collection for the hits

2.1.3.1 Taxonomy

Detailed data is collected for each hit. Collecting information for the expected species that the sequence expressed in. The taxonomy was assigned based on the query match with reference and checking on NCBI genebank to get the taxonomy details.

Script was used to extract reference-hit number to map it to the NCBI database.

From Table 3 of total Data Comparisons that the reads of the files of depth 200 and 700 are smaller than the three other files, therefore, the number of bases. Depth 200m and 700m are almost half the size of depth 1500 and the brine.

2.1.3.2 Conserved Domain

Conserved Domain identification step was essentially needed for Further analysis.
Extracting all hits, and downloading all necessary data needed from NCBI and UniProt using Perl script. Then blasting the hits and aligned them to match conserved domains in the Uniprot conserved Domains database. This step is done to identify the regions where the conserved domains are located.

### 2.1.3.3 Pathways

Pathway and gene ontology identification was performed using Kegg Pathway Database. This extra step was conducted to collect more description for each hit. No filtration was constructed until this step.

### 2.2 Data Analysis

All previous outputs collected are now tabulated in one file for each depth, and files are ready for analysis. In the following steps data filtration and analysis were conducted, Figure 6.

Data Filtration was done before starting analysis. The hits were chosen for further analysis has some criteria. For the match length, only numbers above 50% of the enzyme average length were considered. For enzymes with average amino acid length 500aa, match length considered from 250aa. Data for “E-value below <= .005” were considered, otherwise data were discarded. Then abundance quantitative estimation for the pathways was performed. Each occurrence that satisfies the conditions was counted. Script was used to count the occurrence of the targeted enzymes in each considered file. Followed Taxonomical Abundance was counted. An overview screening of taxonomy spread within each pathway where calculated for each depth. For pathway comparison, special focus was done on the two pathways of Glycine Betaine biosynthesis, and comparison was done between the weights of hits to make an overview of the preferable pathways in each depth. Next, conserved Domain search was done for hits in the 5 depths for enzymes of interests. Output from NCBI CDD and Uniprot were scanned to investigate whether each filtered enzyme sequence is positive or false hit with the help of data was collected in Gene Ontology (GO) for each entry. Also, Choline Pathway Taxonomy Mapping Choline pathway two enzymes, Betaine aldehyde dehydrogenase and choline dehydrogenase taxonomy were extracted from the large file, and mapped together. For each depth this step was repeated. All similarities and differences were monitored and recorded. See Figure 7 a flow chart of methodology.
Figure 6: Work Flow: Steps of the process made, assembled the sequences, blasting the contigs against subgroup of nr, Collecting Detailed Description for the hits from NCBI, uniProt, and Kegg, to describe each reference ontology, and species blast, taxonomy.
Filteration

- Discard all hits that have sequence length less than 50% of protein coverage and evalue more than

Computational Abundance

- Hits were collected for each enzyme in each depth for quantitative description.

Taxonomical Abundance

- Counting each taxonomy for each enzyme

Pathway Element Comparisons

- Comparisons between all six enzymes for each depth

Domain analysis

- Investigating conserved domains for each enzyme

Choline Pathway Taxonomy Mapping

- Taxonomy were extracted for ChA and BADH matched together, and mapped.

Figure 7: Data Analysis workflow. Data analysis done to show the preferred pathway for organism lived in the selected depths
Table 2: Tabulating contigs for LCL.

<table>
<thead>
<tr>
<th>Contig Number</th>
<th>Contig Length</th>
</tr>
</thead>
<tbody>
<tr>
<td>contig13105</td>
<td>18160</td>
</tr>
<tr>
<td>contig17785</td>
<td>12481</td>
</tr>
<tr>
<td>contig17786</td>
<td>11690</td>
</tr>
<tr>
<td>contig17784</td>
<td>10027</td>
</tr>
<tr>
<td>contig08562</td>
<td>9919</td>
</tr>
<tr>
<td>contig13113</td>
<td>9805</td>
</tr>
<tr>
<td>contig06990</td>
<td>9400</td>
</tr>
</tbody>
</table>
CHAPTER 3: RESULTS AND DISCUSSION

3.1 Screening of Glycine Betaine biosynthetic pathways

Sequential computations and analysis were made on Metagenomic libraries for the presence of enzymes involved in the two branches of Glycine Betaine biosynthesis, in order to screen the distribution of Glycine Betaine pathways in the microbial community of Atlantic II in the red sea.

