

American University in Cairo

## AUC Knowledge Fountain

---

Theses and Dissertations

---

6-1-2014

### Functional identification of a Ligase in the Red Sea Atlantis II deepest Layer

Mahera Mohammed

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

---

#### Recommended Citation

##### APA Citation

Mohammed, M. (2014). *Functional identification of a Ligase in the Red Sea Atlantis II deepest Layer* [Master's thesis, the American University in Cairo]. AUC Knowledge Fountain.  
<https://fount.aucegypt.edu/etds/1189>

##### MLA Citation

Mohammed, Mahera. *Functional identification of a Ligase in the Red Sea Atlantis II deepest Layer*. 2014. American University in Cairo, Master's thesis. *AUC Knowledge Fountain*.  
<https://fount.aucegypt.edu/etds/1189>

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](mailto:mark.muehlhaeusler@aucegypt.edu).



The American University in Cairo

School of Science and Engineering

## **Functional Identification of a Ligase in the Red Sea Atlantis II Deepest Layer**

A Thesis Submitted To

The Biotechnology Graduate program

In Partial Fulfillment of the requirements

For the degree of Master of Science

By: Mahera Mohammed Ahmed

Under the Supervision of:

Dr. Rania Siam

May / 2014

The American University in Cairo

**Functional Identification of a Ligase in the Red Sea  
Atlantis II Deepest Layer**

A Thesis Submitted by

Mahera Mohammed Ahmed

To the Biotechnology Graduate Program

May / 2014

In partial fulfillment of the requirements for  
The degree of Master of Science

Has been approved by

Thesis Committee Supervisor/Chair \_\_\_\_\_

Affiliation \_\_\_\_\_

Thesis Committee Reader/Examiner \_\_\_\_\_

Affiliation \_\_\_\_\_

Thesis Committee Reader/Examiner \_\_\_\_\_

Affiliation \_\_\_\_\_

Thesis Committee Reader/External Examiner \_\_\_\_\_

Affiliation \_\_\_\_\_

\_\_\_\_\_  
Dept. Chair/Director

\_\_\_\_\_  
Date

\_\_\_\_\_  
Dean

\_\_\_\_\_  
Date

”It is not the strongest of the species that survives, nor the most intelligent that survives, it is the one that is the most adaptable to change”

Charles Darwin (1809-1892)

## **DEDICATION**

This thesis is dedicated with all my love and respect to my family, especially my Mother and my Father for being there for me throughout the entire master program and whose words of encouragement and support ring in my ears. To my sister Maram, who has never left my side and is very special.

I also dedicate this dissertation to my many friends who have supported me throughout the process. I will always appreciate all they have done.

Last but not least, I dedicate this work and give special thanks to my loving Husband Mohammed, His hours in caring for our wonderful daughter, Lana, enabled the hours of research and writing necessary to complete this project. Both of them are very dear to my heart.

## ACKNOWLEDGEMENTS

It would not have been possible to write this thesis without the help and support of all the people around me.

First the person who influenced me the most in my graduate career has been my advisor Dr. **Rania Siam**, Associate Professor and Chair of the Biology Department in the School of Science and Engineering at the American University in Cairo. I am especially grateful for her faith in me and her guidance over the past four years. She goes so far to make sure students are prepared for whatever the next step in their journeys may be.

I am thankful to my thesis committee members. Thanks for Dr. **Wael Mohammed**, Visiting Assistant Professor in the Biology Department at the American university in Cairo, Dr. **Ramy Aziz**, Assistant Professor in the Microbiology and Immunology Department at the Faculty of Pharmacy Cairo University and Dr. **Ahmed Abdellatif**, Visiting Assistant Professor in the Biology Department at the AUC, Thanks to all of them for agreeing to serve in my committee.

I would like to thank all the professors in the department who taught me through my master program (Dr. **Ahmed Mostafa**, Dr. **Ahmed El-Sayed**, Dr. **Walid Fouad** and Dr. **Asma Amleh**), their feedback made the completion of this research an enjoyable experience and also I would like to thank the administrators in our school division for providing any assistance requested.

I acknowledge the King Abdullah University of Science and Technology (KAUST) for funding the Red Sea Metagenomics project and for the graduate research fellowship that provided the necessary financial support for this research. Amongst the KAUST Red Sea spring 2010 expedition team: Dr. **Rania Siam**, Dr. **Mohamed Ghazy** and Mr. **Amgad Ouf**.

I have had the great pleasure of mentoring on my project from **Nahla Hussien**, a fellow PhD student. I am grateful to her for helping me to learn that there is more than one way to approach a problem. She is a bright scientist, and I am sure she will be an amazing professor.

I would also like to give a special thanks to my colleagues and friends in the American University in Cairo. Through all of the ups and downs that we spend together, we were always there for each other. I am especially grateful to **Hadeel El-Bardisy** and **Aya Medhat** who I had

the great fortune of becoming close friends. They were both willing to talk endlessly with me about my research. I must mention all the support I had from my lab mates who were always kind enough to answer my loads of questions **Salma El Shafie, Sarah Sonbol, Bothaina El-Laimoni, Dina Hassan, Yasmeen Moustafa, Ali Elbehery, Sarah Kamel, Laila Ziko, Ayman Yehia, Rehab Adelallah, Ghada Mostafa, Mustafa Adel, Mariam RizkAllah, Yasmeen Howeedy and Mohamed Maged** ;) Thank you all so much.

I thank my lifetime friends outside the AUC, the most loyal friends a person could ask for, **Mai Safan, Menna Shawky, May Gamal, Noha Fouad, Aya mostafa, Deena Jalal, Zeinab Saad El-Din**. You girls have all been truly amazing friends.

Last, but by no means least, to thank the people who made me into the person who I am now, my **Mom** and **Dad**, I love you both and I wish you all the happiness. My sister and my best friend **Maram**, thank you for your support through my endless favors. My **Grandmother** for her prayers and encourage and definitely my late grandparents who did not have the chance to share these moments with me. I miss you both. To my uncle **Mostafa**, his wife **Nermeen** and their lovely kids **Ahmed** and **Farida**, for always being a great supportive family.

Finally, My dearest husband **Mohammed**, you are my one and only, I could not have done all of this without you by my side and my precious daughter **Lana**, I have done this for you to be proud of your mother. Both of you and your father were my best cheerleaders :) I love you.

# Abstract

The American University in Cairo

## **Functional Identification of a Ligase in the Red Sea Atlantis II Deepest Layer**

By Mahera Mohammed Ahmed  
Under the supervision of Dr. Rania Siam

---

Red sea, described as one of the unique marine ecosystems, incorporates up to 25 deep-sea brine pools. These pools possess multiple extreme conditions influencing the evolution and survival of their inhabiting microbial community. The combination of maximum depth (2194 m), high temperature (68 °C), anoxia, high salinity (26%), high pressure and high concentrations of heavy metals in the lower convective layer (LCL) of the Atlantis II brine pool makes it an ideal environment for identification of novel enzymes with unique characteristics and potential biotechnological applications.

Here we describe the identification and the preliminary *in vivo* functional investigation of the ligase domain of an ATP-dependent DNA ligase from the DNA of the prokaryotic community extracted from water samples of the LCL of Atlantis II brine pool. Previously, these water samples were serially filtered on different membranes and the DNA isolated from the 0.1µm filter was subjected to 454 pyrosequencing. A metagenomic dataset was initiated and used in this study to mine for genes encoding DNA ligases through Pfam search of conserved domains. The search and subsequent bioinformatic analysis resulted in the identification of a contig harboring an ORF of 915 bp (305 amino acids) that encodes a putative DNA ligase (LigATII). Homology search of the putative DNA ligase showed highest similarity to *Erysipelotrichaceae Bacterium* (39% identity, 54% positive). LigATII displays modular architecture that is similar to two distinct domains-(the adenylation domain of *LigD* and the oligonucleotide binding (OB) fold domain)-that are conserved in ATP-dependent DNA ligases.

Functional annotation of the LigATII ORF, identification of the functional conserved amino acids by the Consurf tool, 3D modeling and comprehensive phylogenetic analysis were conducted. These analyses have revealed the relatedness of LigATII to the family of ATP-dependent DNA ligases that has been recently identified through computational studies to exist in prokaryotes. This family is expected to be involved in the specialized form of genomic DNA repair through the non-homologous end joining pathway which acts to join double-stranded breaks (DSBs) or to promote genetic diversity under conditions of selection pressures.

Accordingly, the putative LigATII was amplified from the whole genome DNA amplification of LCL. Sanger sequencing confirmed the sequence of the gene before cloning into pET100 Topo directional expression vector. The cloned LigATII was transformed into a temperature sensitive mutant strain of *Escherichia coli*; strain GR501, with mutation in the DNA ligase gene. LigATII complemented the temperature sensitive strain at the non-permissive temperature (43°C) verifying the *in vivo* functional activity. The biochemical characteristics of the novel LigATII protein will be described.

