Sequence, organization, and genes characteristics of ORFs identified in a metagenomic DNA fragment from microbial community of the deep brine environment of Atlantis II in the Red Sea.

Dina Hassan Assal

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

APA Citation

MLA Citation
Assal, Dina Hassan. Sequence, organization, and genes characteristics of ORFs identified in a metagenomic DNA fragment from microbial community of the deep brine environment of Atlantis II in the Red Sea. 2015. American University in Cairo, Master's thesis. AUC Knowledge Fountain. https://fount.aucegypt.edu/etds/1176

This Thesis is brought to you for free and open access by 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 mark.muehlhaeusler@aucegypt.edu.
School of Sciences and Engineering

Sequence, organization, and genes characteristics of ORFs identified in a metagenomic DNA fragment from microbial community of the deep brine environment of Atlantis II in the Red Sea

A Thesis Submitted to the Biotechnology Program

In partial fulfilment of the requirements for

the degree of Master’s of Science

By:

Dina Hassan Assal

Under the supervision of

Dr. Hamza El-Dorry

January/2015
The American University in Cairo

**Sequence, organization, and genes characteristics of ORFs identified in a metagenomic DNA fragment from microbial community of the deep brine environment of Atlantis II in the Red Sea**

A Thesis Submitted by

Dina Hassan Assal

To the Biotechnology Graduate Program

Month/Year

In partial fulfilment 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

I would like to dedicate this thesis to my dearest family: parents, siblings and their spouses and youngsters who were always the source of encouragement and love throughout my life.

Also I want to dedicate my work to my friends and colleagues inside and outside the AUC who supported me throughout this journey to earn my MSc. degree.

Finally, for everyone who is passionate about science and research.
ACKNOWLEDGEMENT

First, I would like to express my sincere gratitude to my advisor Prof. Dr. Hamza El-Dorry for the continuous support of my MSc. study, research and thesis preparation. Also I would like to acknowledge Dr. Mohamed Ghazy for his extensive guidance that helped me in all the time of research and writing of this thesis.

Besides my advisors, I would like to thank the rest of AUC Academics: Dr. Rania Siam, Dr. Ahmed Mustafa, Dr. Asma Amleh and Dr. Ahmed el Sayed who provided me with continuous support during studying the MSc. graduate courses.

My sincere thanks also goes to Mr. Amged Ouf for his continuous help in the lab, especially during ABI sequencing process and Mr. Mustafa Adel and Mr. Osama el-Said who helped me throughout computational data analysis.

Also I want to extend my gratitude to Ms. Mariam Azmy for her continuous help in the lab and during thesis preparation as well as Ms. Dalia Hamed and Ms. Nahla Hussein for their continuous support during thesis preparation.

I thank my fellow labmates at AUC STRC lab: to all my friends and colleagues in the STRC lab exclusively: Ayman Yehia, Mohamed Maged, Mahera Mohamed, Sara Sonbol, Noha Nagdy, Eman Rabie, Sara Al-Alawi, Hadeel EL-Bardisy for everyone who supported me in the lab and throughout my journey at AUC.

I would also like to thank KAUST for providing us with the funds necessary to do this research work.

Last but not the least; I would like to thank my family: my parents, my siblings and their spouses who supported me and encouraged me throughout my life.
# Table of Contents

**THE AMERICAN UNIVERSITY IN CAIRO**  
الجامعة الأمريكية بالقاهرة

<table>
<thead>
<tr>
<th>Contents</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>School of Sciences and Engineering</td>
<td>i</td>
</tr>
<tr>
<td>LIST OF FIGURES</td>
<td>vii</td>
</tr>
<tr>
<td>LIST OF TABLES</td>
<td>viii</td>
</tr>
<tr>
<td>LIST OF ABBREVIATIONS</td>
<td>10</td>
</tr>
<tr>
<td>LIST OF BIOINFORMATICS TOOLS</td>
<td>12</td>
</tr>
<tr>
<td>Chapter 1: Literature Review</td>
<td>13</td>
</tr>
<tr>
<td>1. Background</td>
<td>13</td>
</tr>
<tr>
<td>2. Atlantis II brine pool in the Red Sea</td>
<td>14</td>
</tr>
<tr>
<td>3. The advent of Metagenomics</td>
<td>16</td>
</tr>
<tr>
<td>4. Construction of metagenomic libraries</td>
<td>18</td>
</tr>
<tr>
<td>5. Targeted Metagenomics</td>
<td>18</td>
</tr>
<tr>
<td>5.1 Functional screening approach</td>
<td>18</td>
</tr>
<tr>
<td>5.2 Sequence based Approach</td>
<td>19</td>
</tr>
<tr>
<td>6. Sequencing technology</td>
<td>20</td>
</tr>
<tr>
<td>7. Data Analysis</td>
<td>22</td>
</tr>
<tr>
<td>8. Metagenomics’ successes in exploring microbial diversity, metabolism and industrial potential of extreme ecosystems</td>
<td>23</td>
</tr>
<tr>
<td>8.1 Assessment of microbial and metabolic diversity</td>
<td>23</td>
</tr>
<tr>
<td>8.2 Industrial Potential of Extreme Ecosystems</td>
<td>26</td>
</tr>
<tr>
<td>8.3 Therapeutics and Health care applications</td>
<td>28</td>
</tr>
<tr>
<td>9. Thesis Objective</td>
<td>30</td>
</tr>
<tr>
<td>Chapter 2: Materials and Methods</td>
<td>31</td>
</tr>
<tr>
<td>1. Sample collection</td>
<td>31</td>
</tr>
<tr>
<td>2. DNA extraction and Fosmid Library Construction</td>
<td>31</td>
</tr>
<tr>
<td>3. Pyrosequencing</td>
<td>32</td>
</tr>
<tr>
<td>4. Data Assembly and Annotation</td>
<td>32</td>
</tr>
<tr>
<td>5. Subcloning of shotgun DNA fragments of the recombinant 99B6 fosmid in bacterial vector</td>
<td>32</td>
</tr>
<tr>
<td>6. DNA Sequencing using dideoxy chain termination technique (Sanger Method)</td>
<td>33</td>
</tr>
<tr>
<td>7. Assembly of the sequenced DNA fragments</td>
<td>33</td>
</tr>
<tr>
<td>8. ORFs identification and gene annotation</td>
<td>33</td>
</tr>
<tr>
<td>9. Other descriptive features for identified ORFs</td>
<td>34</td>
</tr>
</tbody>
</table>
Chapter 3: Results and Discussion ........................................................................................................ 35
  1. Analysis of the ATII-LCL library’s Fosmids by restriction endonuclease ........................................ 35
  2. Establishing a library for fosmid clone 99B6 in pGEM plasmid vector ........................................... 36
  3. Sequencing of 99B6 pGEM library clones ....................................................................................... 37
  4. Assembly of the sequenced reads ...................................................................................................... 38
  5. Open reading frames search for potential genes and annotation of the identified ORFs ............... 42
  6. Structural characteristics of the annotated proteins .......................................................................... 45
  7. Detailed analysis of 10 selected ORFs identified by Interproscan software ..................................... 48
  8. Identified proteins with potential biotechnological applications .................................................... 60
Chapter 3: Concluding Remarks ............................................................................................................ 64
Chapter 4: References .......................................................................................................................... 65
LIST OF FIGURES

Figure 1. Location of Study Area, Atlantis II brine pool ...................................................... 15
Figure 2. Summary of experimental metagenomic processes .................................................. 17
Figure 3. Restriction analysis of six randomly selected recombinant fosmids ......................... 35
Figure 4. Isolation of 1.5 to 2 kb DNA fragments of nebulized 99B6 fosmid clone ................. 36
Figure 5. Analysis of the quality and integrity of clones from 99B6 pGEM library .................. 37
Figure 6. Assigned base quality score value for sequenced reads ......................................... 38
Figure 7. Assembly of the sequenced reads ........................................................................... 39
Figure 8. EcoRI restriction analysis of recombinant 99B6 fosmid clone ............................... 40
Figure 9. Nucleotide sequence composition of un-sequence DNA fragment and sequence end of assembled fosmid 99B6 ........................................................................ 41
Figure 10. Protein sequence analysis for identified ORFs using different protein annotation tools ...................................................................................................................... 42
Figure 11. Prediction of putative secretory signals and THMMs domains in annotated ORFs ................................................................................................................................. 47
Figure 12. Schematic representations of potential full-length ORFs in silico annotated using Interproscan protein analysis software ........................................................................... 48
Figure 13. The prediction of transmembrane domains for the putative membrane associated protease (ORF#34) was detected using the THMM tool ...................................................... 58
LIST OF TABLES

Table 1. Schematic diagram for the location and annotation of identified ORFs on contig9. ............ 43
Table 2. Protein sequence similarity search using blastx tool against nr protein database .................. 44
Table 3. Determination of halophilic potential for identified ORFs ....................................................... 46
Table 4. ORFs identified in contig9 (99B6) with potential industrial or medical applications .......... 60
Abstract

Microbial communities that reside in different natural habitats, particularly those of extreme environments, constitute a rich source for novel industrial enzymes and bioactive compounds. Until the advent of metagenomics technique, extreme environments represented a locked area with huge genetic repertoire that remained unexplored. The Atlantis II brine pool of the Red Sea (ATII) is one of such unexplored extreme environment. The lower part of this pool, the lower convective layer (LCL), has a pH of 5.3, high temperature (68°C), elevated concentration of toxic heavy metals, and extreme salinity (26% salt). To understand the metabolic and the physiological properties of proteins and enzymes that contribute to the survival of microorganisms in this extreme and hostile environment, the structure and characteristics of their genes should be determined. Metagenomics approach helped in this task through two different techniques: 1) mass sequencing of environmental DNA by high throughput sequencing technique such as pyrosequencing technique; 2) sequencing of environmental DNA fragments from metagenomic fosmid library.

The advantage of the first approach is that it produces massive number of reads that can be assembled into long contigs. Its disadvantage is that the majority of the contigs are chimeric i.e. assembled from reads belong to genomes of different microbial species. The second technique has an advantage of establishing the sequence of a contiguous piece of genomic DNA—of around 30 to 40 kb—that most probably is not a chimeric. The major disadvantages however are the high cost of the sequencing process, it involves elaborate steps, and it has a limited output of nucleotide sequences.

In this work we sequenced a contiguous fragment of DNA from the microbial community of the ATII-LCL environment and presenting the structural and potential function of its annotated genes. Interestingly, out of the 39 identified ORFs, 10 ORFs (25%) have no matches in the database. The structure and the function of the potential annotated genes are presented and discussed. In addition, we were able to assembled 28.378 kb out of 33.819 kb of the insert in the recombinant fosmid. The unassembled 5.441 kb is most probably due to the detection of characteristic patterns of low complexity regions, simple repeats as well as gene duplication exists at the end of the assembled sequence.
# List of Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>AT II Deep</td>
<td>Atlantis II deep</td>
</tr>
<tr>
<td>BAC</td>
<td>Bacterial Artificial Chromosome</td>
</tr>
<tr>
<td>BLAST</td>
<td>Basic Local Alignment Search Tool</td>
</tr>
<tr>
<td>CDD</td>
<td>Conserved domain database</td>
</tr>
<tr>
<td>COG</td>
<td>Clusters of Orthologous Groups</td>
</tr>
<tr>
<td>DNA</td>
<td>Deoxyribonucleic acid</td>
</tr>
<tr>
<td>GOS</td>
<td>Global Ocean Sampling</td>
</tr>
<tr>
<td>HGT</td>
<td>Horizontal Gene Transfer</td>
</tr>
<tr>
<td>HMM</td>
<td>Hidden Markov Model</td>
</tr>
<tr>
<td>IPTG</td>
<td>Isopropyl-beta-D-thiogalactoside</td>
</tr>
<tr>
<td>KAUST</td>
<td>King Abdullah University of Science and Technology</td>
</tr>
<tr>
<td>KEGG</td>
<td>Kyoto Encyclopedia of Genes and Genomes</td>
</tr>
<tr>
<td>LCL</td>
<td>Lower Convective Layer</td>
</tr>
<tr>
<td>LB</td>
<td>Luria-Bertani</td>
</tr>
<tr>
<td>NCBI</td>
<td>National Centre for Biotechnology Information</td>
</tr>
<tr>
<td>NGS</td>
<td>Next generation sequencing</td>
</tr>
<tr>
<td>ORF</td>
<td>Open Reading Frame</td>
</tr>
<tr>
<td>PCR</td>
<td>Polymerase Chain Reaction</td>
</tr>
<tr>
<td>PSORT</td>
<td>Protein Subcellular Localization Prediction Tool</td>
</tr>
<tr>
<td>PFAM</td>
<td>Protein Family</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Description</td>
</tr>
<tr>
<td>--------------</td>
<td>-------------</td>
</tr>
<tr>
<td>Psu</td>
<td>Practical Salinity Unit</td>
</tr>
<tr>
<td>rRNA</td>
<td>Ribosomal Ribonucleic Acid</td>
</tr>
<tr>
<td>SSU</td>
<td>Small Subunit</td>
</tr>
<tr>
<td>SB</td>
<td>Subtilisin</td>
</tr>
<tr>
<td>THMM</td>
<td>Transmembrane Hidden Markov Model</td>
</tr>
<tr>
<td>WGS</td>
<td>Whole Genome Shotgun Sequencing</td>
</tr>
<tr>
<td>Web MGA</td>
<td>Web Metagenomic Annotator</td>
</tr>
<tr>
<td>X-Gal</td>
<td>5-bromo-4-chloro-indolyl-β-D-galactopyranoside</td>
</tr>
</tbody>
</table>
# List of Bioinformatics Tools