The enzymes targeted in this research are grouped into two branches in Glycine Betaine biosynthetic pathways. In the first branch, the glycine pathway, there are four enzymes named; glycine N-methyltransferase, glycine/sarcosine N-methyltransferase, sarcosine/dimethylglycine N-methyltransferase and dimethylglycine N-methyltransferase. The second branch, the choline pathway, includes two enzymes, choline dehydrogenase and betaine-aldehyde dehydrogenase. Targeted Enzymes were identified in contigs query extracted from the Metagenomics library in five depths of ATII.

3.1.1 Analysis of the Assembly results

The assembly produced long sequences that is fed to Blastx. Files assembled were from different libraries and different runs from Atlantis II Column 200m, 700m, and 1500m, and Brine UCL and LCL, giving output-assembled fasta files. Contigs produced contains 1.2Kbp and above are around 1000 contigs in each file. Summery of 5 output assemblies are tabulated.

Table 2 shows the contigs produced from assembly and assembly length. In this Table the assembly results displayed are for LCL sequence file, the rest of the results are not shown.

In Table 3 that displays the input files data, it is clear that depth 1500m has the largest reads. The result of the assembly in Table 4 shows the summery of the contigs produced. Upper convective layer read has produced the largest contig number comparing to the sample files. The largest total no of assembled reads is found in depth ATII UCL. It is also clear that N50 value is the highest in ATII 1500m.
3.1.2 Glycine Betaine pathway enzymes identification

All BlastX results are annotated and tabulated in Excel files for each site. Glycine Pathway enzymes were identified. Also Betaine Pathway enzymes were identified. Each pathway showed different weight.

The encoded protein sequence of the genes of Glycine branch consists of 270 to 300 amino acids on average. Enzymes involved are glycine N-methyltransferase, glycine/sarcosine N-methyltransferase, sarcosine/dimethylglycine N-methyltransferase and dimethylglycine N-methyltransferase. The second branch, the choline pathway, includes two enzymes, choline dehydrogenase and betaine-aldehyde dehydrogenase. The encoded protein sequence of the genes of Choline branch consists of 550 and 490 amino acids correspondence. Targeted Enzymes were identified in contigs query extracted from the Metagenomics library in five depths of ATII. Sequences are located on full-length match on contigs.

3.1.3 Distribution of Glycine Betaine biosynthetic pathways

Data are displayed in Figure 8 and Figure 9 to show general distribution Glycine and Choline pathways in the 5 depths of interest.

3.1.3.1 Glycine Pathway

In AT200, BlastX identify 4 occurrence of Dimethylglycine N-methyltransferase, with 152aa, which is 52% of actual length, as the average length of this enzyme is 270aa and 2.00E-06 E-Value, occurred in Aphanothece halophytica. E-value is significant considering that query length is 450aa.

Glycine/sarcosine N-methyltransferase, had 16 occurrences, with maximum match length of 189aa and Evalue and 2.00E-07 occurred in Actinopolyspora saharensis.

Glycine N-methyltransferase occurred twice in AT200m with match length 259 and E-value 6.00E-06 occurred in Mycobacterium marinum. This is significant occurrence and full length, however, small amount of hits.

Sarcosine/dimethylglycine N-methyltransferase had 34 hits, the highest match length occurred in Actinopolyspora xinjiangensis, with match length 192aa, while the full length is 252aa in the enzyme copy of this species.
In AT700, BlastX identify 2 occurrence of Dimethylglycine N-methyltransferase, with 150aa and 0.008 E-Value. The average length of this enzyme is 270aa. E-value is considered mathematically significant; also E-value has a high value because the calculation of E-value considers the length of the query 490. However, considering E-value, Query length, match length that is 52% of actual length, and number of hits, this hit is not significant.