---



## Table of contents:

<i>List of Figures:</i> .....	<i>xi</i>
<i>List of Tables:</i> .....	<i>xii</i>
<i>List of Abbreviations:</i> .....	<i>xiii</i>
<b>Chapter 1: Literature Review</b> .....	<b>1</b>
<b>1. The Red Sea unique ecosystem:</b> .....	<b>1</b>
1.1 Hydrothermal system: .....	1
1.2 Atlantis II brine pool: .....	2
<b>2. Marine Metagenomics:</b> .....	<b>3</b>
2.1 Genetic diversity of microbial communities: .....	4
2.2 Direct sequencing: .....	5
2.3 Shotgun sequencing: .....	6
<b>3. Mining for unique biocatalysts in Metagenomes of extreme environments:</b> .....	<b>6</b>
3.1 Construction of metagenomic libraries: .....	8
3.2 Screening of constructed libraries: .....	8
3.2.1 Sequence-based Screening: .....	9
3.2.2 Function-based screening: .....	9
3.3 Sequence analysis of genes: .....	10
3.4 Expression of genes of interest in appropriate host: .....	11
<b>4. Industrial Biocatalysts from Metagenomes:</b> .....	<b>11</b>
<b>5. DNA Ligase Enzymes and genome repair:</b> .....	<b>13</b>
5.1 Mechanism of action: .....	14
5.2 Ligase enzymes, a branch of the nucleotidyltransferase superfamily: .....	15
5.3 Discovery of NHEJ in bacteria: .....	16
5.4 Sequence Conservation among DNA ligases and the three dimensional structure: .....	19
<b>6. Biotechnological applications of DNA Ligase:</b> .....	<b>22</b>
6.1 Ligase chain reaction (LCR): .....	22
6.2 ELISA: .....	23
6.3 Multiplex LCR: .....	24
6.4 Quantitative real time LCR: .....	24
6.5 Colorimetric nanoparticle based method: .....	24
<b>Chapter 2: Materials &amp; Methods</b> .....	<b>25</b>
<b>1. Sample collection:</b> .....	<b>25</b>
<b>2. DNA Extraction, Sequencing &amp; Construction of the 454 metagenomic database:</b> .....	<b>25</b>
<b>3. Computational analysis:</b> .....	<b>26</b>
3.1. Screening of the ATII-LCL metagenomic database for DNA ligase: .....	26
3.2. Functional Annotation of the resulting contigs: .....	26

3.3. Prediction of the upstream regulatory regions of LigATII: .....	27
3.4. Conserved Domain search of the LigATII: .....	27
3.5. Multiple sequence alignment of LigATII: .....	27
3.6. Identification of functional regions in the LigATII: .....	27
3.7. Phylogenetic analysis for the LigATII: .....	28
3.8. Comparative homology modeling of LigATII: .....	28
3.9 Prediction of the molecular weight and isoelectric point of LigATII: .....	29
<b>4. Isolation of the putative LigATII:.....</b>	<b>29</b>
4.1 PCR based screening method:.....	29
4.2 Cloning and sequencing of the PCR product: .....	29
<b>5. Construction of LigATII expression systems: .....</b>	<b>30</b>
5.1 Expression of LigATII gene using pET SUMO® Expression System: .....	30
5.2 Expression of LigATII gene using Champion™ pET100 Directional TOPO® Expression System:.....	31
5.3 Induction of expression of LigATII: .....	32
<b>6. Functional Identification of the LigATII in the <i>E. coli</i> GR501:.....</b>	<b>33</b>
6.1 Assay for LigATII complementation of temperature sensitive defect of <i>E. coli</i> GR501 on agar plates: .....	34
6.2 Assay for strain viability of the <i>E. coli</i> GR501 by the LigATII.....	34
complementation of the temperature sensitive defect: .....	34
6.3 Assay for LigATII complementation of temperature sensitive defect of <i>E. coli</i> GR501 in liquid cultures: .....	34
<b>Chapter 3: Results &amp; Discussion .....</b>	<b>35</b>
<b>1. Search for DNA ligase in the ATII-LCL Metagenomic Database: .....</b>	<b>35</b>
<b>2. Identification of LigATII:.....</b>	<b>35</b>
2.1 Sequence Analysis of LigATII using BLASTx server:.....	35
2.2 Prediction of the upstream regulatory regions and ribosomal binding site of the LigATII: .....	36
2.3 Search for conserved domains within LigATII:.....	38
2.4 Multiple sequence alignment of LigATII: .....	38
2.5 Identification of functional regions in LigATII using ConSurf: .....	40
2.6 Phylogenetic analysis of LigATII: .....	41
2.7 Comparative homology modeling of LigATII: .....	44
2.8 Molecular weight and isoelectric point prediction: .....	46
<b>3. Isolation of the LigATII enzyme from the DNA of the Atlantis II brine pool: .....</b>	<b>46</b>
3.1 Screening the 1X amplified environmental DNA of the LCL of Atlantis II brine pool: .....	46
3.2 Cloning and sequencing of the PCR product: .....	48
3.3 Expression of LigATII Gene in Champion™ pET SUMO® expression system: .....	48
3.4 Expression of LigATII Gene in Champion™ pET100 Directional TOPO® expression systems:.....	50
<b>4. Functional Identification of the LigATII in the temperature sensitive mutant <i>E. coli</i> (GR501): .....</b>	<b>53</b>
4.1 Assay for LigATII complementation of temperature sensitive defect of <i>E. coli</i> GR501 on agar plates: .....	54
4.2 Assay for strain viability of the <i>E. coli</i> GR501 by the LigATII complementation of the temperature sensitive defect:.....	56

4.3 Assay for LigATII complementation of temperature sensitive defect of <i>E. coli</i> GR501 in liquid cultures:	57
<b><i>Chapter 4: Conclusion &amp; Prospective .....</i></b>	<b>58</b>
<b><i>References: .....</i></b>	<b>60</b>

## List of Figures:

Figure (1): Tectonic activity creating high temperature and saline Red sea brine Pools .....	2
Figure (2): Bathymetric map showing the location of Atlantis II brine pool in the Red sea .	3
Figure (3): Strategies for metagenomic detection of biocatalysts.....	7
Figure (4): Mechanism of action of DNA ligases in the presence of ATP cofactor .....	15
Figure (5): The mechanism of NHEJ in prokaryotes .....	18
Figure (6): Alignment of conserved sequence elements among Bacterial DNA ligases.....	20
Figure (7): the best hit of blastx aligned with LigATII .....	36
Figure (8): Nucleotide sequence and predicted amino acid sequence of the LigATII ORF	37
Figure (9): The conserved domain identified within the LigATII.....	38
Figure (10): Multiple sequence alignment of LigATII .....	39
Figure (11): Identification of functional regions in LigATII on the ConSurf server .....	40
Figure (12): Phylogenetic analysis tree for LigATII.....	42
Figure (13): Structural superimposition of LigATII with Bacteriophage T7 DNA ligase ...	45
Figure (14): Predicted binding site of LigATII with ATP.....	45
Figure (15): PCR product of a single read from LigATII amplified from the 1 X WGA DNA of Atlantis II LCL.....	47
Figure (16): PCR product of LigATII ORF amplified from the WGA DNA.....	47
Figure (17): SDS-PAGE for the analysis of the expression level of LigATII-pET SUMO after induction for 2 hours at 37°C .....	49
Figure (18) SDS-PAGE for the analysis of the induced culture of LigATII-pET SUMO after cell lysis.....	49
Figure (19): PCR product from the amplification of ORF of LigATII .....	50
Figure (20): SDS-PAGE for the analysis of the expression level of LigATII-pET 100 Directional TOPO® .....	51
Figure (21): Growth complementation of <i>E. coli</i> GR501 streaking on LB/ Amp plates .....	55
Figure (22): Growth curves of liquid culture of <i>E. coli</i> GR501 .....	57

**List of Tables:**

<b>Table (1): <i>E. coli</i> GR501 strain viability at 30 °C and 43 °C .....</b>	<b>56</b>
--	-----------

## List of Abbreviations:

ATII	Atlantis II brine pool
LCL	Lower convective layer
KAUST	King Abdullah university of science and technology
rDNA	Recombinant Deoxyribonucleic acid
SSU	Small ribosomal ribonucleic acid subunit
BAC	Bacterial artificial chromosome
NGS	Next generation sequencing
HGP	Human genome project
SIGX	Substrate induced gene expression screening system
gfp	Green fluorescent protein
NCBI	National center for biotechnology information
BLAST	Basic local alignment search tool
IR	Ionizing radiation
DSB	Double strand breaks
HR	Homologous recombination
NHEJ	Non Homologous end joining
ATP	Adenosine triphosphate
NAD	Nicotinamide adenine dinucleotide
NT	Nucleotidyl transferase
OB	Oligonucleotide binding
Lys	Lysine residue
SNP	Single nucleotide polymorphism
LCR	Ligase chain reaction
ELISA	Enzyme-linked immunosorbent assay
CTD	Conductivity, Temperature and Depth
WGA	Whole genome amplification
ORF	Open reading frame
LDF	Linear discriminant function
CDD	Conserved domain database
ML	Maximum Likelihood
IPTG	Isopropyl $\beta$ -D-1-thiogalactopyranoside
SDS-PAGE	Sodium dodecyl sulfate – polyacrylamide gel electrophoresis
ts	Temperature sensitivity

# Chapter 1: Literature Review

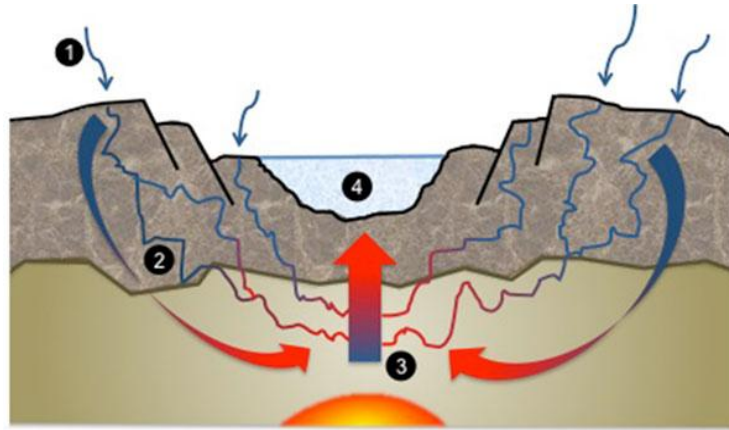
---

## 1. The Red Sea unique ecosystem:

Since the discovery of the first oceanic hydrothermal activity in 1966, attention has been drawn toward the studies that focus on the chemical and physical processes that are happening before and control the ejection of hydrothermal fluid from those vents into the ocean <sup>1</sup>, meanwhile limited studies focused on the impact of these ecosystems on the microbial inhabitants of the hot brines as in the Red sea due to technical constraints <sup>2</sup>. The Red sea is one of the unique ecosystems compared to other marine environment owed to the geochemical and physical parameters of the high rate of evaporation, low level of precipitation and lack of major river inflows. These unique characteristics very much increase the temperature and salinity creating unusual and extreme environment with anaerobic, hypersaline, hyperthermal and metalliferous conditions that characterize about 25 of deep-sea brine pools discovered to date among the four geotectonic regions (southern, central, multi deeps and northern regions <sup>3,4</sup>) representing the ocean basin of the Red Sea <sup>5</sup>.