<table>
<thead>
<tr>
<th>Bioinformatics tool</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>Artemis</td>
<td><a href="https://www.sanger.ac.uk/resources/software/artemis/">https://www.sanger.ac.uk/resources/software/artemis/</a></td>
</tr>
<tr>
<td>Interproscan</td>
<td><a href="www.ebi.ac.uk/Tools/pfa/interproscan/">www.ebi.ac.uk/Tools/pfa/interproscan/</a></td>
</tr>
<tr>
<td>SIGNALP</td>
<td><a href="www.cbs.dtu.dk/services/SignalP">www.cbs.dtu.dk/services/SignalP</a></td>
</tr>
<tr>
<td>Phred/Phrap/consed</td>
<td><a href="www.phrap.org/phredphrap/consed.html">www.phrap.org/phredphrap/consed.html</a></td>
</tr>
<tr>
<td>PFAM</td>
<td><a href="pfam.sanger.ac.uk/">pfam.sanger.ac.uk/</a></td>
</tr>
<tr>
<td>PRED-TAT</td>
<td><a href="www.compgen.org/tools/PRED-TAT">www.compgen.org/tools/PRED-TAT</a></td>
</tr>
<tr>
<td>PSORT</td>
<td><a href="psort.hgc.jp">psort.hgc.jp</a></td>
</tr>
<tr>
<td>THMM</td>
<td><a href="www.cbs.dtu.dk/services/TMHMM">www.cbs.dtu.dk/services/TMHMM</a></td>
</tr>
<tr>
<td>WebMGA</td>
<td><a href="http://weizhong-lab.ucsd.edu/metagenomic-analysis/server">http://weizhong-lab.ucsd.edu/metagenomic-analysis/server</a></td>
</tr>
</tbody>
</table>
Chapter 1: Literature Review

1. Background

The total number of prokaryotic cells in the different habitats on Earth is estimated to be around $4 \times 10^{30}$. They are pivotal as primary producers of essential nutrients and also in salvage and recycling of organic compounds (Whitman, Coleman, & Wiebe, 1998; Wooley, Godzik, & Friedberg, 2010). The majority of these microorganisms are not studied yet and most probably they possess unknown and diverse metabolic and also novel biocatalysts and biomolecules (Schirmer, Gadkari, Reeves, Ibrahim, DeLong, et al., 2005). Their significance is not only recognized as being the primary producers of nutrients and recyclers of organic matter to different carbon sources, but also as a repertoire with relatively unexplored functional and metabolic diversity (Simon & Daniel, 2011). Currently, many microbial derived compounds are considered potentially valuable for multiple industrial disciplines, for example: biocatalysts, antibiotics, anticancer compounds and biofuels are on continuous demand by different biotechnological and pharmaceutical sectors (Zhang & Kim, 2010). However, the potential for 99% of these microbes are remained untapped by traditional genomic techniques that requires prior isolation and culturing of the organism for microbial genome sequencing and analysis (Kennedy, Marchesi, & Dobson, 2008; Schloss & Handelsman, 2003).
2. **Atlantis II brine pool in the Red Sea**

The Red Sea is one of the most distinctive marine environments among others, yet it is considered to be the least studied and explored. The Red Sea was formed as a result of the separation of the Arabian and Asian plates 3-4 million years ago. Its water is considered to be extremely unique due to its high temperature and salinity, which are caused by the increased evaporation rate and low rainfall rate (Bougouffa et al., 2013; Gurvich G. E., 2006). The Red Sea contains at least 25 deep brine pools scattered along its central axial rift (Siam et al., 2012). These brine pools were formed as a result of water perfusion into evaporitic deposits in the shallow layers of the sea floor, thus making water in brine pools denser than seawater. In some of the brine pools, the water is heated by tectonic movement underneath and finally returns to the brine pools by the current flow effect (Noguchi, Taniguchi, & Itoh, 2008).

The largest hyper saline basin, the Atlantis II deep (ATII), is considered one of the intriguing environments located in the central region of the Red Sea (latitude 21°23′ N and longitude 38°04′ E) (Figure 1) at a depth of 2194 m below the sea surface. The ATII brine pool is further stratified into several distinct layers: brine interface, upper convective layer, intermediate and lower convective layer. The layers are characterized by the gradual rise in both salinity and temperature towards deep sediments. Hence, this brine basin is uniquely known to have the highest temperature, be most dynamic, and have the largest “ore – forming” body (Antunes, Ngugi, & Stingl, 2011; Mohamed et al., 2013). The lowest part of the brine pool, the lower convective layer (LCL), has a pH value of 5.3, a temperature of 68 °C and salinity of 25.7% which is 7.5 times more than the salinity of the surface of normal seawater (Antunes et al., 2011).

The Atlantis II deep LCL (ATII-LCL) is an anoxic environment, where dissolved N₂ gas and methane CH₄ are available in high levels with traces of CO₂, ethane and H₂S. The richness of the ATII-LCL sediment with metalliferous sediments, high levels of iron, copper and other toxic heavy metals, as well as its unique geochemical parameters, have made it an attractive site for many geological and geophysical studies (Antunes et al., 2011; Bougouffa et al., 2013). However, little interest was shown from microbiologists in this environment mainly because in the late 1960s it was believed that sustainability of any form of life is not possible under these extreme conditions and thus Atlantis II deep brine was considered to be sterile (Antunes et al., 2011; M. Schmidt et al., 2003). Lately, other findings supported the presence of distinctive microbial consortia that can thrive in these exotic environments. The
extremophilic microorganisms that thrive in ATII-LCL environment should therefore possess an extraordinary biochemical and molecular machinery that is need to survive in such extreme setting.

Figure 1. Location of Study Area, Atlantis II brine pool.

The figure represents the location of Atlantis II brine pool in the Red Sea; the study area is located at latitude 21°23’ N and longitude 38°04’ E. Picture adapted from http://www.emr.rwthaachen.de/aw/EMR/Energy_Mineral_Resources/Zielgruppen/tml/publikationen/~wcw/A_Theoretical_Analysis_of_Hydrothermal_F.
3. **The advent of metagenomics**

Metagenomics represents an alternative to traditional genomics approaches. In this approach, sequencing of total microbial genomes isolated from natural habitat (environmental DNA) enables us to explore the biochemical and physiological process that support growth of microbes in its natural environment. This technique permits to develop an overall view of these processes – biochemical and physiological – without prior need for cultivation. Accordingly, this would presumably accelerate the rate of novel gene discovery from different microbial ecosystems (D.-G. Lee et al., 2007; Uchiyama & Miyazaki, 2009; Verma, Kawarabayasi, Miyazaki, & Satyanarayana, 2013).

The use of metagenomics was first proposed by Pace on 1985 and his colleagues to gain insight into the metabolic and physiological properties of uncultured microbial communities in different environments (Handelsman, 2004; Jeon, Kim, Kang, Lee, & Kim, 2009). Later, many metagenomic projects were conducted such as the Global Ocean Sampling and the acid mine drainage project (Rusch et al., 2007; Tyson et al., 2004). Their primary goal was to investigate questions associated with organisms (who is there) and their metabolic and physiological processes in different complex environments (what are they doing) (Eisen, 2007). As summarized in Figure 2, several methods were taken to tackle these questions in any metagenomic project. In general, the workflow for most metagenomic approaches would either go through constructing, screening and sequencing of metagenomic clones of interest, or direct sequencing of environmental DNA that was facilitated by advances made in next generation sequencing approaches (NGS).
In brief, the experimental design for any metagenomic process begins by the isolation of microbial communities from environmental habitats under investigation, followed by the extraction of total microbial DNA samples. Depending on the objective of the metagenomic projects and questions that being addressed, the following approaches can be used:

1. Targeted metagenomics: the extracted DNA is randomly sheared and cloned into vectors to construct metagenomic libraries and clones are then subjected to either sequence based, or functional based screening analysis.

2. Phylogenetic analysis: assesses microbial diversity of environmental niche using set of primers for conserved genes as 16S rRNA and Rad genes.

3. Random sequencing of environmental DNA to gain insight on the metabolic and physiological processes of microbial community:
   a. Sanger sequencing of metagenomic clones
   b. NGS based metagenomic analysis.

4. Whole (Meta) genome assembly (WGA) of microbial genomes for sequences obtained from Sanger or NGS sequencing processes.
4. **Construction of metagenomic libraries**

In general, the construction of metagenomic libraries and screening processes begins by microbial collection from the environmental niche of interest, such as water, soil, air, and human gut (Culligan, Sleator, Marchesi, & Hill, 2013). Environmental DNA is then isolated from the microbial samples and metagenomic library are constructed in a suitable vectors. Nowadays, many vectors are available for use with capacity to accommodate different sizes of inserts, ranging from smaller size plasmids (accommodate less than 15 kbp) to, moderate size fosmids (accommodate 40 kbp) and bacterial artificial chromosomes (BAC) (>40 kbp). The recombinant vectors are then transformed into appropriate surrogate hosts, usually *E. coli* or others host cell that are appropriate for heterologous gene expression. Once a metagenomic library is created, successive experimental methods and metagenomic approaches are used to explore functional analysis of specific gene of interest, taxonomic analysis of the microbial community, or structural studies of the microbial genomes (Daniel, 2005; Rusch et al., 2007; Simon & Daniel, 2011).

5. **Targeted metagenomics**

5.1 **Functional screening approach**

The functional based analysis relies on the detection of active clones that express characteristic traits of interest in the surrogate host. This approach enables the discovery of novel genes without prior need for sequence information and guarantees the retrieval of both the sequence and activity of isolated genes (Li, McCorkle, Monchy, Taghavi, & van der Lelie, 2009). Several microbial enzymes and biomolecules were successfully identified and isolated using functional screens of metagenomic clones (Jeon, Kim, Kang, et al., 2009; J. Kennedy et al., 2008; Sayed et al., 2014). The limitations of this approach are mainly represented by problems of gene expression in the surrogate host due to gene toxicity, inappropriate selection of the host strain required for heterologous gene expression, or that the promoter of the metagenome gene is not functioning in the host cell (Felczykowska, Bloch, Nejman-Faleńczyk, & Barańska, 2012). The frequency of active clones identified from massive screening of metagenomic libraries is considerably low, it’s not always possible to scale up the functional assay of interest on large libraries and the presence of all genetic elements required for proper gene expression is mandatory (Schloss & Handelsman, 2003; Sharma, Khan, & Qazshari, 2010).
Several attempts were made to address these limitations such as the use of different strains that are either adaptable to extreme conditions or genetically modified strains as those commonly used *E. coli* strains shuttle vector systems to extend their capabilities for gene expression and overcome gene toxicity problems (Kennedy et al., 2011; Li et al., 2009; Lorenz & Eck, 2005; J. Singh et al., 2009).

### 5.2 Sequence based approach
As for the sequence based approach, either a set of PCR primers or hybridization probes are used to identify novel genes, which represent a new variant of already known genes or protein families (Culligan et al., 2013), or using the phylogenetic based analysis where markers of different phylogenetic anchors as 16S ribosomal RNA or rec A sequences link the discovery of novel genes to phylogenetic analysis (Handelsman, 2004; Schloss & Handelsman, 2003; J. Singh et al., 2009). The former approach showed to be successful in identifying new homologs of enzymes and proteins with wide applications in various industries such as: oxidoreductases, chitinases (Simon & Daniel, 2011), mercuric reductases (Sayed et al., 2014), and amylases (Sharma et al., 2010). In addition, the isolation of several polyketide synthases and antibiotics were also reported from different microbial niches (Banik & Brady, 2010; Kerkhof & Goodman, 2009; Schirmer, Gadkari, Reeves, Ibrahim, Delong, et al., 2005).

Regarding phylogenetic marker based sequence analysis, this strategy was frequently used for the discovery of new genes from the sequencing of clones harboring phylogenetic anchors. One of the famous findings was the discovery of the photorodopsin gene which was affiliated to gama-proteobacteria after it was presumed to be only affiliated to archeal lineages (Handelsman, 2004). This enables us to have an insight into ecology and taxonomy of other bacteria in complex environments that have never been cultured (Culligan et al., 2013; Daniel, 2005; Kerkhof & Goodman, 2009). Alternatively, the random sequencing of the metagenomic clones strategy was conducted by large metagenomic sequencing projects such as the Sargasso Sea and acid mine drainage projects, as well as many others that revealed huge inference about microbial genome assemblage, horizontal gene transfer and linkage of traits in addition to other insights that were gained from the studies of uncultured bacterial communities (Handelsman, 2004).
6. **Sequencing technology**

While the arrival of next generation sequencing technologies have revolutionized the field of metagenomics (Dark, 2013; Teeling & Glockner, 2012), the Sanger sequencing was the first technology applied for metagenome sequencing (Hoff, 2009). The Sanger method is based on a chain termination reaction that involves the termination of DNA synthesis using a chemically modified 2, 3-dideoxynucleotides, leading to DNA fragments of varying lengths separated on electrophoresis gel where DNA sequence of the template is determined. Successive advances were made in sequencing technologies using this approach over time as the release of automated Sanger platforms introduced some modifications to old principles by using fluorescent tagged ddNTPs. The fluorophore of ddNTPs are then excited by the laser beam of the sequencer, resulting in the emission of four different colors. The base calling step takes place based on tracking fluorescence signals of the four different emitted colors and quality assessment algorithms are deployed to assess correct base calling process of DNA sequencing (Metzker, 2005; Shendure & Ji, 2008).

The long read length (700 bp or more) obtained, high sequencing accuracy with decreased error range (0.001-1%), and the ability to sequence clones with large insert size (fosmids and BACS) are among the advantages that made Sanger sequencing applicable for whole genome sequencing of low diversity microbial niches. The inherent requirement for metagenomic propagation by vectors into bacterial hosts will probably lead to bias against genes that are toxic for the surrogate hosts, in addition to its labor demanding and exhaustive nature. It is also worth noting that the overall cost of sequence data per Gbp is about USD 400,000 (Thomas, Gilbert, & Meyer, 2012).