Glycine/sarcosine N-methyltransferase, had 22 occurrences, with maximum match length of 122aa and E-value 0.004 occurred in Virgibacillus massiliensis. Full average length of this enzyme is.

Glycine N-methyltransferase occurred twice in AT700m with match length 259 and E-value 6.00E-06 occurred in Mycobacterium marinum. This is significant occurrence and full length, however, small amount of hits.

Sarcosine/dimethylglycine N-methyltransferase had 13 hits, the highest match length occurred in Ruegeria atlantica, with match length 192aa, while the full length is 252aa in the enzyme copy of this species.

In AT1500, BlastX identifies 10 occurrence of Dimethylglycine N-methyltransferase, with 183aa and 7.0E-06 E-Value, Halorhodospira halochloris. The average length of this enzyme is 279aa. E-value is considered mathematically significant, and comparison between query length and reference hit is also considered acceTable.

Glycine/sarcosine N-methyltransferase, had 105 occurrences, with maximum match length of 235aa and E-value 1E-51 occurred in Streptomyces melanosphorofaciens. Same length appeared in Lachnospira pectinoschiza with higher E-value 3.0E-07.

Glycine N-methyltransferase occurred 26 with match length 124 and poor E-value range of 0.53. The reference hit shows occurrence in Otolemur garnettii.

Sarcosine/dimethylglycine N-methyltransferase had 88 hit, the highest match length occurred in, with match length 234aa, occurred in Streptomyces reticuli, Rhodospira trueperi and Prauserella marina.

In Upper convective layer, Dimethylglycine N-methyltransferase, was found in 50% length 138aa 24 times, the highest length occurrence was found Prochlorococcus marinu with E-value of 0.005.
Glycine N-methyltransferase occurred 28 with match length 158 and E-value range of 0.05. The reference hit shows occurrence in Erinaceus europaeus.

Glycine/sarcosine N-methyltransferase, had 290 occurrences, with maximum match length of 229aa and E-value 0.073 occurred in Salinisphaera hydrothermalis, the match length equals to 84% length of the full enzyme.

Sarcosine/dimethylglycine N-methyltransferase appeared 198 times with more than 83% of full length. The highest number appeared in Ruegeria atlantica with 0.0006 E-value, and value 192, and full length of 230aa.

ATIIIILCL, In Lower convective layer, Dimethylglycine N-methyltransferase, was found in 66% length 177aa 24 times, the highest length occurrence was found Halorhodospira halochloris and Planctomyces, with E-value of 0.005.

Glycine N-methyltransferase occurred 24 with match length 229 and poor E-value range of 0.0001. The reference hit shows occurrence in Mycobacterium marinum.

Glycine/sarcosine N-methyltransferase, had 240 occurrences, with maximum match length of 216aa and E-value 0.005 occurred in Coprococcus Comes, the match length equals to 84% length of the full enzyme.

Sarcosine/dimethylglycine N-methyltransferase appeared 230 times with more than 83% of full length. The highest number appeared in Ruegeria atlantica with 0.0006 E-value, and value 192, and full length of 230aa.

Glycine pathway is identified in small amount in Brine Pool layers and 1500m. However, characterization is found in higher percentage in 200M and 700m depth. Glycine pathway enzymes glycine/sarcosine N-methyltransferase, sarcosine/dimethylglycine N-methyltransferase are all recorded.

Four enzymes, glycine N-methyltransferase, glycine/sarcosine N-methyltransferase, sarcosine/dimethylglycine N-methyltransferase and dimethylglycine N-methyltransferase, in Glycine pathway, were identified in all layers but with considerably few hits in the targeted sites, for coverage 100% to 50% full length of the enzyme. Also the presence of their conserved domain is considerably small.
However, dimethylglycine N-methyltransferase glycine N-methyltransferase had small significant recording. This can indicate that sarcosine has source of production other than dimethylglycine N-methyltransferase glycine N-methyltransferase. Table 5 has all recorded numbers for each enzyme in the pathway, considering only hits above 200aa and E-value less than .005.