### 1.1 Hydrothermal system:

Red Sea deep water is heated when it penetrates into fissures in the sea floor, formed from localized tectonic activity. This water circulates through the hydrothermal system and with boiling convective currents it is driven back to the seafloor surface where it becomes enriched with metals leaching from the thick Miocene (NaCl, CaSO<sub>4</sub>) evaporitic beds <sup>6,7</sup>. Therefore salty dense fluids of the Red Sea are trapped in the deeps and their metals are precipitated as layers within the sediments these combined with double diffusive convection induces brine distinctive layering against the surrounding seawater giving the impression of a submarine pool <sup>8</sup>(figure 1).



**Figure (1): Tectonic activity creating high temperature and saline Red sea brine Pools**

[1] Tectonic activity form fissures in the seabed sediment through which the water of the deep-sea penetrates. [2] The submarine magma heats the water after sinking in, which in turn slowly circulates and absorbs minerals during the process. [3] The water becomes heated and driven back to the seafloor surface by convective current. [4] The effect of a submarine pool is given by the dense brine forming a distinctive layer against the surrounding seawater<sup>8</sup>.

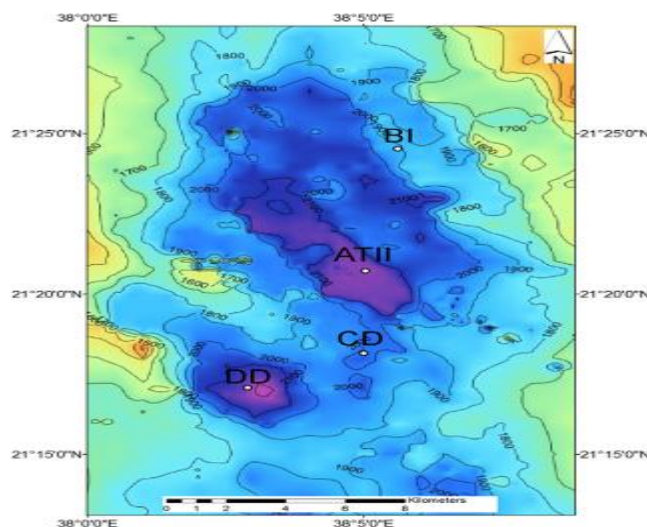
## 1.2 Atlantis II brine pool:

One of the largest, hottest and widely studied hot brines in the Red sea is the Atlantis II Deep (ATII) which is located at a depth of approximately 2200 m near the Central Rift<sup>2</sup> (figure 2), it is about 100 m thick, and covers an area of 60 km<sup>2</sup>. The brine is stratified into layers, with the bottom layer (lower convective layer (LCL)) exhibiting the highest temperature. The temperature of the LCL increased to almost 70<sup>0</sup>C, indicating considerably strengthened hydrothermal activity<sup>9</sup>, according to the observation of the King Abdullah university of science and technology (KAUST) Red Sea Expedition Fall 2008<sup>8</sup>. This expedition also observed a new convective layer UCL4 developing above the three upper convective layers (UCL1, 2 and 3) that display step-wise drops in salinity and temperature<sup>2,10</sup>.

The characteristic anoxia, high pressure, increased temperature, pH value of 5.3 and salinity (250 parts per thousand or 7.5 times that of normal seawater)<sup>2</sup> makes the Atlantis II Deep brine unique and ideally suitable for the study of



extremophiles that produce a wide set of extremozymes<sup>11</sup>. These extremozymes are stable under conditions used during a lot of industrial processes or analytical methods that normally leads to the denaturation of most proteins<sup>7,12</sup>. One example of these enzymes are the group of DNA modifying enzymes, which are predicted to possess desirable properties, including the thermotolerability<sup>13,14</sup>.



**Figure (2): Bathymetric map showing the location of Atlantis II brine pool in the Red sea**

The Atlantis II Deep area (AT II) (between latitudes 21° 13 ' N and 21° 30 ' N and longitudes 37° 58 ' E and 38 ° 9 ' E) is in the central rift zone of the Red Sea, between Saudi Arabia and Sudan<sup>2</sup>.

## 2. Marine Metagenomics:

Marine environments represented in oceans cover almost 70% of the earth's surface and although they are estimated to contain large pool of microbial biodiversity defining the chemistry of these oceans, they are not made full use of yet<sup>15</sup>. However, in recent years due to the growing acknowledgement of the importance of marine ecosystems in human life, interest in marine biology studies have been speeded up to understand the biology of microorganisms adapted to live in these challenging environments. This is through the great influence of the intersecting field of molecular biology that entered the mainstream of biological

thinking in the 1940s, when most molecular and genetic analysis was tied to microorganisms<sup>16</sup>. In the early 90's the genomic studies were leaning on traditional culture based techniques to study individual isolated microbes, which so far added limited information about as many as 99% of the prokaryotes that are not currently culturable in 'unnatural' laboratory conditions and are however dominating the environmental samples<sup>15,16</sup>.

### 2.1 Genetic diversity of microbial communities:

Later, in the past two decades culture independent phylogenetic studies were in the process of becoming larger in order to describe the diversity of microbial communities based on 16S rRNA highly conserved sequence as a marker gene for phylogenetic analysis<sup>17,18</sup>. Where in a given complex sample, PCR based techniques are used to amplify 16S rRNA gene from the bacterial community DNA extracted then cloned in host e.g. *Escherichia coli* (*E. coli*) to construct a clone rDNA library followed by sequencing to develop a database that gives an immense resource for genomic information, like the work done for environmental DNA sequencing of the Sargasso sea<sup>18,19,20</sup>. These techniques averting the requirement for cultivation, have been applied and further more extended to determine the genetic diversity inferring phylogenetic relationships among microorganisms in ecological contexts, these microorganisms require rather difficult conditions that mimic their natural habitat for proliferation during cultivation<sup>21,22</sup>.

So this ecogenomics or microbial population genomics defined as the collective genome of the total microbiota of a specific environment was given the term "Metagenome" by Handelsman and colleagues in 1998<sup>23</sup> and the term "Metagenomics" was used to describe the genome-based analysis of entire communities of microorganisms having impact on each other within a diverse ecological contexts<sup>16</sup>.

To facilitate the access to hereditary material of microorganisms in marine ecosystems, technology for DNA recovery had to be revolutionized toward advancement from the usual methods linked to small ribosomal RNA subunit (SSU) hypervariable regions amplification. In this method probing for ribosomal RNA genes to isolate and characterize other genes physically linked and associated with the phylogenetic targets <sup>24</sup> is done as in the example of its application in the Stein et al. study <sup>25</sup>. Also methods that utilizes bacterial artificial chromosomes (BACs) can yield reads up to 1 kb from a single reaction of the Sanger chain termination/end sequencing of subclones from large DNA fragments (20–400 kb) in an *Escherichia coli* recombinant library, like in the example of Wang et al. BAC library <sup>26</sup>. However, these sequence-based methods were subject to limitations imposed on all Sanger sequencing-based experiments (slow, low throughput and expensive) <sup>27</sup>.

### 2.2 Direct sequencing:

The sequencing methods development was achieved with potential to improve reference databases several orders of magnitude greater than previously achieved, facilitating the exploration of the vast diversity of microbial populations in the human gut, soil and oceans <sup>28</sup> and the rare species within them. This is achieved through a rapid and direct sequencing approach with the low cost per read and high throughput of massively parallel pyrosequencing. This method can provide enormous amounts of short reads (1 million) with an average read length of up to 400 bp in a single 10-h run, Which will uncover more organisms, overcomes cloning requirements and avoids assembly with a great advantage of sensitivity unmatched by conventional sequencing of SSU rRNA genes <sup>5,29</sup>. These Next-Generation Sequencing (NGS) platforms, such as Roche/454 FLX Pyrosequencer, Illumina Genome Analyzer, and Applied Biosystems SOLiD™ Sequencer, have had a tremendous impact on genomic research<sup>30</sup>.