The paradigm shift from classical Sanger sequencing to next generation sequencing technologies is triggered by the need to lower sequencing costs and generate high throughput of sequence data during any metagenomic study (Danhorn, Young, & DeLong, 2012). The 454/ Roche and the Illumina/ Solexa systems are widely used NGS technologies for DNA sequencing to date (Thomas et al., 2012). For the 454/Roche technology, emulsion polymerase chain reaction (ePCR) is applied to “clonally amplified” random DNA fragments of metagenome, where the single stranded DNAs are attached to microscopic beads, then deposited into a picotide plate for further parallel and individual pyrosequencing (Sulsultana & Neelakanta, 2013).
The pyrosequencing process utilizes a “sequencing by synthesis approach” that involves the successive addition of all deoxynucleotides and the incorporation is monitored by conversion of the released pyrophosphate (PPi) into emitted light which is further detected by a CDD camera and converted to sequence data (Dark, 2013).

As for any emerging NGS technologies, there are a couple of limitations associated with 454 pyrosequencing especially when applied to metagenomics. First, the emulsion based PCR method creates artificial replicates of sequencing data that would interfere with estimation of gene abundance; nonetheless those artificial replicates can be detected and filtered out with bioinformatics programs. Secondly, high sequencing error rates are reported as a result of sequencing errors associated with existing homopolymers that might cause deletion or insertion in predicted open reading frames (ORF). However, this could also be tackled by ORF prediction tools to correct these frame shifts. In spite of these disadvantages, the low sequencing costs produced by 454 pyrosequencing (20,000 per Gbp), massive sequencing capacity of ~ 500 Mb per run (multiplexing) along with less man power needed, yet have made it the technology of choice for many metagenomic and other applications (Kim et al., 2013; Lin Liu et al., 2012; Thomas et al., 2012).

Also, one of the major shortcomings associated with other NGS platforms is the shorter read length produced while sequencing microbial genomes. Lately, the advanced version of 454 pyrosequencing technology (GS-FLX Titanium) can produce reasonably longer read-lengths (400-500 bp) compared to the competing NGS platforms (Kim et al., 2013). Moreover, Illumina, and to a lesser extent, SOLiD technologies were found to be applicable for large scale metagenomic surveys owing to their rapidness, huge sequencing data, and more coverage obtained at the same cost when compared to 454 pyrosequencing. It was also noticed in previous studies that Illumina platforms showed comparable results to 454 pyrosequencing in terms of assemblies and high sequencing coverage when studying metabolic and taxonomic diversity of microbial ecosystems (Teeling & Glockner, 2012).

Recently it has been discovered that the utility of the hybrid sequencing approach using data from two or more sequencing technologies represents a promising alternative for many whole microbial genome assembly projects. The primary goal is to obtain cost effective and high quality draft or finished assemblies of microbial genomes. For instance, Craig Vendor and his colleagues investigated the strengths and weaknesses associated with the use of the hybrid
sequencing approach to evaluate assemblies of six microbial genomes against assembly from Sanger sequencing only. It was demonstrated that the incorporation of 454 data into whole genome assembly resulted in the completion of 2 out of 6 microbial genomes and approximately 86% of gaps were reduced for other microbial genomes. The advantage of 454 sequencing over Sanger is represented by no clonal bias and the ability to sequence non-clonal regions. Although the 454 sequencing platform was considered a promising alternative for cost effective and high quality assemblies, it was not recommended for use in de novo sequencing due to the lack of paired end information and ambiguous incorporation of repeats by 454 data. The conclusion of the investigation was that de novo genome sequencing projects should depend on the Sanger platform to make use of pair-end information and deal with large repeats and physical gaps, and that 454 data usage would work best to complement Sanger sequencing for “hard stop” non-clonal regions and gap closure projects (Goldberg et al., 2006).

Despite the progress made in lowering sequence costs, the data analysis phase is still a computational extensive process (L. D. Stein, 2010; Wooley et al., 2010). The following challenges are still considered when analyses of huge metagenomic datasets take place: first, a large storage capacity is needed for data storage and second, the prerequisite for data standardization prior to the analysis and presence of advanced computing resources is required. In addition, the short read length generated results with a higher rate of sequencing errors and difficulties with data assembly. Accordingly, this would presumably limit insights into gene content and correct functional assignment of metagenomes when compared to results obtained by Sanger technology (Kim et al., 2013).

7. **Data analysis**

The emerging NGS technologies are currently introduced in many research aspects including metagenome sequencing. Albeit differences in sequencing chemistries, data output and characteristics of each platform in terms of advantages and limitations, they all share the potential to generate tremendously high throughput sequencing data. Thus, the downstream analysis for these data output is a very complex process in terms of time consumption and lack of bioinformatics expertise that still represent a real challenge for many biologists and researchers of different research facilities (Barzon, Lavezzo, Militello, Toppo, & Palù, 2011). Today, many computational tools are available to analyze sequence data of different microbial genomes and some of them have shown applicable for metagenomic data analysis.
of mostly small scale projects, however additional bioinformatics pipelines with higher computing power and data storage capacity are still highly required, so that sequencing data can be fully exploited and problems associated with metagenome sequencing can be handled (Barzon et al., 2011; Teeling & Glockner, 2012).

In general, a typical framework for metagenomic data analysis to provide insights into functional and taxonomic diversity of microbes might involve data cleaning and processing, assembly, gene calling, taxonomic classification and finally gene annotation or phylogenetic inference. Further data analysis could be achieved on the transcription or translation level as well (Teeling & Glockner, 2012; Wooley et al., 2010).

8. **Metagenomics’ successes in exploring microbial diversity, metabolism and industrial potential of extreme ecosystems**

Metagenomics represent a powerful approach to analyze microbial habitats from different extreme environments in terms of microbial diversity, metabolism and adaptation mechanisms to different extreme conditions as well as comparative studies (Lewin, Wentzel, & Valla, 2013; Wemheuer, Taube, Akyol, Wemheuer, & Daniel, 2013).

Several approaches are used for the application of microbial shotgun metagenomics for current research including the construction of metagenomic libraries in suitable vectors as small size vectors plasmids or large size fosmids and BACs followed by random or targeted sequencing of one or more metagenomic clones. Furthermore, direct sequencing of environmental DNA is carried out using NGS platforms to gain more insights into the microbial genome structure and gene content, as well as microbial genome assembly from assembled metagenomic data (J. a Gilbert & Dupont, 2011).

8.1 **Assessment of microbial and metabolic diversity**

Scientists have initially focused to characterize poorly studied uncultured microbes from different environmental habitat. In order to identify novel microbial lineages or gain more insights into their metabolic functions, this was accomplished by sequencing part of their genomes. This metagenomic approach is presumably have helped us to identify novel microbial lineages and contributed in landmark discoveries from uncultured microbes. For instance, two independent studies were conducted by Beja et.al., and Stein’s group that have led to the discovery of bacterial derived rhodopsin and the first identification of 16S rRNA linked genomic fragment of uncultured marine planktonic archaeon respectively (Béjà et al.,
Interestingly, sequencing of 43kb genomic fragment of uncultured crenarchaeota from the fosmid library of calcareous grassland soil have revealed the functional role of ammonia oxidizing bacteria in different marine habitats using both shotgun metagenomics and inverse transcription PCR based analysis (Treusch et al., 2005).

Another remarkable study by Gilbert et al, have reported the shotgun sequencing of a 32,86 kb fosmid clone from the metagenomic library of a coastal marine microbial habitats. The phylogenetic analysis has revealed high sequence similarity to SAR11 bacteria “C. Pelagibacter unique”. However the results from shotgun metagenome analysis have showed that almost half of putative genes identified in the sequenced fosmid clone were very less similar to the genome of the most closely SAR11 bacteria, unlike the results demonstrated by rRNA based analysis. These results have demonstrated that the shotgun metagenomic approach was superior over 16S based phylogeny to explain the genetic versatility and diversity of ubiquitous C. Pelagibacter across many habitats (J. A. Gilbert, Mühling, & Joint, 2008). More studies were conducted using shotgun metagenomic clones/s sequencing approach as reviewed elsewhere in (Nesbø, Boucher, Dlutek, & Doolittle, 2005; Schirmer, Gadkari, Reeves, Ibrahim, DeLong, et al., 2005; Vergin et al., 1998; G. Y. Wang et al., 2000), as it is estimated that shotgun sequencing of large genomic fragments of uncultured microbes from one or more metagenomic clones will not only participate in the identification of novel microbial lineages, but also can produce large contiguous DNA sequences that provides more insights into gene neighborhood context and identify novel genes variants and biosynthetic clusters with biotechnological prospects that are usually unrecognized during the analysis of short read metagenomic datasets generated by high throughput sequencing technologies (Danhorn et al., 2012; Kunin, Copeland, Lapidus, Mavromatis, & Hugenholtz, 2008).

Lately, an increasing number of shotgun metagenomic studies describing different extreme environments have been reported. The study of Acid mine drainage (AMD) biofilm of Iron Mountain in California represents one of the early metagenomic studies conducted on low complexity microbial community from extreme environments. The sequence data generated by random shotgun sequencing using Sanger technology (76 Mbp) showed dominance by two microbial subgroups; Leptospirillum species and archaeon Ferroplasma acidarmanus. Interestingly, it was the first attempt to determine nearly completes genome from uncultured dominant species, unlocking metabolic pathways for nitrogen and carbon fixation, energy
production and other adaptation mechanisms to extreme environments. Thus, this study provided comprehensive insight into the microbial community and the metabolic structure, as well as adaptation strategies for extreme habitats and this was further regarded as a milestone for subsequent analysis of other extreme acidic environments (Tyson et al., 2004).

The hydrothermal vents represent one of the exotic marine ecosystems from which, only few metagenomic studies were reported to date. For example, Brazelton and his colleagues reported the isolation of DNA from microbial communities of hydrothermal chimney biofilm samples of the lost city hydrothermal field on the Mid-Atlantic Ridge (Brazelton & Baross, 2009). The earlier phylogenetic analysis of chimney microbial community revealed that metagenome of biofilm samples was dominated by methane cycling archaea *Methanosarcinales*, which accounted for more than 80% of the total microbial structure. The pooled metagenome was then randomly sheared, cloned and sequenced by Sanger technology, yielding 46,316 reads of total sequence data size 35 Mbp. Sequence analyses of shotgun reads against nr database using BLAST sequence similarity search tool revealed that more than 8% of the data is represented by transposases, showing a significant abundance and diversity of transposases compared to metagenome of other ecosystems. The identified transposases showed high coverage but yet found small in size, suggesting their location on small extragenomic molecules as phages or plasmids and thus, these mobile elements were thought to contribute in the phenotypic variety of low-complex microbial communities. Further comparative analysis for other deep-sea hydrothermal chimneys by metagenomics have shown common diversity and functions that are selective for chimney biofilms regardless the geochemical differences of studied ecosystems (Sievert & Vetriani, 2012).

The work of Simon and his colleagues represent one of the first attempts to explore microbial diversity and metabolic functions of the Northern Schneeferner (Germany) glacial ice metagenome using 454 pyrosequencing data. The metagenomic analysis of glacial ice DNA revealed potential metabolic versatility of microbial community displaying autotrophic manner in response to the low nutrient content of glacial ice (Lewin et al., 2013). Also, several cold adaptation mechanisms were also determined as cryoprotectants, unsaturated fatty acids and ROS scavengers. Hence, the metagenomic analysis of glacial ice samples provided comprehensive insights into the microbial structure and metabolic functions of psychrophilic community as well as potential adaptations to cold stress on Earth and a
glimpse onto microbial life of other similar extraterrestrial habitats (Simon, Wiezer, Strittmatter, & Daniel, 2009).

8.2 Industrial potential of extreme ecosystems

Microbial enzymes and bioactive molecules from different microbial isolates have been extensively used for many industrial applications to date (Zhang & Kim, 2010). The Global initiatives towards the application of white biotechnology have shifted the paradigm towards novel microbial resources which hold great potential for industries. The advent of metagenomics along with in vitro evolution and high-throughput screening technologies, hold great promises towards microbial enzymes commercialization. Several attempts were made by many biotechnological companies to isolate novel biocatalysts and valuable biomolecules using metagenomics (Lorenz & Eck, 2005). The release of novel nitrilase, phytase and glycosidase enzymes into industry by Diversa company represent one of the success stories accomplished by metagenomics for novel biocatalysts discovery (Urban & Adamczak, 2008).

Recent studies have employed different metagenomic approaches for the isolation of novel microbial enzymes and molecules from different extreme ecosystems. In regard to lipolytic and polysaccharide modifying biocatalysts, Jeon et al. have reported the isolation of a novel low temperature active lipase enzyme from cold sea sediment. One positive recombinant clone was detected with lipolytic activity from the metagenomic library on tricaprylin medium, the clone of interest was then fully sequenced using the shotgun metagenomic approach. Sequence analysis and phylogenetic studies revealed sequence similarity of the sub clone (ORF20) to a new lipase / esterase family. Upon further characterization, the expressed protein displayed lipolytic activity over a range of natural oil substrates and it retained more than 50% activity at 5 degrees and resistance to different detergents. The results described a cold active lipase with interesting features for industry (Jeon, Kim, Kim, et al., 2009). Later, the same co-workers have identified two novel esterases from another extreme low temperature environment, Arctic metagenome samples. These enzymes are potentially used in chiral resolution of heat sensitive substrates (Jeon, Kim, Kang, et al., 2009).

Another example illustrates the isolation of a novel thermostable amylase from screening metagenonomic fosmid library of the Juan De Fuca Ridge Hydrothermal Vent; this novel alpha amylase is a new member of glycosyl hydrolase family 57 and the 3D structural dimensions of the protein were predicted by homology modelling. The enzyme showed maximum activity at pH of 7.5 and temperature of 90 degrees Celsius (°C) with capacity
to retain more than 50% of its activity at this temperature for 4 hours (H. Wang et al., 2011). More investigation into extreme cold and hot ecosystems would presumably reveal a plethora of novel microbial enzymes and molecules that shows maximum activity at extreme temperatures/thermal conditions.