3.1.3.2 Choline Pathway

In ATII200m, Choline dehydrogenase showed seldom occurrence in depth 200m

Betaine Aldehyde dehydrogenase occurred 19 times. Highest hit is 230 aa with E-value 5.00E-37.

In ATII700m, Choline dehydrogenase highest hit 126aa, which is less than half the number of 549aa copy in Amycolatopsis orientalis, so data was discarded according to filtration criteria set to this research.

Betaine Aldehyde dehydrogenase occurred 979 times. Highest hit is 315 aa with E-value 2.00E-83 in Devosia and in Sinorhizobium, which is 70% of the full length and significant E-value.

In ATII1500m, Choline dehydrogenase, had occurrences of 476, with maximum match length of 260aa and E-value 4E-96 occurred in Halobacillus dabanensis, the original copy is 560.

Betaine Aldehyde dehydrogenase occurred 7500 times. Highest hit is 484 aa with E-value 3.00E-128 in Bacillus coagulans, which is 97% of the full length and significant E-value. 9 occurrence appeared with more than 90% full length, in Bacillus coagulans and in Bacillus mycoides with E-value of 2E-124.

Choline Dehydrogenase occurred in ATIIUCL with almost full length as maximum finding of the original copy of the enzyme, value is 533, in Pantoea agglomeran, 2.00E-104, in contig13623

Betaine Aldehyde dehydrogenase occurred 5400 times, between Eukaryotes and Bacteria, with full protein length. We will focus on Bacteria here highest hit is 488aa with E-value 2.00E-79 occurred in Bacillus cytotoxicu 91% of the full length and E-value is significant.

Choline Dehydrogenase occurred in ATIIILCL with 50% length as maximum finding of the original copy of the enzyme.
Betaine Aldehyde dehydrogenase occurred 3700 times, Eukaryotes and Bacteria, with full protein length in contig00191. We will focus on Bacteria here highest hit is 455 aa with E-value 2.00E-92 occurred in Bacillus coagulan 91% of the full length and E-value is significant.

Also, a significant finding of the protein is found in Bacillus amyloliquefaciens 91% coverouge. In Eukaryotes hits exceeded the 92% of the original copy of the enzyme.

Choline pathway is identified in vast amount in Brine Pool layers and 1500m. It is shown in Table 6 that the ATII depth 1500m contains the highest weight of Betaine aldehyde dehydrogenase significant 7500 hits after filtration. And the least amount appears in AT II depth 200m, significant 19 hit.

The enzymes involved in this branch have been screened in significant amount in ATII Brine Pool Upper and Lower Convictive layers. The pathway of choline has more abundant occurrence in the scope of UCL and LCL where the salinity is very high. Also in column 1500m, there are plenty amount of choline Dehydrogenase and betaine-aldehyde dehydrogenase. A significant drop is found in 700m and 200m. In Choline dehydrogenase, there are almost consistence of the number of the three depth of LCL, UCL, and 1500m an observed significant decrease of Betaine aldehyde dehydrogenase appears in the depth 200m. In Figure 10, it is very clear that depth 200 has seldom appearance of Betaine aldehyde dehydrogenase.

Then a sudden drop of choline is very clear in Figure 10 and Table 7 in depth 200m and 700m.

This could be due to drop of salinity concentration in this depth. Also, another interpretation can be that organism in this depth use another mechanism to adapt to the high salinity. In depth 200m the input file used in the research has considerably fewer reads, maybe due to smaller community in this site or to poor sequencing. Adding that we cannot ignore that the results of assembly that shows small N50 in depth 200m. In Table 7 all data are collected to show the comparisons for each branch of the pathway against the depth.