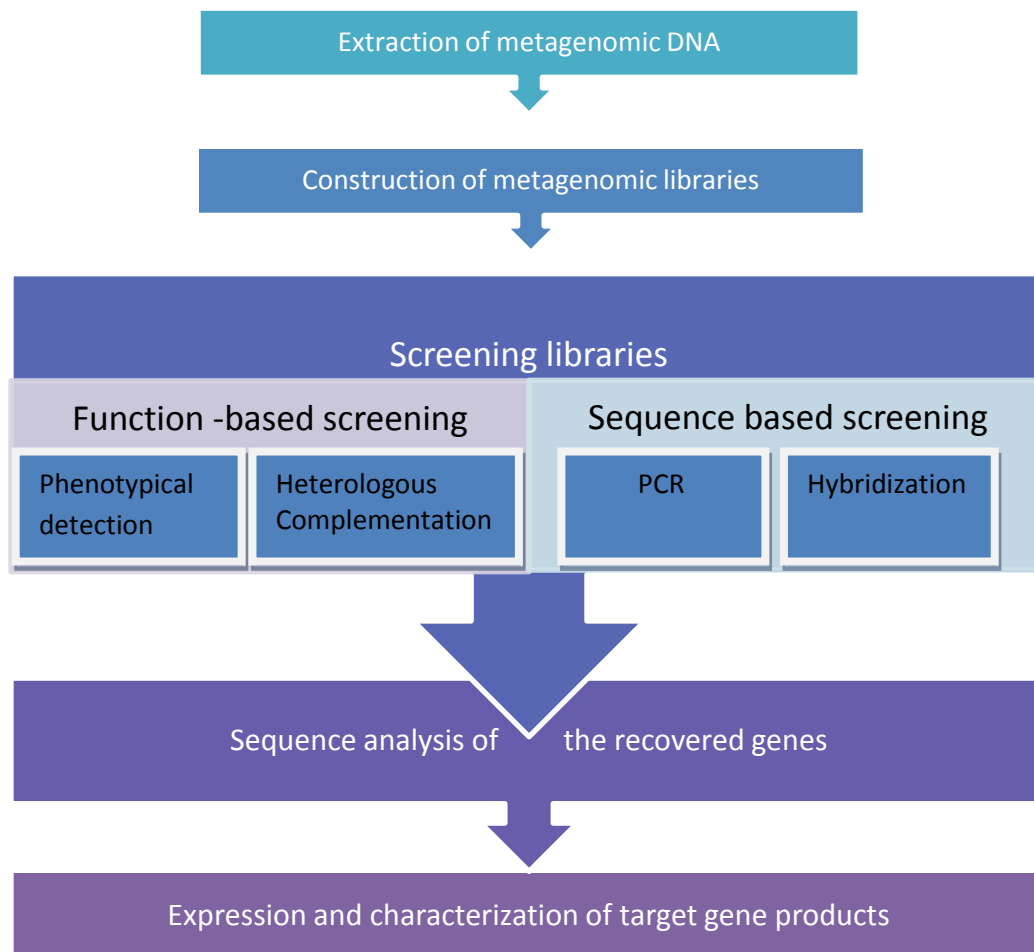
**2.3 Shotgun sequencing:**

The essence for the development of the massive parallel sequencing of the Next-Generation Sequencing (NGS) platforms was the Shotgun sequencing. In this sequencing approach developed during the Human Genome project (HGP), environmental DNA is isolated and mechanically subjected to fragmentation<sup>31</sup>. The generated small DNA fragments are cloned into vectors that will be individually and randomly sequenced by Sanger sequencing then subjected to extensive computational analysis to reassemble the fragments giving the overall consensus sequence. Not only to characterize and enumerate different taxa through protein-based phylogeny and mining for genes but also inferring to their biological functions through the identification of metabolic pathways; as in the study of the planktonic microbial community in the North Pacific Gyre<sup>32</sup>.

**3. Mining for unique biocatalysts in Metagenomes of extreme environments:**

The huge amount of information found within the genomes of uncultured microorganisms that are able to grow in a vast range of environments, like extreme conditions in hydrothermal vents, needed a key technology as metagenomics to allow the investigation of the wide diversity of individual genes and their biocatalysts. These biocatalysts are of great interest due to their potential applications in industry owed to their high stability. Metagenomics concerns with the extraction, cloning and analysis of the entire genetic complement of a habitat which majority of microorganisms are usually reluctant to cultivation although they are active. This overcomes past limitations to the wide spreading application of biocatalysts in industry that were traditionally only obtained from bacterial isolates<sup>33,34</sup>.

The work in Metagenomics field involves the four following stages (i) construction of metagenomic library, (ii) screening of metagenomic library, (iii) sequence analysis of genes, and (iv) expression of genes of interest in appropriate hosts, to produce the desired biocatalysts (summarized in figure 3).



**Figure (3): Strategies for metagenomic detection of biocatalysts**

The metagenomic process involves: construction of metagenomic library from the DNA extracted of the environmental samples and these libraries are subjected to screening for functional enzymes or subjected to sequence screening of the target genes. After the sequence analysis of the gene, expression of the gene product in appropriate host will facilitate the characterization of the desired biocatalyst.

### 3.1 Construction of metagenomic libraries:

As for the first stage, genes are collected from an environmental sample then with the help of several protocols construction of metagenomic libraries from the extracted DNA. A combination of various physical and chemical methods that many laboratories have worked out for their own studies are used with the aim of optimising extraction and reducing bias caused by unequal lysis of different members of the microbial community. In addition, attention is drawn to avoid severe DNA shearing and sample contamination to increase recovery and efficiency of molecular analysis preventing the formation of chimeric products or interference with downstream processes<sup>35,36</sup>.

The method of metagenomic library construction involves insertion of large inserts of approximately 40–200 kb in size through blunt-end or T–A ligation into large cloning vectors like cosmid, fosmids or BACs. This differs from traditional cloning of small insert in plasmid libraries and determined according to the design of the study. The first cloning library allows the detection of large gene clusters and operons, while the later can be an alternative in case of genes that exhibit low expression levels. The following process of the transformation of clones is generally engaged to *E. coli* as a host strain for the ease of genetic manipulation and downstream processes, despite the limitation of the inability to express specific genes from environmental microorganisms, to overcome this, a number of alternative cloning hosts such as *Bacillus subtilis*, *Pseudomonas spp.* or eukaryotic expression systems could be used<sup>37</sup>.

### 3.2 Screening of constructed libraries:

In the second stage, the metagenomic library constructed will need to be screened to detect transformants containing phylogenetic markers (usually a 16S rRNA gene which is not expressed into protein) through a sequence based screening in which PCR is employed, or detecting functional genes (coding for a corresponding protein/ biocatalyst) by a function based screening which can be done by assaying the activity of the expressed biocatalysts (enzymes)<sup>35</sup>.

### 3.2.1 Sequence-based Screening:

The sequence-driven screening approach is limited to polymerase chain reaction (PCR) based methods using primers or hybridization-based approaches using probes designed on the base of conserved regions of known genes and gene products. These approaches only identify new members of known gene families that harbor regions with similarity to the sequences of the probes and primers and this line to the reality of not detecting any new genes that should carry novel sequences undiscovered before. Besides, these approaches cannot recover or select for full-length genes and functional gene products unless genome walking is applied next. However the advantage of this screening approach is the freedom from production of foreign genes in heterologous host strain, also with the increase in collection of phylogenetic markers, more genes will be linked to phylogenetic markers and more genomes are being reconstructed<sup>36,38</sup>.

### 3.2.2 Function-based screening:

For this screening method to accomplish its intended purpose, it requires the use of an appropriate host that is able to express the transformed gene of concern in the surrogate host's heterologous background. Also this method requires the analysis of more clones than sequence-based screening for the recovery of the positive clones. However, the major advantage of this screening approach is the ability to identify new classes of full-length genes with known functions<sup>38</sup>.

There is a variety of strategies to detect the desired expression products (biocatalysts) but the simplest mainly is based on the use of indicator media containing certain substrates to cultivate the metagenomic library, which are then acted upon by the present enzyme and the reaction product release is verified by the presence of a clearing zone around transformants, as in the case of hydrolases. Other high-throughput strategies involve chromogenic or fluorogenic substrates usage to visually detect the converted reaction products *in vivo*<sup>35</sup>.

A different strategy is the use of host strains that lacks the target genes and requires heterologous complementation by functional foreign genes for growth under selective conditions; one example is the identification of functional DNA ligase-encoding genes <sup>39</sup>. In a study the *E. coli* GR501 mutant strain, which carries a temperature-sensitive lethal mutation in the *ligA*, was employed as host at a growth temperature of 43°C. Only recombinant *E. coli* strains complemented by a gene encoding the T4 or *E. coli* DNA ligase conferring DNA ligase-activity are able to grow.

The third function-driven strategy is based on induced gene expression, substrate-induced gene expression screening system (SIGEX) for the identification of novel catabolic genes. The gene for a promoter green fluorescent protein (gfp) will be constructed to be found in operon-trap expression vector in adjacent to the target gene, when expression of the target gene is induced by the substrate, the *gfp* gene will be coexpressed, and positive clones can rapidly be separated from other clones by the fluorescent-activated cell sorting technique <sup>38</sup>.

Nevertheless, significant differences in the predicted expression modes between distinct groups of organisms suggest that only 40% of the enzymatic activities may be recovered by random cloning in *E. coli* due to biased usage of codons. As in the case of the AUG codon, in the translation initiation, it is favored over GUG and UUG, commonly used by most organisms. And despite the several trails in different approaches to overcome the failure of the host's genetic machineries to recognize transcriptional/translational signals in the metagenome formed from a variant array of microorganisms, however getting a hit from the metagenome relays mainly on luck <sup>40</sup>.

### 3.3 Sequence analysis of genes:

As for the third stage of the metagenomic study, it involves the characterization of the biocatalysts through bioinformatic analysis of the amino acid sequences to predict the biocatalyst's properties taking advantage of



bioinformatic softwares. Most of those softwares can be found freely in public databases, such as National Center for Biotechnology Information (NCBI, <http://www.ncbi.nlm.nih.gov/>). The sequences coding for biocatalyst of interest can be subjected to BLAST (Basic Local Alignment Search Tool) to search for homologous sequences in databases, these significant homologous sequences are aligned with the query sequences through the Clustal Omega program to give information about the structural and functional features of target biocatalysts <sup>35</sup>.

### **3.4 Expression of genes of interest in appropriate host:**

The fourth and final stage of the process of metagenomics implicate the expression of such recombinant genes in a host characterized by the ability to grow easily in inexpensive media, genetically well characterized, availability of compatible cloning vectors, lack of restriction enzymes that affects downstream processes and the variety of mutant strain related to the host. That is why *E. coli* appears to be the most suitable host as a microbial system for gene expression <sup>35</sup>.

During the process of expression, different factors affecting the expression level have to be taken in consideration these include: features of the target gene sequence, the efficiency of transcription of mRNA, correct protein folding by the suitable chaperones, degradation of the recombinant protein by proteases, preferential codon usage and last but not least whether the protein is toxic to the host cell or not <sup>41</sup>.