Furthermore, metagenomics has been successfully employed to explore enzyme potential with other desirable characteristics from different extreme habitats as well. As for example, the thermo-alkali-stable xylanase encoding gene was identified by Verma and her co-workers by shotgun sequencing of metagenomic subclone of composite soil samples. The recombinant enzyme displayed activity over a wide range of pH and with optima activity at pH 9.0 and temperature at 80 degrees Celsius (°C). This study represent one of the first metagenomic attempts to screen for xylanases that tolerate harsh conditions predominant in many industrial processes, as required by feed as well as paper and pulp industries (Verma et al., 2013). Another novel thermostable pectinase was identified by Singh and his collaborators from screening soil metagenome from hot springs of northern India. The gene encoding for pectinase was expressed into E.coli strain and characterized over a broad range of temperature and pH. The enzyme was found to be thermostable at temperature of 60 degrees Celsius (°C) and activity over wide pH range, reaching temperature and pH optima at 70 degrees Celsius (°C) and 7.0 respectively which make it applicable for different industries as waste water treatment and textile processing (R. Singh, Dhawan, Singh, & Kaur, 2012).

As for proteases, they are widely used in different industrial applications and represent more than 60% of sales of the global enzyme industry (Zhang & Kim, 2010). The metagenomic survey for novel environmental niches as Gobi and the Death Valley deserts, resulted in the detection of 17 putative proteolytic clones, two proteases belong to the subtilisin (S8A) family and were further purified and characterized displaying different biochemical activities in terms of pH and temperature. Protease M30 revealed optimum activity at alkaline pH >11 and temperature of 40°C, whereas protease DV1 showed an optimum activity at pH of 8 and temperature of 55°C, and thus these unique characteristics make them suitable for biotechnological uses (Neveu, Regeard, & DuBow, 2011).
Interestingly, Mohamed et al., have reported the isolation and characterization of a thermostable, metal-resistant esterase from most exotic layer of the Atlantis II deep in the Red Sea, the lowest convective layer of the brine pool (Mohamed et al., 2013). The above-mentioned features make them a potential biocatalysts for industry. Recently, the richness of the Atlantis II deep with heavy metals have intrigued further interests to explore an Atlantis II metagenomic dataset for novel heavy metal detoxifying enzymes using “synthetic metagenomics” approach (Culligan et al., 2013). The work revealed the first novel thermostable, halo tolerant and metal resistant mercuric reductase (Sayyed et al., 2014).

8.3 Therapeutics and health care applications

Today, many microbial enzymes have been successfully employed in many clinical and healthcare applications. For example, many thermostable superoxide dismutases (SOD) are highly required for different cosmetics preparations. A study by He and his colleagues reported the isolation of a novel thermostable Fe-superoxide dismutase enzyme from screening metagenomic libraries constructed from hot spring sample. The identified gene was expressed into E. coli and characterized by pyrogallol method as an iron dependant superoxide dismutase (Fe-SOD) enzyme. The enzyme displayed optima temperature at 80 degrees and stability at a broad pH range 4-11. Comparative modelling results revealed that the presence of a large number of inter-subunit ion pairs and hydrogen bonds and decreased solvent accessible hydrophobic surfaces accounted for its high thermostablity (He et al., 2007). The microbial derived production of melanin pigments represents a hot topic for R&D scientists in cosmetics industry as an inexpensive alternative for producing melanin (Gabani & Singh, 2013). The discovery of a melanin producing clone from screening metagenomic library of deep sea sediment represents one of the accomplishments made from exploring untapped microbial repertoire of extreme environments (Huang et al., 2009).

In highlight to the potential use of proteases in clinical applications, some metalloproteases were identified as potential thrombolytic agents for treatment and prevention of cardiovascular diseases. As the use of current thrombolytic agents is associated with undesirable side effects and lower specify to fibrin, thus mining for novel thrombolytic enzymes from extreme environments as deep sea and other extreme environments that represent a unique microbial repertoire for mining novel fibrinolytic enzymes for treatment of thrombosis. Lee et al. described the isolation of a novel zinc-dependant metalloprotease from deep sea sediments of clam bed metagenome in the west coast of Korea. Primary screening of
metagenomic libraries showed one metagenomic clone with proteolytic activity on azocasein substrate. The purified enzyme was further characterized displaying optimal activity at 50 °C for 1 hour and pH 7.0. The functional assay was performed using azocasein and fibrin as substrates and enzyme activity was further inhibited by different metal chelating agents (D.-G. Lee et al., 2007).

Also, the discovery of novel antibiotics and anticancer compounds has gained special interest by metagenomics experts. Several bioactive compounds were successfully isolated from temperate environmental niches by metagenomics (Gillespie et al., 2002; E. W. Schmidt et al., 2005; G. Y. Wang et al., 2000), however still little is known about extreme environments as a potential source for novel therapeutics for treatment of cancer and other life threatening diseases. Recently, three recent independent metagenomics studies of desert sand soil have reported the identification of anticancer agent indolotryptoline and two polyketide type II synthases derived molecules; fluostatins and erdacin respectively. The study of desert soil metagenome represents an interesting example for extreme ecosystems that hold promises towards novel bioactive molecules discovery (Chang & Brady, 2013; Feng, Kim, & Brady, 2011; King, Bauer, & Brady, 2010 reviewed in M. H. Lee & Lee, 2013).

The contribution of microbes to pharmaceutical and health care sector goes back to the early discovery of antibiotics by Lewis Pasteur and later from other microbial isolates. The emerging resistance against current antibiotics and the increasing potential of microbial enzymes in clinical applications (Zhang & Kim, 2010), have provoked the need to fully explore novel microbial habitats for novel therapeutics and bioactive compounds using successful approaches as the use of metagenomic approach.

These findings are supporting evidence that a unique genetic and microbial repertoire are showing high endurance to extreme conditions of different extreme ecosystems, and thus necessitate further exploration by metagenomics.
9. **Thesis objective**

Several metagenomic approaches are currently used to find novel microbial products in different environmental niches. However, despite the extensive screening efforts made for novel genes discovery, most of the targeted metagenomic strategies that involve screening of full libraries containing thousands of clones have usually resulted in few to zero positive hits, leaving behind a plethora of metagenomic clones (negative hits) that could be useful for other applications. As a consequence, the discovery rate and application of novel molecules has remarkably decreased (Mária Džunková, Giuseppe D’Auria, David Pérez-Villarroya, 2012). In addition, the genes most distantly related to the consensus for a gene family are most probably of great interest for biotechnological applications. However, they are the least likely to be detected by probes or primers designed based on the database entries. Thus, the undirected approach is likely to unmask a homologues set of useful genes that would not have been detected by directed metagenomic studies. Moreover, even with huge metagenomic data generated from emerging NGS technologies, the functional potential of a large number of generated sequences are yet remained unexplored. This is due to some pitfalls in assembly and scaffolding strategies of NGS data as a result of short read length and lack of mate-paired sequence information (Jackson, Rounsley, & Purugganan, 2006). To address these problems, and to establish in our laboratory the processes to obtain full length, contiguous, and un-chimeric assembled metagenomic sequences, the Sanger chain termination technique was used to sequenced the insert of one recombinant fosmid clone from an LCL-ATII metagenomic library. Therefore, the objective of this work is to determine the sequence, gene organization and to characterize identified ORFs of long de-novo contiguous metagenomic DNA fragment from the deep brine environment of Atlantis II in the Red Sea.
Chapter 2: Materials and Methods

1. **Sample collection**
Samples were collected by the AUC research team from the LCL of Atlantis II brine pool (at 21° 20.72 North and 38° 04.59 East) during the King Abdullah University for Sciences and Technology (KAUST) (Thuwal, Kingdom of Saudi Arabia); Woods Hole Oceanographic Institute (WHOI) (Woods Hole, MA); and Hellenic Center for Marine Research (HCMR) (Anavisso, Greece) oceanographic cruise of the research vessel Aegaeo in March/April 2010. Microbial communities were collected by serial filtration of the water samples through 3 µm, 0.8 µm and 0.1 µm of mixed cellulose-ester membrane filters (nitrocellulose/cellulose acetate). The filters were then stored in sucrose buffer at -20 °C (Mohamed et al., 2013).

2. **DNA extraction and fosmid library construction**
Isolation of environmental DNAs was performed as described (Rusch et al., 2007). Fosmid DNAs were extracted by phenol/chloroform method and purified using Miniprep fosmid extraction kit (Promega) according to manufacturer’s instructions.
To construct the metagenomic fosmid library, the CopyControl™ HTP Fosmid Library Production Kit (Epicentre) was used according to the manufacturer’s instructions. In brief, prokaryotic DNA isolated from LCL was further separated on 1% low-melting point agarose by gel electrophoresis to purify the fragments that have an approximate size of 40 Kb, end-repaired to have 5’-phosphorylated blunt ends, and then ligated to the pCC2FOS vector. The ligated mix was then packed into lambda phage using lambda packaging extract and then used to transfect EPI300-T1R phage T1-resistant E. coli cells provided with the kit. The transformants were then plated on Luria-Bertani (LB) agar plates containing 12.5µg of chlormaphenicol/ml and transformantes were collected individually into 96-well plates. Using this approach a fosmid library containing 111plates comprising 10,656 clones were assembled. These steps were performed in our laboratory in collaboration with Dr. Mohamed Ghazy and Mr. Amged Ouf.
3. **Pyrosequencing**
Environmental DNA isolated from the ATLII-LCL water samples were prepared and then sequenced according to the 454 pyrosequencing GS FLX Titanium library guidelines using Roche GS FLX Titanium 454 sequencing platform located in the AUC genomics facility. The sequencing was conducted by Dr. Ahmed El Sayed, Dr. Mohamed Ghazy and Mr. Amged Ouf from AUC.

4. **Data assembly and annotation**
Data generated from 454 pyrosequencer were further assembled based on Newbler® GS assembler version 2.6 using the default parameters and minimum overlap identity value of 90%. The software Metagene Annotator (MGA) was used to detect potential ORFs within 454 assembled contigs and further annotated using different annotation pipelines as MGRAST.

5. **Subcloning of shotgun DNA fragments of the recombinant 99B6 fosmid in bacterial vector**
Six recombinant fosmid clones were randomly selected from fosmid plate #99. Subsequently, their fosmid DNA were extracted and analyzed by restriction digestion using BamHI restriction enzyme to check for recombinant clones’ heterogeneity. For the purpose of full clone insert sequencing, fosmid B6 was further arbitrary selected and fosmid DNA isolated for this purpose. Fosmid DNA was nebulized at 10 psi for 45” (Life Tech, K7025-05) and fragmented DNA between 1.5 to 2 kb were localized on the gel, and the gel containing the DNA fragments between these sizes were then excised. DNA fragments were then purified using QIAquick Gel Extraction Kit (Qiagen, 28704) and end-repaired using T4 DNA polymerase. Addition of 3’ dA overhangs was done by incubating end-repaired fragments with dATP and Taq polymerase for 1 hour at 72 °C. Finally, the A-tailed DNA fragments were ligated to pGEM®-T Easy (Promega) according to manufacturer’s instructions. Next, a 1ul of the ligation reaction was used to transform 50 ul of electrocompetent *E. coli* TOP10 cells using MicroPulser™ Electroporation Apparatus (Bio-Rad). Finally, the transformed cells were then plated on LB/Ampicillin/IPTG/X-Gal agar plates.
6. **DNA Sequencing using dideoxy chain termination technique (Sanger Method)**

Subclones were randomly sequenced using in the Applied Biosystems 3730xl DNA Analyzer using the BigDye® Terminator v3.1 Cycle Sequencing Kit. Two reactions were prepared for each plate containing: Big dye terminator, 5X Sequencing Buffer, universal primers for pGem-T easy vector (forward or reverse) and, DNA template. The primers used were T7 prom and Sp6 as universal vector primers. Thermocycler conditions were adjusted at initial denaturation step at 96°C for 1 minute, 25 cycles (denaturation at 96°C for 10 seconds, annealing at 55°C for 5 seconds and extension at 60°C for 4 minutes).

7. **Assembly of the sequenced DNA fragments**

Data assembly was done using phredPhrap consed program. All raw Sanger reads were screened from pGem-T easy vector and fosmid sequences and low quality reads were masked before assembly. Phred/Phrap/Consed was used to mask raw sequences with quality score limit with error probability 0.01(Q20). The assembly of masked raw reads was performed by Phred/Phrap/Consed using the software default parameters. Base quality score of 20 or more was assigned for bases in consensus sequence and overlap length of minimum 15 nucleotides and 95% identity were used for effective assembly.