**3.1.4 Domain analysis for identified enzymes**

Further analysis was done on the hits, to insure enzyme screening, which is searching for the known conserved domains of the enzymes and gene ontology. After blastx was done, conserved domains were mapped from NCBI with full sequence length. If blastx showed the hits, then the conserved domains of the enzymes were found among the hits, gives clearer screening, that the targeted enzyme is detected.
3.1.4.1 Glycine Pathway

The conserved domain search was made to detect the enzyme occurrence and its core activity in the sequences. Glycine pathway domain analysis was identified in the five layers of investigation. For Glycine pathway, SAM-dependent methyltransferase, Methyltransferase domain, Tocopherol O-methyltransferase (PLN02244), and S-adenosylmethionine-dependent methyltransferases S, and Acetyltransferase (GNAT) domains were detected for Glycine/Sarcosine N-Methyltransferase. Methyltransferase domain, SAM-dependent methyltransferase, Glycine N-methyltransferase activity domain, Mycolic acid cyclopropane synthetase, and L-asparagine transporter and related permeases made hits, which suggest the activity of Glycine N-Methyltransferase. SAM-dependent methyltransferase, Methyltransferase domain, S-adenosylmethionine-dependent methyltransferases, Tocopherol, O-methyltransferase, Mycolic acid cyclopropane synthetase were screened to point for Sarcosine/Dimethylglycine N-Methyltransferase activity Table 10.

3.1.4.2 Choline Pathway

Blastx had detection for Choline pathway enzymes. Conserved domains also where blasted, to confirm that the enzyme hits actually point to the correct enzymes. The following domains where recorded for Choline Dehydrogenase, PRK02106, GMC oxidoreductase region_name="GMC_oxred_C"GMC oxidoreductase, Rossmann-fold NAD(P)(+)-binding proteins, and pyranose oxidase.

For Betaine Aldehyde Dehydrogenase, active domain recorded are NAD(P)+-dependent aldehyde dehydrogenase, Pseudomonas fluorescens 4-hydroxymuconic, uncharacterized Candidatus pelagibacter aldehyde, NAD+-dependent betaine aldehyde, Rhodococcus ruber 6-oxolauric acid, aldehyde dehydrogenase family 2 member and Rhodococcus ruber 6-oxolauric Acid, and oxidoreductase activity, acting on the aldehyde or oxo group of donors, NAD or NADP as acceptor, Rhodococcus ruber 6-oxolauric acid, and Pseudomonas fluorescens 4-hydroxymuconic Table 9.

3.1.5 Taxonomical Screening and Organism Counts

We have collected information about each hit. For example, we know that this hit belongs to one enzyme of Glycine pathway and that is derived from a specific species. It was a confirmation step to compare between species that the hit for enzymes in each pathway exists.
in. The confirmation is that to check if enzymes of each pathway appear in same species or not. If all enzymes of each pathway exist in the same species that gives better information that the pathway is active. Also this comparison can tell which species appear more frequently.

**3.1.5.1 Choline pathway**

For Choline pathway, organism counting was performed for UCL, LCL, 1500, and 700.

For LCL, 249 Taxa made appears in both Betaine Aldehyde Dehydrogenase and Choline Dehydrogenase, 139 in Betaine ADH only, and 49 in Choline, Figure 11.

For UCL and 1500m, 30 species type shared the two enzymes, and 12 species type shared the two enzymes.

The two enzymes of the choline pathways appeared together in many organisms, which suggest that this pathway is functioning in different species. The highest number of sharing, appeared in the LCL layer, where salinity is very high in the lower layer and organisms operate their surviving mechanisms.

**3.1.6 Glycine Betaine Pathway Data Normalization**

In Table 8, data normalization was needed to compare between the abundance of each integrated pathway through the sea depth, because data in each depth varies in sample size.

The results of the normalization appear in Figure 12. Choline pathway hits increases from depth 200 gradually to depth LCL with a sharp raise in 1500m. Going down in sea depth the salinity increases, which increase the need of osmoregulation pathway. In Glycine pathway the increase is steady and increase with the increase of salinity. In Choline Pathway the increase is linear except for depth 1500m, as the an enormous amount of positive hits encountered by Betaine Aldehyde Dehydrogenase, could be due to high diversity of the enzyme, or could be high demand by many species in osmoregulation or other functionality.