### **4. Industrial Biocatalysts from Metagenomes:**

Different industries are now more interested in exploiting the resource of uncultivated microorganisms that has been identified through the potentials of metagenomics, through exploration of large-scale environmental genomics utilizing the appropriate screening tools to access novel enzymes and biocatalysts that can impact industrial production. These industrial process use enzymes in a

wide range of applications only in minute quantities to synthesize kilograms of challenging products and these “ideal enzymes” are required to function well according to specific performance parameters<sup>37</sup>. Therefore in designing aged industrial application, enzymatic constraints limited these processes and it is more thinkable now to evolve *in vitro* technologies that benefit from natural enzymes of the uncultivated microbial diversity, to fit industrial process requirements.

Following the report of the first metagenomic library prepared directly from environmental soil DNA. Research focused on the enriched consortia for screening and expression of biocatalysts useful for industrial and biotechnological applications, example the group of DNA modifying enzymes, to benefit from the fact that most environmental bacteria are rebellious to cultivation but will be accessible by direct cloning of DNA fragments collected from the environment<sup>37</sup>.

Screening metagenomic libraries depends strongly on the final application of the biocatalyst, whether the level of enzyme activity is important, or their stability under a wide range of conditions is important. The existence of enzymes with these unique characteristics have only been possible because of the vast unlimited likelihood of prokaryotes to adapt and flourish in every environment, from hydrothermal vents on the ocean such as the Atlantis II brine pool in this study to acid mine drainage sites. And here comes the importance of metagenomic studies to emphasis on improving the range of tests aimed for the biochemical characterization of novel biocatalysts as well as improve screening and detection strategies taking into consideration the final application of the biocatalysts and its specific requirements in the biotechnology processes<sup>33</sup>.

Still to remain are the drawbacks of metagenomics, that lies not only in the expensive data analysis, assembling reads, chimeric sequences production and the detection of rare communities. But also that only very restricted number of enzymes obtained through metagenomics are used in any of the founded

biotechnological processes. Moreover metagenomics is facing quite challenges to uncover new functions encoded by the already identified enzymes that still yet to be detected and executed in industrial production processes<sup>34</sup>.

### 5. DNA Ligase enzymes and genome repair:

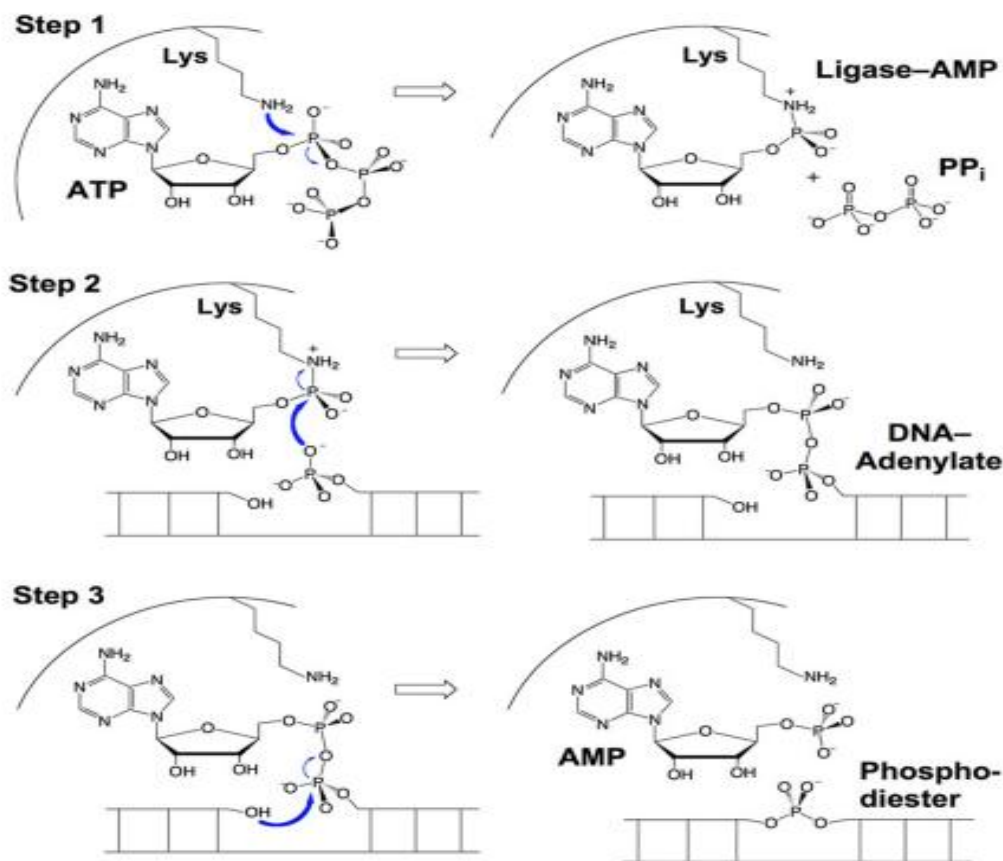
Breaks in the DNA molecule are frequent and generated by a variety of reasons. Whether during necessary processes like DNA replication and recombination, endogenously from reactive oxygen species, or even from exogenous DNA damaging agents as ionizing radiation (IR)<sup>42</sup>. One of the most lethal forms of these breaks is the double-strand breaks (DSBs) because its accumulation leads to gross chromosomal rearrangements and ultimately, cell death<sup>42,43</sup>. This requires enzymes to keep the genome integrity by sealing DNA breaks<sup>44,45</sup>. Previous reports describing conditional lethal mutants of *Escherichia coli* with mutations in the ligase gene, have confirmed the importance of this essential enzyme, due to deficiency in both repair and replication of DNA<sup>46,47</sup>.

Therefore, DNA ligases are described as a group of rather important enzymes to manage genome repairs through two major DSB-repair pathways: HR (homologous recombination) and NHEJ (non-homologous end-joining)<sup>43,48</sup>. In HR, an intact copy of the broken chromosome segment serves as a template for DNA synthesis at the break site and faithfully guides this repair pathway in the dividing cells. Conversely, the NHEJ pathway directly joins and ligates the two identified ends of the break after a processing stage if required<sup>42,43,48</sup> and the term “non-homologous” was assigned to this pathway because no homology is required<sup>49</sup>. This means that the NHEJ can operate in situations when only one chromosomal copy is available as in the G<sub>1</sub> phase of the cell cycle or as a repair pathway for the DSBs induced in dormant spores as a defense mechanism for survival in difficult environmental stresses<sup>48,50</sup>.

Mainly the NHEJ pathway in eukaryotes requires core constituents to aid in the process of the break ends recognition, approximation and stabilization through the Ku70/80 heterodimer that will recruit the specified ligase complex for the process of end sealing<sup>42,43</sup>. Additional factors may also be required to aid in the processing of the break termini to restore the complementary ends. These factors include polymerases ( $\mu$ ,  $\lambda$ , and Pol4), nucleases, polynucleotide kinase and phosphatase<sup>42,43</sup>. Extending the ends by polymerases or restricting them by nucleases prior to sealing, results in mutagenic NHEJ unlike the HR which is mainly error free. The mutagenesis generated in the DNA genome from NHEJ is considered rather an advantage, especially in situations where the introduction of new nucleotide is evolutionary crucial under the selection pressures to promote genetic diversity<sup>49</sup> or when death is the only alternative for the organism<sup>48,51</sup>.

### 5.1 Mechanism of action:

The mechanism of nick sealing, which is common for all members of the DNA ligase family, can be divided into three steps. First, in the absence of the DNA substrate, an activated Ligase – nucleotide mono phosphate adduct is generated by the transfer of the adenyl group of a cofactor to an active site of a lysine residue on the enzyme forming a phosphoamide bond with the release of an inorganic pyrophosphate ( $PP_i$ ) or a nicotinamide mononucleotide (NMN). In the second step, the 5' phosphate group of the adenyl group on the enzyme forms a pyrophosphate linkage with the 5' end phosphate at the single-strand break site of the DNA substrate to activate it for the third step of the phosphodiester bond formation by the attack of the 3' hydroxyl on the adjacent part of the nicked strand on the 5' phosphorylated end to release the adenyl group and covalently join the DNA Strands. All three chemical steps depend on a divalent cation cofactor<sup>52,53,54</sup> (figure 4).



**Figure (4): Mechanism of action of DNA ligases in the presence of ATP cofactor**

Step1: The transfer of the adenyl group of the ATP cofactor to an active site of a lysine residue on the enzyme generates an activated Ligase – nucleotide mono phosphate (AMP) adduct with the release of an inorganic pyrophosphate ( $PP_i$ ). Step2: The 5' phosphate group of the adenyl group on the enzyme forms a pyrophosphate linkage with the 5' end phosphate at the single-strand break site of the DNA substrate to activate it and generate DNA-adenylate intermediate. Step3: The attack of the 3' hydroxyl on the adjacent part of the nicked strand on the 5' phosphorylated end with the phosphodiester bond formation and the release of the adenyl group which covalently join the DNA Strands<sup>53</sup>.

## 5.2 Ligase enzymes, a branch of the nucleotidyltransferase superfamily:

The covalent nucleotidyltrans-ferase superfamily is comprised of members that catalyze the chemical addition reactions of nucleotidyl monophosphate at 5' ends through the lysine adenylate intermediate complexes. This superfamily branches include DNA ligases, RNA ligases and mRNA capping<sup>54</sup>. DNA ligases

are grouped into two families according to the nucleotide substrate requirements for ligase adenylate formation, ATP-dependent DNA ligases and NAD-dependent DNA ligases<sup>55,47</sup>. The NAD-dependent DNA ligases were surmised to be exclusively encoded in all known bacteria (are monomeric proteins of 70-80 KDa), while ATP-dependent DNA ligase are found in eukarya and archaea<sup>56</sup>. These enzymes display a broad range of molecular mass, 30-100 KDa, due to the diversity of the N-terminal region of each DNA ligase<sup>57</sup>, also those ligases from bacteriophages and viruses are classified in the ATP-dependent DNA ligase family<sup>58</sup>.