8. **ORFs identification and gene annotation**

Putative open reading frames (ORFs) were then identified using the MetaGene Annotator tool (MGA). Artemis genome browser was used as visualization, annotation and six-frame translation tool to determine genomic features and protein signatures of identified ORFs. This was achieved using a collection of embedded sequence similarity search tools as Basic Local Alignment Search Tool (BLAST) and functional domain analysis using PFAM search engine. Furthermore, the identified ORFs were also submitted to KEGG annotation/COG search tools to assign potential metabolic pathways and detection of orthologous groups associated with identified sequences based on a collection of manually curated databases within Kyoto Encyclopedia of Genes and Genomes (KEGG) database resource and Clusters of Orthologous groups (COGs) (Wu, Zhu, Fu, Niu, & Li, 2011).
9. **Other descriptive features for identified ORFs**

Several bioinformatics tools were used to determine other descriptive features for identified ORFs. The transcriptional regulatory elements [-35,-10] of the promoter regions were traced by Bprom software ([www.softberry.com](http://www.softberry.com)) and the ribosomal binding site (RBS) was searched for manually RBS consensus sequence AGGAG and other A/G rich motifs. Furthermore, SignalP4 program (Petersen, Brunak, von Heijne, & Nielsen, 2011) was primarily used to predict possible signal peptide triggered protein secretion pathways. Alternatively, sequences were further analyzed to identify other non-classical secretory pathways and Twin Arginine secretory signals using SecretomeP (Bendtsen, Kiemer, Fausbøll, & Brunak, 2005) and PRED-TAT (Bagos, Nikolaou, Liakopoulos, & Tsirigos, 2010) respectively. The subcellular localization of proteins was predicted by PSORTb v3.0.2 (Yu et al., 2010) software and the likelihood of proteins to contain transmembrane domain/s was checked by THMM (Krogh, Larsson, von Heijne, & Sonnhammer, 2001) analysis tool. Finally, the halophilicity of potential coding sequences were estimated by calculating the ratio of glutamate and aspartate residues for pairwise alignments between ORF and corresponding reference sequence of blastx similarity search results.

Furthermore, the potential full-length ORFs were selected for detailed protein sequence analysis using a collection of protein sequence analysis tools as CDD search, Interproscan, phmmer software. For instance, the EBI Interproscan (Jones et al., 2014) is a web interface that integrates different protein sequence analysis search tools and databases into one user-friendly platform for the detection of potential functional domains and other important protein signatures.
Chapter 3: Results and Discussion

1. **Analysis of the ATII-LCL library’s fosmids by restriction endonuclease**

A metagenome fosmid library of environmental DNA isolated from microbial communities collected from the ATII-LCL was constructed in our laboratory. This library comprises 10,656 clones distributed in 111 96-well plates. To test if the fosmid clones do not have a high level of redundancy of the inserted DNAs, we randomly selected 6 clones from plate #99 for restriction endonuclease analysis of the inserted DNAs. Electrophoresis of the isolated DNAs from the clones 99A1, 99A10, 99B6, 99B12, 99F7, and 99G8, showed DNAs with high quality and high molecular weight (Figure 3). The DNAs of the 6 recombinant fosmids were digested by BamHI enzyme and the result is shown in Figure 3. As it can be seen from the restriction patterns of the 6 clones, all the clones analysed gave different restriction patterns with different sizes DNA fragments. This result indicates that no particular tendency for insertion of specific fragment of DNA occurs during construction of the library.

![Fosmid marker 40 kb](image)

<table>
<thead>
<tr>
<th>Fosmid marker 40 kb</th>
<th>1 = Undigested</th>
<th>2 = Bam HI digest</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>A10</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>B6</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>B12</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>F7</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>G8</td>
<td>1</td>
<td>2</td>
</tr>
</tbody>
</table>

Figure 3. Restriction analysis of six randomly selected recombinant fosmids.

Six clones were randomly selected; 99A1, 99A10, 99B6, 99B12 and 99G8, and their recombinant fosmids were isolated and digested with BamHI and different DNA fragments were then analysed on 0.8 % agarose gel electrophoresis as indicated in the figure.

35
2. Establishing a library for fosmid clone 99B6 in pGEM plasmid vector

To sequence the entire DNA insert from one fosmid clone, we arbitrary selected clone 99B6 for this purpose. Purified DNA from clone 99B6 was nebulized using Nebulizer (Life Tech, K7025-05) for 45 seconds at 10 psi. The condition of nebulization was adjusted to yield DNA fragments of approximate size of 1.5-2 kb (Figure 4). DNA fragments of size 1.5-2 kb were gel purified after electrophoretic separation without exposure to UV light and then ligated into pGEM-T easy vector (Promega). The ligated mixture was used to transform *E. coli* Top10 competent cell (Invitrogen) using electroporation technique. This transformation yielded 469 recombinant clones. The 469 clones were inoculated into five 96-well plates and stored at -80 °C.

![Figure 4. Isolation of 1.5 to 2 kb DNA fragments of nebulized 99B6 fosmid clone.](image)

Isolated fosmid 99B6 was nebulized for 15” at 10 psi as described in the Material and Methods section. The nebulized fragments were separated on 0.8% agarose gel electrophoresis. The fragments sizes, 1.5 to 2 kb, were isolated from the excised gel section shown in the figure without exposure to UV light.
Plasmid DNAs from 12 subclones were isolated using R.E.A.L. Prep 96 Plasmid Kit (Qiagen) and analysed on 0.8 % agarose gel electrophoresis. As it can be seen from Figure 5, the plasmids DNAs have good quality. pGEM-T easy vector has 3015 bp and each recombinant plasmid was found to have an estimated size of about 4.0 to 4.5 kb (Figure 5). Therefore, each plasmid has an insert size that range from 1 to 1.5 kb. This result indicates that the size of the inserts in our pGEM-T 99B6 library is adequate for complete sequencing of each insert using the Sanger technique using the pGEM-T easy forward and reverse primers.

![Molecular weight marker](image)

Figure 5. Analysis of the quality and integrity of clones from 99B6 pGEM library.

Recombinant plasmid isolated from 12 clones of the 99B6 pGEM library was randomly selected and their plasmids were isolated and analysed on 0.8% agarose electrophoresis.

3. **Sequencing of 99B6 pGEM library clones**

The plasmids from the 469 clones of the 99B6 pGEM library were isolated and purified as described earlier. Plasmids were sequenced using the chain-termination methods of Sanger and analysed on Applied Biosystems 3730xl DNA sequencer. Plasmid’s inserts were sequenced from both ends using the forward and the reverse primers of the pGEM-T easy vector. Out of 952 forward and reverse sequencing reaction, only 476 reads were mate-paired, i.e., sequences from the forward and the reverse reaction that was overlapped. The rest of the reads, 298, were found to be sequences of low quality, vector sequences or remained as singletons during the assembly process.
4. **Assembly of the sequenced reads**

DNA sequences from the ABI 3730xl DNA sequencer files were traced and each base was assigned a quality value using the Phred software utilizing the default parameters (a cut-off value of 20 or more represents a good), the base quality score values before and after screening process are provided in Figure 6 (Ewing & Green, 1998). The quality reads were then assembled, visualized, and edited by the Phrap and Consed softwares (Gordon, Abajian, & Green, 1998).

**A. Quality (Pre-screening)  B. Quality (Post-screening)**

![Graphs showing quality assessment](image)

Figure 6. Assigned base quality score value for sequenced reads. (A) Quality assessment for the sequenced reads before low sequence quality screening process (B) Quality for the sequenced reads after screening process using Phred/Phrap/Consed software.

The 672 assembled reads is presented in Figure 7. The assembled reads has a single contig of 28,378 bp with average read depth of coverage value of 20 times. The orientation of reads is denoted by color codes; red color represents reverse read and green represents forward reads as denoted by CLC genomics workbench tool (Figure 7).
Figure 7. Assembly of the sequenced reads. Sequenced reads were traced and bases were assigned a quality values using Phred software (Ewing & Green, 1998), and then assembled and visualized using Phrap, consed softwares and CLC genomics workbench (Gordon et al., 1998; Sequencing, 2011). For more details refer to the text and the materials and methods section.

As mentioned above, around 298 reads did not assembled in the major contig. The estimated average insert size in our fosmid library is between 30 to 40 kb. Our assembled contig has 28,378 bp, therefore, it is apparent that we did not sequence the entire insert. To estimate the size of the inserted DNA in the sequenced fosmid, and to clarify the size of the un-sequenced DNA fragment, we digested the recombinant 99B6 fosmid clone with Eco RI. Four DNA fragments were observed with the following sizes: 25, 8, 6, and 3 kb (Figure 8A). From the assembled 28,378 bp, we mapped two Eco RI sites within the inserted DNA fragment, and one Eco RI site within the fosmid vector located at 300 bp upstream from the insertion site (Figure 8B). From the sizes of the Eco RI DNA fragments, we estimated that our recombinant vector 99B6 has a size of 42 kb. Since the fosmid vector has a size of 8.181 kb, the inserted DNA fragment, therefore, should have 33.819 kb. Based on this analysis, the un-sequenced DNA fragment should have an estimated size of 5.441 kb (Figure 8B).
Figure 8. *Eco* RI restriction analysis of recombinant 99B6 fosmid clone. A) Recombinant fosmid 99B6 was isolated, purified, and digested with *Eco* RI restriction enzyme. The digested fragments were analyzed on 0.8% agarose gel electrophoresis. B) *Eco* RI restriction map of recombinant 99B6 fosmid. The sketch shows the un-sequenced fragments highlighted in red.

Several attempts were made to finish sequencing of the fosmid clone insert. As for instance, we manually inspect the base composition of the assembled 99B6 sequence (28,378 bp) and sequenced un-assembled singletons for the presence of low complexity regions and repeats. The results have shown a characteristic patterns of simple repeats and low complexity regions at the nucleotide sequences of some of the unassembled singletons as well as at the end of the assembled 99B6 sequence (28,378bp) (Figure 9). It has been estimated that regions of low complexity and simple repeats, compositional bias of sequence data (AT-rich, poly-purines etc. are presumably accompanied with DNA polymerase slippage that takes place during DNA sequencing when a nearby repeat is present (Frith, 2011). These simple repeats and low complexity regions are thus most likely associated with assembly problems and spurious annotations (Treangen & Salzberg, 2012).
Furthermore, annotation of 28,378 bp sequence from 99B6 recombinant clone, have demonstrated a gene which encodes for a methyl transferase enzyme from an uncultured microbe is duplicated, these results are further displayed in the annotation section. Accordingly, we have presumed that the presence of simple repeats, low complexity regions as shown in Figure 9, along with gene duplication at the end of the assembled 99B6 consensus sequence, have enforce us to recognize the difficulty to recover the un-assembled DNA fragment 5.441 kb of the recombinant fosmid clone 99B6 insert.

Figure 9. Nucleotide sequence composition of sequence end of assembled fosmid 99B6 and un-sequenced DNA fragment and. Detection of low complexity regions and simple repeats in both assembled sequence 99B6 (approx. last 400 bp of assembled sequence) and unassembled (sequence I and II) respectively. The estimated low complexity regions and simple repeats are highlighted with different colour codes and underlined respectively.
5. Open reading frames search for potential genes and annotation of the identified ORFs

Using Metagene Annotator (Noguchi et al., 2008), we identified 39 ORFs displayed along the 28,378 kb assembled contig. To annotate the 39 identified ORFs, different annotation tools were deployed simultaneously for data analysis including NCBI-CDD sequence search tool for conserved sequence detection of domains, Interproscan integrated pipeline for detection of protein domains, families and other protein signatures against different databases (Jones et al., 2014), Phmmmer for protein sequence similarity search based on the more sensitive hidden Markov model (hmmer.janelia.org), and finally assigning ORFs to different metabolic networks databases using COG/KEGG for orthologous group detection and infer ORF metabolic functions (Wu et al., 2011).

The outcome of the annotation processes using these programs is presented in Figure 10. Out of the 39 identified ORFs, 29 have matches; While 10 ORFs have no recognized matches using the pipelines mentioned earlier.

The 39 annotated ORFs and their orientations are presented in Figure 9 and Table 1 respectively.

![Figure 10. Protein sequence analysis for identified ORFs using different protein annotation programs.](image-url)
It is not surprisingly that about 26% of identified ORFs remained unannotated during our analysis, as previous studies have estimated that about 50% of all metagenomic data generated to date are considered as hypothetical proteins or proteins of unknown function. For instance, it was estimated that about 50% of 1.2 million protein coding genes identified from shotgun sequencing of Sargasso Sea metagenome were predicted as hypothetical proteins (Venter et al., 2004). Hence, there is a global initiative toward implementing more powerful computing resources and integrated platforms to alleviate bottlenecks of huge dataset analysis to accelerate novel gene discovery processes (Teeling & Glockner, 2012).

Table 2 presents a detailed analysis of the annotated ORFs including the % of query coverage to the best hit and the % of sequence identity and similarity. The table also presents the identification of -10 and -35 consensus sequences required for RNA polymerase interaction, potential ribosomal binding site and the start and stop codons for the translation process.

<table>
<thead>
<tr>
<th>Annotation of the 39 identified ORFs using NCBI blastx search tool</th>
</tr>
</thead>
<tbody>
<tr>
<td>1) Unknown</td>
</tr>
<tr>
<td>2) DeoR transcriptional regulatory protein</td>
</tr>
<tr>
<td>4) Transcription initiation factor IFB</td>
</tr>
<tr>
<td>6) Ribonuclease 2H</td>
</tr>
<tr>
<td>7) ECF family sigma-70 type sigma factor</td>
</tr>
</tbody>
</table>

Table 1. Schematic diagram for the location and annotation of identified ORFs on contig9. (Contig9, 28,378 kb)
Table 2. Protein sequence similarity search using blastx tool against nr protein database.