In Figure 12, an increase of choline pathway appears in 1500m and UCL which conflicts with the hypothesis that the more salinity the more need of Betaine Glycine for osmoregulation and cope with the high salinity. The expected result was that a linear increase from 1500m to LCL, but a negative linear slope is found. Figure 11, present a possible interpretation for this appearing phenomenon. Both Betaine Aldehyde Dehydrogenase and Choline Dehydrogenase
appear separately in an amount of species, not intersected, which means that not all hits involve in the pathway.

3.2 Prevailed Species

After performing the species count for Choline Pathway, another feature was conducted. Bacteria; Firmicutes; Bacilli; Bacillales; Bacillaceae; Bacillus had the highest abundance in LCL, UCL, 1500, and 700. Many prevailed species were found as well. At LCL the prevailed species that had Betaine Aldehyde Dehydrogenase shown in Figure 13.
Table 3: Total reads and bases. This Table combines the files of the sample, showing the size of each file in number of reads and number of bases. AT1500 shows the largest number of reads. AT0200 and 0700m show a smaller sample, which indicates smaller community.

<table>
<thead>
<tr>
<th>Site</th>
<th>AT0200m</th>
<th>AT0700m</th>
<th>AT01500m</th>
<th>ATBRUCL</th>
<th>ATBRLCL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reads Number</td>
<td>190351</td>
<td>184048</td>
<td>494236</td>
<td>451675</td>
<td>405363</td>
</tr>
<tr>
<td>Bases Number</td>
<td>69058938</td>
<td>65908986</td>
<td>243626393</td>
<td>168594599</td>
<td>162751879</td>
</tr>
</tbody>
</table>

Table 4: Summary of Assembly. Main properties of sample assembly are illustrated; Consensus results are displayed in this integrated Table against ATII 200, 700, 1500, and LCL and UCL.

<table>
<thead>
<tr>
<th>Sample Site / Properties of Assembly</th>
<th>ATII UCL</th>
<th>ATII 200m</th>
<th>ATII 700m</th>
<th>ATII 1500m</th>
<th>ATII LCL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total no of assembled reads</td>
<td>326010</td>
<td>12771</td>
<td>19534</td>
<td>98573</td>
<td>68925</td>
</tr>
<tr>
<td>Total no of single tons</td>
<td>59731,</td>
<td>162418,</td>
<td>147488,</td>
<td>344709</td>
<td>289638</td>
</tr>
<tr>
<td>Largest contig</td>
<td>35746</td>
<td>5072</td>
<td>5790</td>
<td>10286</td>
<td>10575</td>
</tr>
<tr>
<td>N50 contig size</td>
<td>1018</td>
<td>759</td>
<td>943</td>
<td>1236</td>
<td>1017</td>
</tr>
<tr>
<td>Total no of contig</td>
<td>18660</td>
<td>2863</td>
<td>1379</td>
<td>12694</td>
<td>4639</td>
</tr>
</tbody>
</table>
Table 5: Quantitative Hits For Glycine Pathway. Hit are displayed, depth LCL and UCL has the largest quantities.

<table>
<thead>
<tr>
<th>Enzyme</th>
<th>LCL</th>
<th>UCL</th>
<th>AT1500</th>
<th>AT700</th>
<th>AT200</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sacrocin DGNM</td>
<td>57</td>
<td>198</td>
<td>88</td>
<td>13</td>
<td>34</td>
</tr>
<tr>
<td>Glycin SNM</td>
<td>40</td>
<td>290</td>
<td>105</td>
<td>22</td>
<td>16</td>
</tr>
<tr>
<td>GNM</td>
<td>13</td>
<td>28</td>
<td>26</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>DIG</td>
<td>14</td>
<td>24</td>
<td>10</td>
<td>2</td>
<td>4</td>
</tr>
</tbody>
</table>

Table 6: Hits For Choline Pathway. BetaineADH has the largest hits in the LCL, UCL, and 1500m

<table>
<thead>
<tr>
<th>Enzyme</th>
<th>LCL</th>
<th>UCL</th>
<th>AT1500</th>
<th>AT700</th>
<th>AT200</th>
</tr>
</thead>
<tbody>
<tr>
<td>BADH</td>
<td>3700</td>
<td>5400</td>
<td>7500</td>
<td>979</td>
<td>19</td>
</tr>
<tr>
<td>CDH</td>
<td>466</td>
<td>422</td>
<td>476</td>
<td>0</td>
<td>3</td>
</tr>
</tbody>
</table>