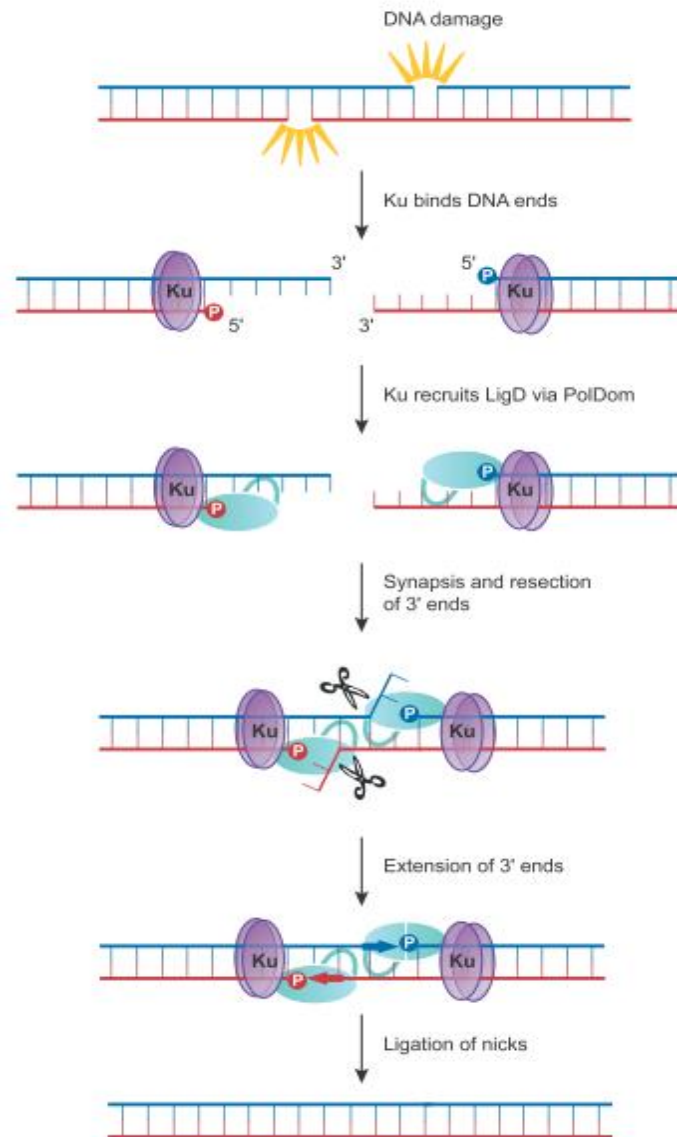
However, with the identification and characterization of new ligases, the presumption that bacteria only encode NAD-dependent DNA ligase (LigA) is overruled by the demonstration of a bacteria in 1997, *Haemophilus influenzae*, to possess ATP-dependent DNA ligases; a possibility of the result of viral genes integrating in the bacterial genomes through horizontal gene transfer. Also recently the discovery of coexistence of ATP dependent DNA ligases along with NAD dependent DNA ligases in several other bacterial species including *Neisseria meningitidis*, *Y. pestis*, *Vibrio cholerae*, *Pseudomonas aeruginosa*, overturn this presumption<sup>47</sup>. Specifically the rich assortment of ATP-dependent ligases (named LigB, LigC and LigD) found in *Mycobacterium tuberculosis* and *Agrobacterium tumefaciens*<sup>48</sup>.

### 5.3 Discovery of NHEJ in bacteria:

The identification of ATP-dependent DNA ligases in bacterial species derived the sequence search and the bioinformatic studies that led to revealing the existence of bacterial Ku homologues genes that are encoded by a variety of prokaryotes<sup>42,43</sup>. Interestingly, in most but not all bacterial species, the prokaryotic Ku genes are found in operons containing a conserved ATP-dependent DNA ligase, LigD (ligase D). However the absence of these genes from certain bacterial species as *E. coli*, raise inquires about the evolution of the NHEJ machinery in bacteria and its relation to horizontal gene transfer that may explain the

distribution of these genes among phylogenetically different organisms. Another explanation would be that these genes are descendants' of an ancient common ancestor to eukaryotes and bacteria and were subsequently lost in some bacterial species during evolution<sup>43</sup>.

LigD is considered the key agent of the prokaryotic non-homologous end-joining (NHEJ) pathway of DNA double-strand break (DSB) repair<sup>43,59</sup> (figure 5). It is a multifunctional ATP-dependent DNA ligase that often contains a core ligase domain (LigDom), a DNA polymerase domain (PolDom) and a nuclease domain (NucDom), where the later poses 3'-5' exonuclease activity that aids in processing the break termini by resection<sup>43</sup>. The order of these three catalytic domains differ among bacterial species<sup>60,61</sup>. It is possible to invoke that the mutagenic signatures are likely due to the activity of the polymerase domain that has been displayed *in vitro* to be related to non templated single-nucleotide addition to the 5'phosphate overhang DSB as a role in the fill-in synthesis. Studies of polymerase-defective mutant gene of LigD of *Mycobacterium tuberculosis* produced in *E. coli*, with abolished nontemplated nucleotide additions, demonstrated the polymerase LigD direct role in the mutagenic nucleotide addition during plasmid joining<sup>61,62</sup>. In addition, in these studies the isolation of the C- terminal ligase domain of LigD alone retained the nick sealing properties of the full length Lig D protein which entails the accessory role of the N-terminal domains compared to the main role of the ligase domain (LigDom) in the nick sealing through the NHEJ pathway<sup>47</sup>.



**Figure (5): The mechanism of NHEJ in prokaryotes**

After the DNA damage, Ku binds to the break termini and recruits the LigD where the PolDom will specifically recognize the 5' phosphate end (P). Microhomology pairing of complementary ends is followed by resection of the 3' nonextendable ends through the NucDom activity. Resynthesis by the polymerase activity and ligation of the nick through LigDom will complete the process of end-joining repair<sup>43</sup>.



#### **5.4 Sequence Conservation among DNA ligases and the three dimensional structure:**

The DNA ligases primary sequences were identified through structure functional analyses to illustrate five conserved peptide motifs (I, III, IIIa, IV, and V), those motifs give a very well alignment with no insertions or gaps between viruses, eukarya and archaea <sup>57</sup>. These motifs play crucial roles in nucleotide binding, nick recognition, and the transfer of nucleotidyl <sup>63</sup>. These motifs are found and shared between the ATP-dependent DNA ligases and the NAD-dependent DNA ligases, within the nucleotidyltransferase domain (NT) along with an oligonucleotide binding domain (OB) fold that contain within the motif VI <sup>54</sup>. However this motif is poorly conserved in the bacterial ATP-dependent DNA ligases <sup>64</sup> (figure 6). Despite sharing the presence of the common catalytic domains, the little sequence similarity found between the ATP-dependent DNA ligases and the NAD-dependent DNA ligases, in comparison to the high level of sequence homology conserved within each class of DNA ligases, implicates that these two families are not closely related in structure <sup>65</sup>.

NAD <sup>+</sup> -Dependent DNA ligases													
1. Ec LigA:	112-	<u>ELKLDGLA</u>	-47-	LEVRGEVF	-45-	TFFCYGVGVLEG	-52-	DGVVI	-20-	AVAFKFPAGEQ	-64-	PQVVNVLSER	-276
2. Tfi Lig:	113-	<u>EHKVDGLS</u>	-42-	LEVRGEVY	-47-	TFYALGLGLGLE	-52-	DGVVL	-20-	ALAYKFPAGEK	-64-	PEVLRVLKERR	-274
3. BSt Lig:	111-	<u>ELKIDGLA</u>	-44-	LEARGEAF	-45-	DLFVYGLADAEA	-51-	DGIVI	-20-	AIAYKFPAGEV	-64-	PEVGVVVDRR	-280
4. Mt LigA:	120-	<u>ELKIDGVA</u>	-50-	LEVRGEVF	-47-	ICHGLGHVEGFR	-49-	DGVVV	-20-	AIAYKYPPAEA	-64-	PEVLGPVVELR	-286
5. Bbu Lig:	109-	<u>EPKIDGCS</u>	-44-	LVLERGEVY	-43-	FIYDFLNAGLEF	-51-	DGVVL	-20-	AMAYKFEALSG	-64-	PAVEMVINKFS	-274
6. Ec yicF:	123-	<u>QPKVDGVA</u>	-45-	STLQGEIF	-38-	FVNANPDGPQLM	-44-	DGVVV	-18-	LVAWKYQFVAQ	-64-	FRIDDDVWR--	-177
7. Tpa Lig:	97-	<u>QHKLDGVS</u>	-52-	GGVRGEVI	-37-	VCYDAVPSTPGK	-55-	DGLVV	-17-	QIAFKFSTQEA	-63-	PKIEALVSTPA	-447
NAD <sup>+</sup> Consensus:		<u>E.KhDghs</u>		hEhRGEhh		.h.sh.....		DghVh		AhA.Kh.s...		Pph...s....R	
ATP-Dependent DNA ligases													
8. Hi URF1:	38-	<u>SEKLDGVR</u>	-28-	FAIDGELF	-24-	KLYVFDVPADEG	-47-	EGVVV	-14-	ILKLTARGE	-73		
9. Vc 1542:	47-	<u>SEKLDGIR</u>	-28-	YSLEGELW	-26-	SLMLFDMPPAAG	-47-	EGVVL	-14-	LLKLRHQDAE	-76		
10. Cj1669c:	39-	<u>SEKLDGVR</u>	-28-	FAIDGELW	-26-	TYNIFDVNPAGE	-54-	EGIVI	-14-	ATKLPYDDAE	-77		
11. NmB2048:	43-	<u>SEKLDGVR</u>	-28-	YPLDGEY	-23-	RLHVFDPKAAQ	-47-	EGVVL	-14-	LLKLSQYDDE	-75		
12. Aq 1394:	248-	<u>EYKYDGER</u>	-38-	FIVELEAV	-33-	AGFLFDIILYDG	-51-	EGLVC	-17-	WIKYKRDYKSV	-118-	PRFTGRYRFDK	-25
13. Mt LigB:	208-	<u>EAKLDGAR</u>	-38-	LVADGEAI	-33-	SVFFFDILHRDG	-48-	EGVMA	-15-	WLKVKPVHTLD	-96-	ARVV-RYRADK	-15
14. Bs ykoU:	21-	<u>EVKYDGYR</u>	-43-	LTLDGEIV	-34-	CFLAFDILLERSG	-57-	EGIVA	-15-	WLKYKNFKQAY	-82-	IGFEFQMDWTE	-304
15. Bh 2209:	20-	<u>EVKYDGYR</u>	-43-	ITLDGELV	-34-	TLLAFDILELKG	-57-	EGVVA	-15-	WLKKKNFRQVT	-81-	HRFRLDVKPAQ	-306
16. Mt Lig C:	26-	<u>EPKWDFGR</u>	-38-	CVIDGEII	-32-	SFIAPFDLLALGD	-54-	DGVIA	-13-	MPKIKHLRTAD	-114-	TAQFNRRWRPDR	-26
17. Bs yocV:	22-	<u>ELKFDGIR</u>	-35-	TVLDGEVI	-26-	VYCVFDVIYKDG	-47-	EGIVI	-15-	WLKVINYDYTE	-81		
18. Pa 2138:	235-	<u>ELKLDGYR</u>	-38-	SWLDGELV	-35-	LYVLFDPYHEG	-49-	EGVIG	-14-	WIKLKCQLRQE	-111-	AREVTGERPAG	-313
19. Mt Rv0938:	478-	<u>EGKWDFGYR</u>	-38-	VVLDGEAV	-22-	EFWAFDILLYDG	-46-	EGVIA	-15-	WVKDKHWNTQE	-98-	-SSWRGLRPDK	-8
Bact ATP Consensus:		<u>s.KhDghR</u>		..hpGEhh		.h.hFDh....s		EGhhh		hhK.K.....		.....	
20. T7 Lig:	31-	<u>EIKYDGYR</u>	-48-	FMLDGEIM	-49-	HIKLYAILPL--	-62-	EGLIV	-14-	WWKMKPENEAD	-96-	PSFVM-FRGTE	-7
Motif		I		III		IIIa		IV		V		VI	