<table>
<thead>
<tr>
<th>ORFs in Contig 9</th>
<th>Similarity</th>
<th>Query Coverage</th>
<th>Identity%</th>
<th>RBS</th>
<th>-10</th>
<th>-35</th>
<th>Start Codon</th>
<th>Stop codon</th>
<th>Annotation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Contig 9 ORFs</td>
<td>39%</td>
<td>NA</td>
<td>NA</td>
<td>YES</td>
<td>NA</td>
<td>GTG*</td>
<td>YES</td>
<td>No similarity</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>43%</td>
<td>87%</td>
<td>30%</td>
<td>YES</td>
<td>YES</td>
<td>TTG*</td>
<td>YES</td>
<td>DeoR family transcriptional regulator [Bacillus megaterium QM B1551]</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>69%</td>
<td>91%</td>
<td>44%</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>hypothetical protein [Natronorubrum bangense]</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>72%</td>
<td>94%</td>
<td>53%</td>
<td>NO</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>initiation factor IIB [Archaeoglobus fulgidus DSM 4384]</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>53%</td>
<td>65%</td>
<td>41%</td>
<td>YES*</td>
<td>NO</td>
<td>NO</td>
<td>TTG*</td>
<td>YES</td>
<td>membrane protein [uncultured organism]</td>
</tr>
<tr>
<td>5</td>
<td>56%</td>
<td>58%</td>
<td>35%</td>
<td>NO</td>
<td>YES</td>
<td>YES</td>
<td>TTG*</td>
<td>YES</td>
<td>ribonuclease H [Cellulomonas fimii ATCC 484]</td>
</tr>
<tr>
<td>6</td>
<td>60%</td>
<td>25%</td>
<td>43%</td>
<td>YES*</td>
<td>YES</td>
<td>YES</td>
<td>GTG*</td>
<td>YES</td>
<td>ECF family sigma-70 type sigma factor [Rhodococcus opacus]</td>
</tr>
<tr>
<td>7</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NO</td>
<td>YES</td>
<td>YES</td>
<td>TTG*</td>
<td>YES</td>
<td>NO similarity</td>
</tr>
<tr>
<td>8</td>
<td>53%</td>
<td>67%</td>
<td>29%</td>
<td>YES</td>
<td>YES</td>
<td>GTG*</td>
<td>YES</td>
<td>serine/threonine kinase family protein [uncultured bacterium A1Q1_fos_1807]</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>62%</td>
<td>37%</td>
<td>49%</td>
<td>YES*</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>NADPH-dependent F420 reductase [Natraliba chabanaeomnis]</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>54%</td>
<td>50%</td>
<td>29%</td>
<td>NO</td>
<td>YES</td>
<td>YES</td>
<td>GTG*</td>
<td>YES</td>
<td>AhpC/TSA family protein [Zunongwangia profanda SM-A87]</td>
</tr>
<tr>
<td>11</td>
<td>56%</td>
<td>54%</td>
<td>42%</td>
<td>NO</td>
<td>NO</td>
<td>NO</td>
<td>GTG*</td>
<td>YES</td>
<td>Glutathione S-transferase [Pseudomonas pseudoalcaligenes]</td>
</tr>
<tr>
<td>12</td>
<td>61%</td>
<td>52%</td>
<td>39%</td>
<td>NO</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>serine protease [Streptococcus mitis B6]</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>59%</td>
<td>74%</td>
<td>33%</td>
<td>YES*</td>
<td>YES</td>
<td>YES</td>
<td>TTG*</td>
<td>YES</td>
<td>2-hydroxyxypenta-2,4-dienoate hydratase [Pseudomonas stutzeri]</td>
</tr>
<tr>
<td>14</td>
<td>62%</td>
<td>58%</td>
<td>44%</td>
<td>YES*</td>
<td>YES</td>
<td>YES</td>
<td>GTG*</td>
<td>YES</td>
<td>hypothetical protein GUTHDRAFT_88488 [Guillardia theta CCMP2712]</td>
</tr>
<tr>
<td>15</td>
<td>51%</td>
<td>36%</td>
<td>30%</td>
<td>NO</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>Hore_22870 hypothetical protein [ Halothermothrix oreinii H 168 ]</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>59%</td>
<td>79%</td>
<td>43%</td>
<td>YES*</td>
<td>YES</td>
<td>YES</td>
<td>GTG*</td>
<td>Yes</td>
<td>PREDICTED: similar to GTP cyclohydrolase I [Tricholobium castanenum]</td>
</tr>
<tr>
<td>17</td>
<td>49%</td>
<td>46%</td>
<td>37%</td>
<td>YES*</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>sterol 3beta-glucosyltransferase [Puccinia graminis f. sp. tritici CRL 75-36-700-3]</td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>52%</td>
<td>56%</td>
<td>35%</td>
<td>YES*</td>
<td>YES</td>
<td>GTG*</td>
<td>YES</td>
<td>hypothetical protein [Bacillus cereus]</td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>48%</td>
<td>60%</td>
<td>33%</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>GTG*</td>
<td>YES</td>
<td>60 kDa chaperonin [uncultured pig faeces bacterium]</td>
</tr>
<tr>
<td>20</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>YES*</td>
<td>YES</td>
<td>YES</td>
<td>GTG*</td>
<td>YES</td>
<td>No similarity</td>
</tr>
<tr>
<td>21</td>
<td>50%</td>
<td>43%</td>
<td>46%</td>
<td>NO</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>hypothetical protein Deide_07800 [Denococcus deserti VCD115]</td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>39%</td>
<td>65%</td>
<td>27%</td>
<td>NO</td>
<td>YES</td>
<td>YES</td>
<td>TTG*</td>
<td>YES</td>
<td>PREDICTED: collagen alpha-(V1) chain [Orcinus orca]</td>
</tr>
<tr>
<td>23</td>
<td>41%</td>
<td>92%</td>
<td>28%</td>
<td>NO</td>
<td>YES</td>
<td>YES</td>
<td>GTG*</td>
<td>YES</td>
<td>hypothetical protein LEPBI_L2319 [Leptospira biflexa serovar Patoc strain Patoc 1 (Paris)]</td>
</tr>
<tr>
<td>24</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>YES</td>
<td>YES</td>
<td>NO</td>
<td>NO</td>
<td>No similarity</td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>YES</td>
<td>YES</td>
<td>GTG*</td>
<td>NO</td>
<td>No similarity</td>
<td></td>
</tr>
<tr>
<td>26</td>
<td>47%</td>
<td>89%</td>
<td>31%</td>
<td>YES*</td>
<td>YES</td>
<td>YES</td>
<td>GTG*</td>
<td>YES</td>
<td>arabinosyltransferase [Mycobacterium leprae TN]</td>
</tr>
<tr>
<td>27</td>
<td>52%</td>
<td>19%</td>
<td>39%</td>
<td>NO</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>hypothetical protein GYMC_1090 [Paeucibacillus sp. Y412M10]</td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>47%</td>
<td>29%</td>
<td>37%</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>GTG*</td>
<td>YES</td>
<td>filamentous hemagglutinin outer membrane protein [Enterobacter sp. SST3]</td>
</tr>
<tr>
<td>29</td>
<td>47%</td>
<td>29%</td>
<td>37%</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>GTG*</td>
<td>YES</td>
<td>hypothetical protein Thimo_0605 [Thioflavicoccus mobilis 8321]</td>
</tr>
<tr>
<td>30</td>
<td>92%</td>
<td>97%</td>
<td>30%</td>
<td>NO</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>GH13668 [Drosophila grumshawi]</td>
<td></td>
</tr>
<tr>
<td>31</td>
<td>62%</td>
<td>24%</td>
<td>43%</td>
<td>NO</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>Putative subtilisin peptidase [Micromonospora lupina]</td>
<td></td>
</tr>
<tr>
<td>32</td>
<td>50%</td>
<td>12%</td>
<td>35%</td>
<td>NO</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>hypothetical protein Thino_0605 [Thioflavicoccus mobilis 8321]</td>
<td></td>
</tr>
<tr>
<td>33</td>
<td>58%</td>
<td>97%</td>
<td>39%</td>
<td>NO</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>Polyprenyl synthetase [Acidiphilum profundum boonei T469]</td>
<td></td>
</tr>
<tr>
<td>34</td>
<td>64%</td>
<td>18%</td>
<td>52%</td>
<td>NO</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>membrane-associated protease [Prochlorococcus marinus str. AS9601]</td>
<td></td>
</tr>
<tr>
<td>35</td>
<td>50%</td>
<td>56%</td>
<td>32%</td>
<td>NO</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>membrane or secreted protein [uncultured organism]</td>
<td></td>
</tr>
<tr>
<td>36</td>
<td>64%</td>
<td>59%</td>
<td>43%</td>
<td>YES*</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>Methyltransferase type 11 [uncultured bacterium]</td>
<td></td>
</tr>
<tr>
<td>37</td>
<td>49%</td>
<td>49%</td>
<td>26%</td>
<td>NO</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>glycosyl transferase family protein [Rhodobacter sphaeroides ATCC 17025]</td>
<td></td>
</tr>
<tr>
<td>38</td>
<td>47%</td>
<td>98%</td>
<td>33%</td>
<td>NO</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>hypothetical protein [Halorubrum tebenquichense]</td>
<td></td>
</tr>
<tr>
<td>39</td>
<td>64%</td>
<td>59%</td>
<td>43%</td>
<td>YES*</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>Methyltransferase type 11 [uncultured bacterium]</td>
<td></td>
</tr>
</tbody>
</table>

* ‘highlighted’ ORF descriptions: Blastx hits resulted using significant E-value (10E-5).
6. Structural characteristics of the annotated proteins

Potential halophilic features of the annotated proteins — Microorganisms that survive in halophilic environments are classified into slightly, moderately, and extremely halophilic, with an optimal growth in 0.5, 0.5–2.5, and around 4 M NaCl respectively (Oren, 2002; Sayed et al., 2014). Two mechanisms are known that are adapted by halophilic microorganisms to cope with the high salinity of its surrounding environment and keep the osmotic pressure of the cell within the physiological range, the salt-out and the salt-in mechanisms. Microorganisms that adapted the salt-out approach accumulate within their cytoplasm osmotic solute such as glycine, betaine, and ectoine (Oren, 2008). The second approach, the salt-in, the microorganism accumulates molar concentration of potassium chloride to balance the high salt concentration of the surrounding environment (Christian & Waltho, 1962; Chromatography et al., 1992; Lanyi, 1974). In this case the proteins are adapted to function at high salt concentration reaching up to 4 molars. Such adaptations include increase of acidic amino acid residues on the surface of the protein and decrease the in hydrophobic amino acids (Paul, Bag, Das, Harvill, & Dutta, 2008; Siglioccolo, Paiardini, Piscitelli, & Pascarella, 2011).

The halophilicity of the annotated proteins was estimated by calculating the ratio between the numbers of acidic amino acid residues (D+E) in each protein coding sequence and the number of acidic amino acid residues (D+E) in corresponding homolog blastx match. A protein is set to be potentially halophilic if possess a halophilic ratio of 1.25 or more. As it can be seen from, Table 2, 14 out of the 39 Proteins have ratios of D+E at least 1.25 times higher than their corresponding homologous proteins indicating that they may have some level of halophilic properties.
Table 3. Determination of halophilic potential for identified ORFs. The halophilicy of the annotated orfs was estimated by calculating the ratio between the numbers of acidic amino acid residues (D+E) in each protein coding sequence and the number of acidic amino acid residues (D+E) in corresponding blastx matches.