Table 7: Distribution of Glycine Betaine Pathway Quantitative

<table>
<thead>
<tr>
<th>Enzyme</th>
<th>LCL</th>
<th>UCL</th>
<th>AT1500</th>
<th>AT700</th>
<th>AT200</th>
</tr>
</thead>
<tbody>
<tr>
<td>Choline PW</td>
<td>4166</td>
<td>5822</td>
<td>796</td>
<td>979</td>
<td>22</td>
</tr>
<tr>
<td>Glycin PW</td>
<td>124</td>
<td>540</td>
<td>229</td>
<td>39</td>
<td>56</td>
</tr>
</tbody>
</table>

Table 8: Normalization integrated data for Glycine Betaine Pathways in 5 depths

<table>
<thead>
<tr>
<th>Site</th>
<th>AT0200m %</th>
<th>AT0700m %</th>
<th>AT01500m %</th>
<th>ATBRUCL %</th>
<th>ATBRLCL %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Choline PW</td>
<td>28.39%</td>
<td>0.0115</td>
<td>96.3%</td>
<td>1.6</td>
<td>91.6%</td>
</tr>
<tr>
<td>Glycin PW</td>
<td>71.6%</td>
<td>0.029</td>
<td>3.6%</td>
<td>0.02</td>
<td>2.79%</td>
</tr>
<tr>
<td>Salinity</td>
<td>4</td>
<td>4</td>
<td>8</td>
<td>9-16</td>
<td>26</td>
</tr>
<tr>
<td>Betaine Aldehyde Dehydrogenase</td>
<td>Choline Dehydrogenase</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>--------------------------------</td>
<td>-----------------------</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NAD(P)+-dependent aldehyde dehydrogenase</td>
<td>PRK02106</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pseudomonas fluorescens 4-hydroxymuconic</td>
<td>GMC oxidoreductase</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Uncharacterized Candidatus pelagibacter aldehyde</td>
<td>GMC_oxred_C''</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bacillus</td>
<td>GMC oxidoreductase;</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>subtilis NAD+-dependent betaine aldehyde</td>
<td>Rossmann-fold NAD(P)(+)-binding proteins</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rhodococcus ruber 6-oxolauric acid</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>aldehyde dehydrogenase family 2 member''</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rhodococcus ruber 6-oxolauric acid</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DIMETHYLGLYCINE N-METHYLTRANSFERASE</td>
<td>GLYCINE/SARCOSINE N-METHYLTRANSFERASE</td>
<td>GLYCINE N-METHYLTRANSFERASE</td>
<td>SARCOSINE/DIMETHYLGLYCINE N-METHYLTRANSFERASE</td>
<td></td>
<td></td>
</tr>
<tr>
<td>-----------------------------------</td>
<td>--------------------------------------</td>
<td>-----------------------------</td>
<td>-----------------------------------------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mycolic acid cyclopropane synthetase</td>
<td></td>
<td></td>
<td>Mycolic acid cyclopropane synthetase</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tocopherol O-methyltransferase (PLN02244) Dimethylglycine N-methyltransferase</td>
<td></td>
<td></td>
<td>Mycolic acid cyclopropane synthetase</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Figure 8: Choline Pathway abundance. BADH shows the most abundance in AT1500, UCL, LCL and less abundance in AT700 and AT200.
Figure 9: Glycine Pathway abundance. Here the figure shows all five sites together and how each enzyme of the Glycine pathway is distributed.
Figure 10: Each circle resembles an enzyme. The area represents the number of hits produced. The scale was plotted proportionately to compare the numerical quantity of each enzyme. In LCL, UCL, 1500M, 700M, 200M comparing the six enzymes of the two branches of Glycine Betaine Pathway.
Figure 11: Shows the intersection and disjointed between the two proteins in Choline pathways in species' occurrence. The intersection area shows that both enzymes exist in the same species.
Figure 12: Comparison done in abundance between the two pathways Choline and Glycine. Glycine pathway shoes consistent increase with the increase of salinity. Data showed in percentage.
Figure 13: Taxonomy counts of Betaine Aldehyde Dehydrogenase in LCL
CHAPTER 4: CONCLUSIONS AND RECOMMENDATIONS