**Figure (6): Alignment of conserved sequence elements among Bacterial DNA ligases**

Six conserved sequence elements have been identified in DNA ligases and mRNA capping enzymes (motifs I, III, IIIa, IV, V and VI) although the alignment for motif VI is poor in the Bacterial ATP-dependent sequences. The sequence of T7 DNA ligase is provided for comparison of the Bacterial DNA ligases. Homology among all NAD-dependent DNA ligases is high. The active-site motif of DNA ligases that contain the adenylated lysine (KXDG) is underlined. Coloured capital letters denote residues that are identical in all Bacterial DNA ligases. Ec, *Escherichia coli*; Mt, *Mycobacterium tuberculosis*; Tfi, *Thermus filiformis*; Bst, *Bacillus stearothermophilus*; Bbu, *Borrelia burgdorferi*; Tpa, *Treponema pallidum*; Hi, *Haemophilus influenzae*; Vc, *Vibrio cholerae*; Cj, *Campylobacter jejuni*; NmB, *Neisseria meningitidis* serogroup B; Aq, *Aquifex aeolicus*; Bs, *Bacillus subtilis*; Bh, *Bacillus halodurans*; Pa, *Pseudomonas aeruginosa*<sup>64</sup>.

The adenylation nucleotidyl transferase domain (NT) that contains the active-site lysine (known as the KXDG (Lys-Xaa-Asp-Gly) motif) and the oligonucleotide-binding OB-fold domain (OBD), are jointly called the catalytic core<sup>57</sup>. The OB domain binds non-specifically to DNA and positions the nucleotidyltransferase domain (NT) on DNA substrate to form a ring-shaped protein structure enclosing the substrate as a critical step for the enzyme nucleotidylation and coordination of the PP leaving group<sup>54</sup>. The presence of

both of these domains in ATP-dependent DNA ligases and NAD-dependent DNA ligases, is probably the reason why both classes utilize same catalytic mechanism despite the fact that they have different cofactor requirement <sup>45</sup>.

When in complex with DNA, the OBD is required to rotate between two active conformations during the course of DNA end ligation, where residues on one face of the OBD engage the substrate, while residues on the opposite face of the OBD aid in the adenylation step. When catalysis reaction is terminated, the closed conformation formed by the encirclement of the N-terminal domain to the DNA substrate must open to permit the release of products and enable multiple turnovers <sup>66</sup> as explained in the elucidation of the three dimensional structure of the *Pyrococcus furiosus* DNA (PfuLig) <sup>57</sup> and *Chlorella virus* DNA ligases <sup>67</sup>.

Examination of the positions of the conserved sequence motifs within the characterized crystal structures of the T7 bacteriophage DNA ligase, that represent the simplest DNA ligases encoded by viruses and bacteriophages, revealed to have the two-domains (The adjacent nucleotide-binding domain (ATP binding in this case)) and the (OB-fold domain)<sup>53</sup>. These domains are rather analogous in structure to the catalytic core of more complex multidomain DNA ligases found in bacteria, where the motif V forms the bridge between the two domains and consists of two  $\beta$  strands one on each side of the two domains<sup>52</sup>. The adenosine nucleotide binds in a pocket that is located entirely within the N-terminal domain and is composed of two anti parallel  $\beta$  sheets flanked by  $\alpha$  helices, this pocket is lined with the conserved motifs. These motifs participate in hydrogen bonding interactions with the ribose sugar hydroxyls of the adenylate cofactor. The active site lysine (Lys) residue lies in motif I and forms the enzyme-(lysine) AMP intermediate. The glutamate (Glu) residue in the motif III, binds ribose sugar of the ATP with hydrogen bonds while the tyrosine (Tyr) residue stack against the cofactor ring in motif IIIa. And the essential Lysine binds the  $\alpha$  phosphate group within the motif V <sup>56</sup>.

## 6. Biotechnological applications of DNA Ligase:

The DNA based diagnostics have been engaging the developed molecular biology tools for detection of bacterial and viral infectious agents and diseases, through the identification of insertions, deletion or substitutions (single nucleotide changes), in genes of medical concern that is denoted as single nucleotide polymorphism (SNPs) <sup>68,69,70</sup>. Therefore a worthy of reliance DNA diagnostics method will demand amplification of target sequences with precise single-base discrimination, low background, a plus advantage of complete automation and high sensitivity technology <sup>68</sup>.

There is also the obstacle of detecting mutations in genes controlling the cell cycle and apoptosis that leads to cancer. They require the development of technology that is able to identify exactly one or more low abundance mutations present in only one of the two chromosomes of a tumor cell. These tumor cells depict only a 15 % of the DNA sequence present in a sample for that gene that is contaminated with at least 70% of normal cells from the total cells in the early stages of tumors <sup>71</sup>.

### 6.1 Ligase chain reaction (LCR):

Ligases became crucial reagents in molecular cloning development of the DNA biotechnology field, including molecular diagnostics since their discovery in molecular biology laboratories at 1967 <sup>52</sup>. This has led to broad diversity of diagnostic applications evolved to genotype SNPs, some of these include LCR (ligase chain reaction) alone or coupled to PCR and other techniques. Thermostable DNA ligases obtained from thermophilic and hyperthermophilic organisms are the main component of the ligase chain reaction (LCR) for amplification and detection of single-base genetic diseases, due to their high thermostability feature that make them perfectly fit for the typing of single nucleotide polymorphism (SNP). Therefore they have been utilized as model systems for structural and mechanistic studies of DNA ligation <sup>72</sup>.

As for the fundamental of the LCR technique, it can be explained with the following example: Two sets of primers are synthesized such that they completely hybridize to the denatured target double DNA strands in such a way that the joining of each set on one strand lies at the nucleotide that represents a possibility of having the single base pair polymorphism. Only when these two nucleotides are complementary the two primers can be joined by the activity of the ligase, after a few cycles where the ligated products can serve as templates for the next following cycles this product will increase while if a mismatch is present in the joining position of the primers, no product will be distinguished. LCR mainly employs ligase enzyme that is characterized by thermostability to minimize target-independent ligation in oligonucleotide probe based assays<sup>73</sup>.

Along with the utilize of ligase chain reaction in finding cancer cells in the presence of normal cells example K-ras Oncogene Mutations and detecting p53 Mutation<sup>74,75,76</sup>, modification of the main principle was applied for identification of pathogenic organisms from normal flora in various infections as detection of *Neisseria gonorrhoeae*<sup>77</sup>, *Chlamydia trachomatis*<sup>78</sup>, *Mycobacterium tuberculosis*<sup>79</sup> and *Listeria monocytogenes*<sup>80</sup> in addition to identification of a number of infectious viruses example West Nile Virus<sup>81</sup>, HBsAg Mutants<sup>82</sup> and HCV RNA<sup>83</sup>.

This traditional LCR method is recently being developed to cope with the advances in the diagnostic methods:

### 6.2 ELISA:

This method involves constructing ligase chain reaction (LCR) primers that allow the differentiation of species down to a single base pair difference. Coupled with PCR to increase sensitivity of detection and by labeling LCR primers with biotin and digoxigenin, the LCR was combined with an ELISA-based detection system to improve sensitivity and ease of application<sup>84</sup>.