<table>
<thead>
<tr>
<th>ORFs in Contig 9</th>
<th>Annotation</th>
<th>D+E ORF</th>
<th>D+E(REF)</th>
<th>orf/ref ratio</th>
<th>Halophilicity</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>No similarity</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>DeoR family transcriptional regulator [Bacillus megaterium QM B1551]</td>
<td>3</td>
<td>8</td>
<td>0.375</td>
<td>NO</td>
</tr>
<tr>
<td>3</td>
<td>hypothetical protein [Nitronorurum bangense]</td>
<td>14</td>
<td>8</td>
<td>1.75</td>
<td>NO</td>
</tr>
<tr>
<td>4</td>
<td>initiation factor IIB [Archeoglobus fulgidus DSM 4304]</td>
<td>46</td>
<td>41</td>
<td>1.12</td>
<td>NO</td>
</tr>
<tr>
<td>5</td>
<td>membrane protein [uncultured organism]</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>NO</td>
</tr>
<tr>
<td>6</td>
<td>ribonuclease H [Cellulomonas fimii ATCC 484]</td>
<td>5</td>
<td>4</td>
<td>1.25</td>
<td>YES</td>
</tr>
<tr>
<td>7</td>
<td>ECF family sigma-70 type sigma factor [Rhodococcus opacus]</td>
<td>5</td>
<td>6</td>
<td>0.833</td>
<td>NO</td>
</tr>
<tr>
<td>8</td>
<td>No similarity</td>
<td>4</td>
<td>9</td>
<td>0.444</td>
<td>NO</td>
</tr>
<tr>
<td>9</td>
<td>serine/threonine kinase family protein [uncultured bacterium A1Q1_fos_1807]</td>
<td>27</td>
<td>16</td>
<td>1.687</td>
<td>YES</td>
</tr>
<tr>
<td>10</td>
<td>NADPH-dependent F420 reductase [Natralba chahannaensis]</td>
<td>5</td>
<td>6</td>
<td>0.833</td>
<td>NO</td>
</tr>
<tr>
<td>11</td>
<td>AhpC/TSA family protein [Zangongwanga profunda SM-A67]</td>
<td>13</td>
<td>9</td>
<td>1.44</td>
<td>YES</td>
</tr>
<tr>
<td>12</td>
<td>Glutathione S-transferase [Pseudomonas pseudoalcaligenes]</td>
<td>7</td>
<td>5</td>
<td>1.4</td>
<td>YES</td>
</tr>
<tr>
<td>13</td>
<td>serine protease [Streptococcus mitis B6]</td>
<td>1</td>
<td>0</td>
<td>no ratio</td>
<td>N/A</td>
</tr>
<tr>
<td>14</td>
<td>2-hydroxypenta-2,4-dienoate hydratase [Pseudomonas stutzeri]</td>
<td>8</td>
<td>7</td>
<td>1.14</td>
<td>NO</td>
</tr>
<tr>
<td>15</td>
<td>hypothetical protein GUTHDRAFT_88488 [Guillardia theta CCMP2712]</td>
<td>10</td>
<td>6</td>
<td>1.66</td>
<td>YES</td>
</tr>
<tr>
<td>16</td>
<td>Hore_22870 hypothetical protein [Halothermothrix orenii H 168 ]</td>
<td>23</td>
<td>21</td>
<td>1.09</td>
<td>NO</td>
</tr>
<tr>
<td>17</td>
<td>PREDICTED: similar to GTP cyclohydrolase I [Tribolium castaneum]</td>
<td>3</td>
<td>2</td>
<td>1.5</td>
<td>YES</td>
</tr>
<tr>
<td>18</td>
<td>stereol 3beta-glycolytransferase [Puccinia graminis f. sp. tritici CRL 75-36-700-3]</td>
<td>13</td>
<td>14</td>
<td>0.92</td>
<td>NO</td>
</tr>
<tr>
<td>19</td>
<td>hypothetical protein [Bacillus cereus]</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>NO</td>
</tr>
<tr>
<td>20</td>
<td>60 kDa chaperonin [uncultured pig faeces bacterium]</td>
<td>8</td>
<td>9</td>
<td>0.889</td>
<td>NO</td>
</tr>
<tr>
<td>21</td>
<td>No similarity</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>22</td>
<td>hypothetical protein Deide_07800 [Deinococcus deserti VCD115]</td>
<td>5</td>
<td>6</td>
<td>0.833</td>
<td>NO</td>
</tr>
<tr>
<td>23</td>
<td>PREDICTED: collagen alpha-5(VI) chain [Orcinus orca]</td>
<td>37</td>
<td>20</td>
<td>1.85</td>
<td>YES</td>
</tr>
<tr>
<td>24</td>
<td>hypothetical protein LEPBI_I2319 [Leptospira biflexa serovar Patoc strain Patoc 1 (Paris)]</td>
<td>27</td>
<td>21</td>
<td>1.28</td>
<td>YES</td>
</tr>
<tr>
<td>25</td>
<td>No similarity</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>26</td>
<td>No similarity</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>27</td>
<td>arabinosyltransferase [Mycobacterium leprae TN]</td>
<td>3</td>
<td>9</td>
<td>0.333</td>
<td>NO</td>
</tr>
<tr>
<td>28</td>
<td>hypothetical protein GYMC10_1090 [Paenibacillus sp. Y412MC10]</td>
<td>9</td>
<td>10</td>
<td>0.9</td>
<td>NO</td>
</tr>
<tr>
<td>29</td>
<td>filamentous hemagglutinin outer membrane protein [Enterobacter sp. SSTS]</td>
<td>10</td>
<td>7</td>
<td>1.42</td>
<td>YES</td>
</tr>
<tr>
<td>30</td>
<td>GH35068 [Drosophila grimshawi]</td>
<td>15</td>
<td>23</td>
<td>0.65</td>
<td>NO</td>
</tr>
<tr>
<td>31</td>
<td>Putative subtilisin peptidase [Micromonospora lupini]</td>
<td>10</td>
<td>15</td>
<td>0.667</td>
<td>NO</td>
</tr>
<tr>
<td>32</td>
<td>hypothetical protein Thino_0605 [Theclavicoccus mobilis 8321]</td>
<td>27</td>
<td>20</td>
<td>1.35</td>
<td>YES</td>
</tr>
<tr>
<td>33</td>
<td>Polypropenyl synthetase [Acidifilum profundum boonei T469]</td>
<td>70</td>
<td>61</td>
<td>1.14</td>
<td>YES</td>
</tr>
<tr>
<td>34</td>
<td>membrane-associated protease [Pochlorococcus marinus str. AS6001]</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>YES</td>
</tr>
<tr>
<td>35</td>
<td>membrane or secreted protein [uncultured organism]</td>
<td>18</td>
<td>16</td>
<td>1.125</td>
<td>NO</td>
</tr>
<tr>
<td>36</td>
<td>Methyltransferase type 11 [uncultured bacterium]</td>
<td>20</td>
<td>11</td>
<td>1.81</td>
<td>YES</td>
</tr>
<tr>
<td>37</td>
<td>glycosyl transferase family protein [Rhodobacter sphaeroides ATCC 17025]</td>
<td>12</td>
<td>7</td>
<td>1.71</td>
<td>YES</td>
</tr>
<tr>
<td>38</td>
<td>hypothetical protein [Halobacterium tebenquichense]</td>
<td>40</td>
<td>64</td>
<td>0.625</td>
<td>NO</td>
</tr>
<tr>
<td>39</td>
<td>Methyltransferase type 11 [uncultured bacterium]</td>
<td>20</td>
<td>11</td>
<td>1.81</td>
<td>YES</td>
</tr>
</tbody>
</table>

#Putative Proteins are considered potentially halophilic protein if the halophilic ratio is 1.25 or more.
Detection of potential secretory signals — Detection of potential secretory signals and THMM domains within annotated ORFs were performed using different detection online tools as well as predictions for protein subcellular localization of putative proteins (Bagos et al., 2010; Bendtsen et al., 2005; Krogh et al., 2001; Petersen et al., 2011; Yu et al., 2010). Figure 11 shows the different secretory pathways that were identified in the 39 potential proteins based on the presence of sequences and domains involved in the corresponding secretory process.

The results indicate that the majority of the potential secreted proteins identified using the above mentioned tools use a non-classical secretory pathway (20 out of 39; around 50%), which uses signal peptide independent secretion pathways (Bendtsen et al., 2005). The second secretory pathway that is potentially used by the microbial community in LCL environment employ transmembrane domains (16 out of 39; around 40%). Interestingly, just very few proteins (4 out of 39; around 10%) were found a signal peptide using the classical secretion process.

Figure 11. Prediction of putative secretory signals and THMMs domains in annotated ORFs.
7. **Detailed analysis of 10 selected ORFs identified by Interproscan software**

Out of the 39 ORFs identified, 10 potential full-length ORFs were annotated using Interproscan web interface software. The detailed analysis of each ORF including the identified specific domains for the putative proteins is presented in Figure 12.

Figure 12. Schematic representations of potential full-length ORFs in silico annotated using Interproscan protein analysis software.

1. **ORF#4**

   a) Gene level
   - Position : start: 1887, end: 2837
   - Blast best-hit: Transcription factor TFIIB [*Archaeoglobus profundus DSM 5631*]
   - Query coverage: 94%
   - Identity: 53%, Fraction conserved: 72%

   b) Protein level (length 316 amino acids)
   - Zinc finger, TFIIB type IPR013137 (18-59)
   - Transcription factor TFIIB, Cyclin-like domain IPR013150 (134-202, 228-298)
   - Transcription factor TFIIB, conserved site IPR023486 (260-275)
   - Non-classical protein secretion detected by SecretomeP
   - No transmembrane regions detected
   - Cytoplasmic protein
   - Halophilic ratio 1.25
2. ORF#9

a) Gene level
- RBS: GGAGG
- Position: start: 4278, end: 4676
- Blast best-hit: Serine/Threonine Kinase family protein [uncultured bacterium AIQ1_fos_1807]
- Query coverage: 67%
- Identity: 29%, Fraction conserved: 53%

b) Protein level (length 132 amino acids)
- AraC-type arabinose-binding/dimerization domain IPR003313 (20-104)
- Possible coil-coil region (74-102, 107-128)
- Non-classical protein secretion detected by SecretomeP
- No transmembrane regions detected
- Cytoplasmic protein
- Halophilic ratio 1.687

3. ORF#14

a) Gene level
- RBS:
- Position: start: 5921, end: 6151
- Blast best-hit: 2-hydroxypenta-2, 4-dienoate hydratase [Pseudomonas stutzeri]
- Query coverage: 74%
- Identity: 33%, Fraction conserved: 59%
b) Protein level (76 amino acids)
- Ribbon-helix-helix protein, copG (Arc-type repressor) domain IPR02145 (10-43)
- No secretory signals detected
- No transmembrane regions detected
- Cytoplasmic protein
- Halophilic ratio 1.14

4. ORF#28

a) Gene level
- Position: start: 13974, end: 15800
- Blast best-hit: Hypothetical protein GYMC10_1090
  [Paenibacillus sp, Y41MC10]
- Query coverage: 19%
- Identity: 39%, Fraction conserved: 52%

b) Protein level (608 amino acids)
- Carbohydrate-binding-like fold IPR013784 (492-571)
- Domain of unknown function 4480 (DUF4480), IPR027804 (497-575)
- Non-classical protein secretion detected by SecretomeP
- Two transmembrane regions detected
- Unknown localization
- Halophilic ratio 0.9
5. **ORF#31**

a) Gene level
- Position: start: 17001, end: 18302
- Blast best-hit: Putative subtilisin peptidase \([Micromonospora lupini]\)
- Identity: 43%, Fraction conserved: 62%

![Gene level diagram]

b) Protein level (length 433 amino acids)
- Peptidase S8/S53 domain IPR000209 (117-382).
- Peptidase S8, subtilisin, Ser-active site IPR023828 characterized by catalytic triad Asp, His and Ser (342-352).
- Signal peptide detected by Secretome P, Pred-TAT and THMM
- One transmembrane region detected, most likely detected as signal peptide
- Cytoplasmic membrane protein
- Halophilic ratio 0.667

![Protein level diagram]
6. **ORF#32**

a) **Gene level**
- **RBS**: GGAGG
- **Position**: start: 18351, end: 22940
- **Blast best-hit**: Hypothetical protein Thimo_0605 \( [Thioflavicoccus mobilis 8321] \)
- **Query coverage**: 12%
- **Identity**: 35%, **Fraction conserved**: 50%

b) **Protein level** (length 1529 amino acids)
- Concanavalin A like lectin/glucanase, subgroup domain IPR013320 (82-296)
- Glycosyl hydrolase, five bladed beta propeller domain IPR023296 (585-827)
- Signal peptide detected by SignalP4, Secretome P and Pred-TAT and THMM
- Two transmembrane regions detected, one of them most likely detected as signal peptide
- Probable multiple localization sites (as cytoplasmic membrane or outer membrane protein)
- Halophilic ratio 1.36
7. **ORF#33**

a) **Gene level**
- Position: start 22994, end 24010
- Blast best-hit: Polytprenyl synthetase [*Aciduliprobundum boonei T467*]
- Query coverage: 97%
- Identity: 39%, Fraction conserved: 58%

b) **Protein level (338 amino acids)**
- Terpenoid synthase domain [IPR008949 (10-337)]
- Polytprenyl synthase_1 PS00723 (81-95) and polytprenyl synthase_2 PS00444 (222-234) aspartic acid rich patterns
- Non-classical protein secretion detected by SecretomeP
- No transmembrane regions detected
- Cytoplasmic protein
- Halophilic ratio 1.14
8. ORF#34

a) Gene level
- Position: start 24011, end 24772
- Blast best-hit: Membrane associated protease [Prochlorococcus marinus str.AS9601]
- Query coverage: 18%
- Identity: 52%, Fraction conserved: 64%

b) Protein level (length 253 amino acids)
- CAAX amino terminal protease family IPR003675 and Abi domain PF02517 (152-244)
- Non classical protein secretion detected by Signal peptide detected by Secretome P
- Eight transmembrane region detected
- Cytoplasmic membrane
- Halophilic ratio 1
9. ORF#38

a) Gene level
   - RBS: AGxAGG/AGGxGG
   - Position: start: 26948, end: 27778
   - Blast best-hit: Hypothetical protein *Halorubrum tebenquichense*
   - Query coverage: 60%
   - Identity: 33%, Fraction conserved: 47%

b) Protein level (length 276 amino acids)
   - Alkaline-phosphatase-like, core domain IPR017850 (86-135, 207-258)
   - No Signal peptide detected
   - No transmembrane regions detected
   - Cytoplasmic protein
   - Halophilic ratio 0.625
10. ORF#39

a) Gene level
- RBS : GGA/GAG/AGG
- Position : start: 27768, end : 28256
- Blast best-hit : Methyltransferase type 11 [uncultured bacterium]
- Query coverage : 59%
- Identity: 43%, Fraction conserved: 64%

b) Protein level (length 162 amino acids)
- S-adenosyl-L-Methionine-dependent methyltransferases superfamily SSF53335 (3-140), Methyltransferase_23 domain PF13489 (9-127) and no IPR
- Non classical protein secretion detected by Signal peptide detected by SecretomeP
- Cytoplasmic protein
- No transmembrane regions detected
- Halophilic ratio 1.81

Methyltransferase_23 domain (9-127)
The translation start and the stop codons of each ORF and the upstream transcription -35 and -10 sequences were identified using MGA and PBROM programs respectively. The RBS sequence was searched for manually. The RBS consensus sequence AGGAG was detected for 6 out of the 10 analyzed ORFs, suggesting that RBS of the uncultured metagenome could possess more diverse RBS sequences than commonly known RBS conserved sequences.

Furthermore, the 10 selected full-length ORFs were further investigated for the presence of potential secretory signals and proteins localization sites were also predicted using a different set of bioinformatics tools. The results from Signal4P analysis demonstrated that, out of the ten selected full-length ORFs, two putative proteins were predicted to have signal peptides needed for classical protein secretion. While for the remaining ORFs that possess no signal peptides, five were found to be secreted through non-classical pathway, and two were not secreted.

Furthermore, the protein localization sites were predicted for the two aforementioned secreted proteins using PSORT analysis. The analysis results suggest that ORF#31 which potentially encodes for subtilisin was secreted to the cytoplasmic membrane, whereas the subcellular location of the hypothetical Thimo_0605 protein encoded by ORF#32 was not exactly determined, suggesting that it might have multiple localization sites and hence, was further regarded as a cytoplasmic membrane or an outer membrane protein.

As for the remaining 8 ORFs that possess no secretory signals, they were further analysed by SecretomeP that further categories them into non-classically secreted proteins and proteins that are not secreted (Yu et al., 2010).

Results from both SecretomeP and PSORT analysis revealed that ORF#14 (2-hydroxypenta-2, 4-dienoate hydratase) and ORF#38 (Hypothetical protein) were potentially a cytoplasmic proteins.