4.1 Conclusion

In conclusion, we succeeded in this study to establish an assembled Metagenomics dataset of the Red Sea Atlantis II five different depths and perform a targeted screening for osmoprotectants in different sites. The work is considered the first step towards the establishment of a large database of assembled datasets from a novel and unexplored environment which will provide a tool for further studies regarding the community structure, function, and mechanisms that adapt microbial community to survive in these exceptional harsh conditions.

After conducting our study of distribution of Glycine Betaine pathways in the microbial community of the red sea. We conclude that the two branches of Glycine Betaine pathways are detected in Atlantis II but with different concentration and distribution in each depth, in the Red Sea. Choline pathway large abundance suggests preferable consumption.

4.2 Recommendation for Further Research

Functional metagenomics is a culture independent approach, which is based on the construction of gene libraries using environmental DNA and subsequent functional screening of the resulting clones to search for enzymatic activities. Advantages of this approach include the identification of functional genes during the screening and also that the nucleotide sequences retrieved are not derived from previously sequenced genes, which enables the identification of both novel and known genes (Simon and Daniel, 2009; López-Pérez and Mirete, 2014). For this study, further identification needed with Functional Metagenomics and wet lab by extracting the regions from the contig and test the functionality of the enzymes.
References

6. 2-BitBio 2017
7. CIRCOS TABLE VIEWER V0.63-9. 2008-2017 MARTIN KRZYWINSKI
8. CLONING, EXPRESSION, AND CHARACTERIZATION OF AN ALDEHYDE DEHYDROGENASE FROM ESCHERICHIA COLI K-12 THAT UTILIZES 3-HYDROXYPROPIONALDEHYDE AS A SUBSTRATE. APPL. MICROBIOL. BIOTECHNOL. 81, 51–60
15. JASON R. MILLER, SERGEY KOREN, AND G. SUTTON, ASSEMBLY ALGORITHMS FOR NEXT-


22. MOHAMED, YASMIN; GHAZY, MOHAMED; SAYED, AHMED; OUF, AMGED; EL DORRY, HAMZA; SIAM, RANIA; ISOLATION AND CHARACTERIZATION OF A HEAVY METAL-RESISTANT, THERMOPHILIC ESTERASE FROM A RED SEA BRINE POOL. NATURE. 2013 NOVEMBER. SREP03358


26. RATHNASABAPATHI, B; BURNET, M; RUSSELL, BL; GAGE, DA; LIAO PC; NYE GJ; SCOTT P; GOLBECK JH; HANSON AD (1997). CHOLINE MONOOXYGENASE, AN UNUSUAL IRON-SULFUR ENZYME CATALYZING THE FIRST STEP OF GLYCINE BETAINES SYNTHESIS IN PLANTS: PROSTHETIC GROUP CHARACTERIZATION AND CDNA CLONING. PROC NATL ACAD SCI U S A. 1997 APR 1;94 (7):3454-8.


30. SONBOL, SARAH, AND ET AL. RED SEA ATLANTIS II BRINE POOL NITRILASE WITH UNIQUE THERMOSTABILITY PROFILE AND HEAVY METAL TOLERANCE. 2013.

31. SWIFT, S.A., A.S. BOWER, AND R.W. SCHMITT, VERTICAL, HORIZONTAL, AND TEMPORAL CHANGES IN TEMPERATURE IN THE ATLANTIS II AND DISCOVERY HOT BRINE POOLS,
34. YANG, JIAN, ET AL. SALINITY SHAPES MICROBIAL DIVERSITY AND COMMUNITY STRUCTURE IN SURFACE SEDIMENTS OF THE QINGHAI-TIBETAN LAKES. SCIENTIFIC REPORTS 6, ARTICLE NUMBER: 25078 (2016)