Regarding ORF#28, which encodes for Hypothetical protein GYMC10_1090, was predicted by SecretomeP to be secreted through a non-classical secretory pathway. However, the subcellular localization of the protein was not known.
Similarly, ORFs#4 (Serine/Threonine kinase family protein), ORF# 9 (Transcription factor TFIIB), ORF#33(Polyprenyl synthetase), and ORF#39 (Methyltransferase type 11), were regarded as cytoplasmic proteins by PSORT analysis. However, these findings are contradicted by SecretomeP results which indicated their potential non-classical secretion of proteins.

As for the membrane associated protease which was encoded by ORF#34, it was denoted as a potential cytoplasmic membrane protein by both PSORT and SecretomeP; these findings would presumably indicate that the protein is secreted non-classically.

In addition, the full-length ten ORFs were assessed for the presence of transmembrane domains using the THMM search tool. The results have shown that, only ORFs# 28, 32 and 34 have displayed true positive signals for the presence of transmembrane regions within proteins. For instance, THMM analysis of the putative membrane associated protease (ORF#34) has predicted 8 transmembrane helices within the protein. Moreover, THMM assigned the first transmembrane domain detected in the N-terminus region as a potential signal peptide sequence (Figure 13).

Figure 13. The prediction of transmembrane domains for the putative membrane associated protease (ORF#34) was detected using the THMM tool.

The putative enzyme displayed 8 transmembrane helices with Topology=i2-20o53-70i77-99o119-141i153-175o185-204i209-226o230-252i. Moreover, THMM assigned the first transmembrane domain detected in the N-terminus region as a potential signal peptide sequence
Finally, the halophilic potential of the 10 full-length annotated ORFs was further estimated. By comparing the number of acidic amino acid residues (D+E) in our selected ORFs to number of acidic amino acid residues in the corresponding best matches of pairwise alignment; ORFs #4, 9, 32 and 39 have displayed halophilic ratios greater than 1.25, and thus were regarded as putative halophilic proteins (Danson & Hough, 1997).

It has been estimated that bacteria and archaea mainly depend on two mechanisms to protein machinery characterized by high acidic amino acid residues to high salt concentration and osmotic stress of extreme environments; these acidic proteins are stabilized through increase of intracellular K+ ions levels to balance excess salt concentration outside (salt-in) which is common for most halophilic archaea and some extremely halophiles from domain bacteria. Otherwise, the accumulation of compatible organic solutes (osmolytes) inside the cell without altering protein machinery is capable to maintain osmotic stability across cell. A good example for the use of organic molecules as osmolytes by different domains of life is represented by the accumulation of ectoine and glycine betaine by halophilic bacteria and use of other osmolytes by several archaeal lineages as halophilic methangens (salt-out) (a Oren, 2002). As for the molecular adaptation of secreted halophilic proteins to high salt concentration, it is noteworthy that they were not only characterized by increased acidic amino acid residues as Aspartate and Glutamate, decreased hydrophobic amino acid residues as cysteine and thus less positive charges and hydrophobic surfaces, but also it is presumed that some transport systems as TAT-pathways are usually involved to maintain structural stability of these acidic proteins as noticed for several haloarchaea (Rose, Brüser, Kissinger, & Pohlschröder, 2002). However, according to our findings there were no TAT secretory signals detected for any of the identified orfs using PRED-TAT software.
8. **Identified proteins with potential biotechnological applications**

Table 4 shows description and potential applications of some putative genes of interest identified in our metagenomic study.

Table 4. ORFs identified in contig9 (99B6) with potential industrial or medical applications.

<table>
<thead>
<tr>
<th>Protein Name</th>
<th>ORF Number</th>
<th>Potential Applications</th>
</tr>
</thead>
<tbody>
<tr>
<td>Serine peptidase</td>
<td>31</td>
<td>Pharmaceutical and Industrial applications in detergents and paper industries (Pushpam, Rajesh, &amp; Gunasekaran, 2011)</td>
</tr>
<tr>
<td>Glycosyl hydrolase family 43 (GH43)</td>
<td>32</td>
<td>Valuable products (e.g. Arabinan) used in Food Industry (Shallom &amp; Shoham, 2003; Shi et al., 2014)</td>
</tr>
<tr>
<td>Na+/K+ antiporter</td>
<td>25</td>
<td>Salt tolerance; agriculture and industrial applications(Khan, 2011; Waditee, Hibino, Nakamura, Incharoensakdi, &amp; Takabe, 2002)</td>
</tr>
<tr>
<td>Geranyl pyrophosphate synthase</td>
<td>34</td>
<td>Isoprenoids with cosmetic, pharmaceutical and food applications (Holstein &amp; Hohl, 2004)</td>
</tr>
<tr>
<td>Methylase enzyme</td>
<td>36 and 39</td>
<td>VitK2 and coenzyme Q10 biosynthesis; they maintain coagulation homeostasis and prevention of neurodegenerative disorders (Cluis, Burja, &amp; Martin, 2007; Conly, Stein, Worobetz, &amp; Rutledge-Harding, 1994).</td>
</tr>
<tr>
<td>Biotin synthase</td>
<td>16</td>
<td>Biotin with pharmaceutical and livestock industries (Ikeda et al., 2013).</td>
</tr>
<tr>
<td>Hyaluronan synthase (GT2)</td>
<td>37</td>
<td>Production of Hyaluronic acid used in cosmetics and pharmaceutical preparations (Long Liu, Liu, Li, Du, &amp; Chen, 2011).</td>
</tr>
<tr>
<td>Antimicrobial peptide</td>
<td>15</td>
<td>Antibacterial and antifungal peptide involved in pharmaceutical preparations and as a food preservative against food pathogens (Pushpanathan et al., 2012; Pushpanathan, Gunasekaran, &amp; Rajendhran, 2013).</td>
</tr>
</tbody>
</table>

For example, the potential *in-silico* characterization of novel member of subtilisin S8/S53 serine peptidase superfamily from our metagenomic dataset (ORF#31, NCBI Blast CDD, Interproscan and COG) that contains peptidase S8/S53 domain and Ser active site (Asp, His and Ser catalytic triad residues) essential for activity. As previously mentioned, industrial applications of proteases account for two-thirds of sales in the worldwide enzyme industry, this is due to their potential role in different pharmaceutical and industrial applications (Pushpam et al., 2011).
Another remarkable finding from our ATLCL brine pool metagenomic library is the identification of a novel glycosyl hydrolase like protein (ORF #32, CDD and Interproscan), where the domain structure analysis of the putative protein by Interproscan suggests that it most probably belongs to glycosyl hydrolase, family 43. In general, O-Glycosyl hydrolases represent a group of enzymes that are involved in glycosidic bond cleavage between carbohydrates or carbohydrate and non-carbohydrate moieties. They are further classified based on sequence homology into 85 protein families as provided in the CAZy (CArbohydrate-Active EnZymes) database. Glycosyl hydrolase, family 43(GH43) involves different GH members of varying enzymatic activities, such as; beta-xylosidase, alpha-L-arabinofuranosidase, arabinanase and xylanase which are generally involved in the cleavage of different polysaccharides. The detection of glycosyl hydrolase, five-bladed beta propellor domain in ORF#32 further describes it as a member of GH43 family. This domain was first displayed in the three-dimensional structure of a-L-arabinanase enzyme isolated from C. japonicus (Shallom & Shoham, 2003). However, experimental analysis for enzyme activity and substrate specificity is required for full characterization. Members of GH43 family are potentially used for many industrial applications, such as: pulp and paper industry (xylanases), the food industry (arabinanase and alpha-L-arabinofuranosidase) (Shallom & Shoham, 2003; Shi et al., 2014) and xylitol manufacturing processes (xylosidase) (Prakasham, Rao, & Hobbs, 2009).

Another protein of interest is Na+/K+ antiporter found in our identified metagenomic clone (ORF#25, COG). Na+/K+ antiporter is a ubiquitous membrane protein that can be located at the cytoplasm or cellular membrane of many microbial origins as plants, animals microbes and humans (Liew, Illias, Mahadi, & Najimudin, 2007). Their main physiological role is to maintain pH homeostasis of cells in the presence of high salt concentration environment. In addition, the application of Na+/K+ antiporter proteins in transgenic plants and engineered microbes for salt tolerance in agriculture and industry is currently reviewed in several studies (Khan, 2011; Waditee et al., 2002).
Our results were not only exclusive for the isolation of novel biocatalysts but they also included the identification of other important enzymes as putative geranyl geranyl pyrophosphate (GGPP) synthase (ORF#34), which is involved in the biosynthetic pathway of several biologically important isoprenoids. The Intermediate product GGPP acts as a precursor for several monoterpenes and monoterprenoids that are primarily found in herbs and plant extracts and commonly used in the food and cosmetic industries. Recently, some of these derivatives such as: limonene and perillyl alcohol, were assessed for their potential role as chemopreventive agents (Holstein & Hohl, 2004).

We found a gene that encodes for a putative methylase enzyme involved in vitamin K2 and coenzyme Q10 biosynthesis (ORF#39, NCBI CDD, Interproscan and phmmer). Vitamin K2 is an essential cofactor involved in the production of various clotting factors to maintain coagulation homeostasis of the body. Hence, deficiency in vitk2 production from diet of inadequate production by intestinal microbiota might lead to significant coagulopathy (Conly et al., 1994). As for coenzyme Q10, there is growing evidence that coenzyme Q10 supplementations help in treatments of various disease conditions such as cardiomyopathy, diabetes and Alzheimer’s. In addition, it has been acknowledged in cosmetics applications owing to its antioxidant properties (Cluis et al., 2007).

We also report the identification of a gene classified as a biotin synthases (ORF#16, COG). Biotin (vitamin B7) is not only regarded as an important cofactor but also have significant contribution in many industrial applications such as cosmetics, pharmaceuticals and livestock industries. It was estimated that the global market for biotin production have reached 10-30 tons with a hundred million U.S. dollars per year. Recent attempts investigate the potential use of molecular biology and genetic engineering to produce an ecofriendly biotin, instead of more hazardous multistep chemical synthesis (Ikeda et al., 2013).

It is noteworthy; a gene was annotated according to COG classification as a putative glycosyl transferase with hyaluronic acid synthase activity (ORF#37). Hyaluronic acid is one of the valuable natural sugar polymer produced by glycosyl tranferases_2 (GT) which is a member of large protein families that are involved in the biosynthesis and transfer of different polysaccharides, oligosaccharides and glycoconjugates which mediate various functions as ranging from structure and storage to signalling. Hyaluronic acid is highly diverse in prokaryotes and also present in eukaryotes. The most common forms of GTs are involved in sugar residue transfer from activated nucleophile sugar donor to specific acceptor molecule
that takes place in new sugar biosynthesis through the retention of the inversion of the configuration of anomeric carbon (Breton, Snajdrová, Jeanneau, Koca, & Imberty, 2006). Owing to its unique rheological, physiological, biological characteristics and safety, it is currently implicated in many Industrial and medical applications. It is estimated that the global market for Hyaluronic acid to have reached over $1 billion with prominent increase in the fields of dermal fillers and viscosupplementation. Recently, the recombinant microbial production of hyaluronic acid has gained an increasing attention by the industrial society to alleviate the unnecessarily costs and potential toxicity associated with the traditional Hyaluronic acid isolation from rooster combs (Long Liu et al., 2011).

Until now, the continuous rise in infectious diseases and development of microbial resistance has provoked the need for novel antimicrobial compounds. The use of metagenomics for the isolation of novel antimicrobial compounds from different environmental resources have proven to be an effective approach as it has been reported in earlier studies (Gillespie et al., 2002). We found an ORF that potentially encodes for a antimicrobial peptide that showed homology to well characterized alpha-helical antimicrobial peptide (Maganin) isolated from an African clawed frog *Xenopus laevis* (ORF#15, Phmmer.pl.). Antimicrobial peptides (AMPs) represent a diverse set of ubiquitous molecules with different biological functions. The structural properties and their proposed specific mode of action against microbes, which evade possible resistant mechanisms, are amongst the factors that made them a promising alternative to currently available antibiotics (Pushpanathan et al., 2013). Many AMPs were reported in literature as antibacterial, antifungal, immune-modulatory and anticancer peptides. For instance, a novel antifungal peptide was identified from marine metagenome through functional metagenomic approach, and then its antifungal activity and mode of action against most common fungal infections were further assessed (Pushpanathan et al., 2012, 2013).
Chapter 3: Concluding Remarks

The continuous need for novel gene inventories with unusual biochemical characteristics will hold the attention of the scientific community towards mining exotic natural microbial habitats that were early thought to be deemed of any forms of life.

In this study, a 28.378 kb DNA fragment of recombinant fosmid clone from the deep brine environment of Atlantis II in the Red Sea was characterized. The Shotgun sequencing of 469 subclones of the recombinant 99B6 fosmid followed by assembly process, have produced 28.387 Kb of a contiguous DNA fragment encompassing 39 ORFs.

Out of the 39 ORFs identified, 10 (26%) have no match in the database, whereas at least 8 out of the 39 ORFS, according to literature, were shown to have potential for biotechnological applications.

Using restriction endonuclease analysis, it was estimated that the recombinant fosmid 99B6 has an insert of around 33.819 kb indicating that 5.441 kb was not assembled at the end of the 28.378 kb contig.

Analyses of the end of the assembled contig show a high degree of repetitions (simple repeats and gene duplication) that most probably explain the difficulties of recruiting more reads at this end.

Despite high costs and laborious process, the use of Sanger sequencing procedure for recovery of long de-novo contiguous assemblies yet represent the method of choice for obtaining accurate gene prediction and annotation.
Chapter 4: References


Pushpanathan, M., Rajendhran, J., Jayashree, S., Sundarakrishnan, B., Jayachandran, S., & Gunasekaran, P. (2012). Identification of a Novel Antifungal Peptide with Chitin-
doi:10.2174/092986612803521620