### **6.3 Multiplex LCR:**

This novel SNP genotyping method: gap ligase chain reaction (Gap-LCR) PCR amplifies allele-specific fragments followed by the use of a Luminex 100 fluorescent microsphere-based liquid assay. This method achieves simultaneous parallel detection using several tag-attached microspheres, in the Luminex detection platform <sup>85</sup>.

### **6.4 Quantitative real time LCR:**

Sensitive quantitative method of real time LCR is dependent on fluorescent dye that is DNA specific which is used to quantify the polymorphism from mutations of encephalopathy and can be accurately used in detection of drug resistance that results from point mutations & genetically modified organisms in food along with the previously mentioned applications <sup>86</sup>.

### **6.5 Colorimetric nanoparticle based method:**

where a method is developed that take advantage of the properties of gold nanoparticles emerging as a novel applied branch of science along with the ligation reaction in a colorimetric detection method of single-base discrimination can be performed at a relatively high temperature to discriminate between differently colored purple solution of perfectly matched assembled nanoparticles and the red colored solution of mismatched separated nanoparticles <sup>87</sup>.

## Chapter 2: Materials & Methods

---

### 1. Sample collection:

Water samples were collected from the convective layers of Atlantis II brine pool during the KAUST Red Sea Spring 2010 expedition. Sample collection was performed by utilizing Niskin bottles. These bottles are associated with Conductivity, Temperature and Depth (CTD) meter tool, for assessment of the salinity, temperature, pH, dissolved oxygen and depth. Almost 100 liters of water samples from each layer of the brines were collected and subjected to serial filtration using different sized pores 3  $\mu\text{m}$ , 0.8  $\mu\text{m}$  and 0.1  $\mu\text{m}$  mixed cellulose ester (Millipore) filters. The filters were preserved in sucrose buffer and kept in a -20°C freezer on the research vessel until transport back to the laboratory, where they were stored at -80°C till the time of DNA extraction. These previous steps were done by KAUST Red Sea Spring 2010 expedition members.

### 2. DNA Extraction, Sequencing & Construction of the 454 metagenomic database:

Prokaryotic DNA was extracted from 0.1  $\mu\text{m}$  filters using the Metagenomic DNA Isolation Kit for Water (Epicentre® Biotechnologies; Cat.No. MGD08420). 1X whole genome amplification (WGA) for the DNA sample was applied using the REPLI-g Mini Kit (Qiagen, USA) to obtain DNA concentration adequate for direct sequencing. DNA concentration was determined using the Quant-it™ PicoGreen® dsDNA Kit (Invitrogen, USA) and the Thermo Scientific NanoDrop™ 3300 Fluorospectrometer.

A single stranded (ss) DNA library was constructed by nebulization and attachment on emulsion beads, then was amplified by emulsion PCR (emPCR). Afterward, Pyrosequencing was done by Roche 454 GS-FLX genome analyzer using GS-FLX Titanium pyrosequencing kit to generate the metagenomic

database. For the assembled 454 metagenomic database, generated reads were collectively assembled into contigs by 454 life science corporation Newbler® GS assembler version 2.6. Then using MetaGeneAnnotator (MGA)<sup>88</sup>, potential open reading frames (ORFs) were predicted for the assembled contigs. These previous steps were performed at the Biology Department, American University in Cairo.

### **3. Computational analysis:**

#### **3.1. Screening of the ATII-LCL metagenomic database for DNA ligase:**

A search with the keyword “DNA ligase” was done in the pfam protein family signature database 27.0 (<http://pfam.sanger.ac.uk/>)<sup>89</sup>. Pfam accessions corresponding to domains of the N and C terminal of DNA ligases were selected to curate and build a concatenated HMM model from the raw HMM of each domain. The model was used in an HMM search against the 454 assembled ORFs metagenomic database for improving the quality of annotation. Simultaneously, the full length sequences of the alignment for each domain were concatenated and used for a tblastn against the 454 assembled ORFs metagenomic database. The resulted reads of both searches were extracted and filtered for unique reads then subjected to assembly.

#### **3.2. Functional Annotation of the resulting contigs:**

Following the assembly of reads, ORFs within contigs were detected by Artemis annotation & visualization tool<sup>90</sup> and then annotated by searching against the National Centre for Biotechnology Information (NCBI) non-redundant protein database with BLASTx using the BLOSUM45 substitution matrix<sup>91</sup>. The result revealed a “contig 5” that contained within an ORF encoding a putative DNA ligase which was chosen for further investigations and was assigned the name “LigATII”.



### **3.3. Prediction of the upstream regulatory regions of LigATII:**

The prediction of the regulatory promoter regions (-35 and -10 regions) was done by using BPPROM (<http://linux1.softberry.com>) based upon the highest Linear discrimination function (LDF) score. The ribosomal binding site “Shine delgarno” region was predicted manually 7 bps upstream of the ORF’s start codon as the site rich in G and A nucleotides. The amino acid sequence corresponding to the identified ORF was predicted using the translation tool from ExPASy resources online ([http://web.expasy.org/compute\\_pi/](http://web.expasy.org/compute_pi/))<sup>92</sup>.

### **3.4. Conserved Domain search of the LigATII:**

The deduced amino acid sequence of LigATII was then submitted to the Conserved Domain Database (CDD) provided by the NCBI<sup>93</sup> that is to identify the conserved domains within the LigATII that displays similarities to identified domains in the DNA ligase families.

### **3.5. Multiple sequence alignment of LigATII:**

Multiple sequence alignment of LigATII with sequences of ATP dependent DNA ligases was carried out using Clustal Omega software<sup>94</sup> to identify the presence of the conserved motifs and the amino acids residues that are putatively involved in catalysis.

### **3.6. Identification of functional regions in the LigATII:**

The amino acid sequence of LigATII was submitted to the ConSurf web server (<http://consurf.tau.ac.il/>)<sup>95</sup> used for identification of functionally important regions in proteins by estimating the degree of conservation of the amino-acid sites among their close sequence homologues.

**3.7. Phylogenetic analysis for the LigATII:**

For the phylogenetic analysis, sequences of 66 DNA ligases from bacteria, archaea, viruses & eukaryotes were acquired from the NCBI protein sequence database. Phylogenetic tree was constructed using phylogeny.fr web service <sup>96</sup> ([http://www.phylogeny.fr/version2\\_cgi/index.cgi](http://www.phylogeny.fr/version2_cgi/index.cgi)).

Multiple sequence alignment of the retrieved sequences and LigATII was performed using full mode run MUSCLE. Alignment was curated manually using Jalview version 2.8. Phylogenetic relations were inferred by Maximum likelihood (ML) estimation approach using PhyML version 3.0 programs. WAG amino acid substitution model was selected with gamma distribution parameter estimated and the number of substitution rate categories was four. The constructed tree confidence was estimated by a bootstrap of 100 pseudo-replicates. The tree was visualized using the Interactive Tree Of Life (iTOL) v2 online server (<http://itol.embl.de/index.shtml>).

**3.8. Comparative homology modeling of LigATII:**

The prediction of three dimensional (3D) models for the LigATII protein was done using MODELLER version 9.13. The template used to generate the model was the crystal structure of Bacteriophage T7 complexed with ATP adenosine Tri-phosphate (PDB ID: 1AO1|A) determined as the simplest DNA ligase that contain the catalytic core domains. This template was retrieved from BLASTP search of the LigATII against PDB database. Comparative homology modeling predicted the most satisfying model of LigATII with the template by structural alignment in regards of spatial restraints and molecular geometry.

Structural superimposing and prediction of substrate's binding site in LigATII were carried out by Discovery Studio® visualizer 4 (Accelrys Software Inc.).

### 3.9 Prediction of the molecular weight and isoelectric point of LigATII:

The theoretical isoelectric point and the molecular weight of LigATII were calculated using the Compute pI / Mw analysis tool on the ExPASy server<sup>92</sup>.

## 4. Isolation of the putative LigATII:

### 4.1 PCR based screening method:

The 1X whole genome amplification of the LCL environmental DNA was screened using two pairs of primers. The first pair contg5\_F (5'-CGTGATTACTGGGCTTTTGA-3') & contg5\_R (5'-GCAGTTCTTTATTGAAACCCG-3'), designed based on one read within the "LigATII" ORF's middle region, starting from position 431bp to 835bp. After successful amplification, the full ORF of LigATII was amplified by using a second pair of primers, ORF5\_F (5'-GTCAAGACTAT CGAACCTAT- 3') & ORF5\_R (5'-TTATCCACTTCTTATTGCCCCC-3'). The PCR reactions were carried out using the Applied Biosystems® Veriti® 96-Well Fast Thermal Cycler (Life Technologies, USA). The PCR conditions for all primers were: an initial denaturation step at 94 °C for 5 minutes, then 30 cycles of denaturation at 95 °C for 30 seconds, annealing at 55 °C for 30 seconds and extension at 72 °C for 30 seconds then a final extension step at 72 °C for 7 minutes. The PCR products were then analyzed using 1% agarose gel electrophoresis.

### 4.2 Cloning and sequencing of the PCR product:

The PCR product amplified with ORF5\_F and ORF5\_R primers was purified using QIAquick® PCR Purification kit (Qiagen, USA). The purified amplicon was then ligated to pGEM®-T Easy cloning vector (Promega, USA) according to the manual instructions. Ligation products were transformed into electrocompetent *E. coli* Top10 cells using MicroPulser™ Electroporation Apparatus (Bio-Rad, USA). Transformed cells were plated on LB agar containing 0.5 mM isopropyl  $\beta$ -D-1-thiogalactopyranoside (IPTG), 40  $\mu$ g/ml 5-bromo-4-chloro-indolyl- $\beta$ -D-galactopyranosid (X-gal) and 200  $\mu$ g/ml ampicillin

to allow blue/white selection. A number of 5 to 10 white colonies were carefully picked and colony PCR was done to confirm correct insert orientation using the T7 promoter forward primer and the ORF5\_R. Recombinant plasmids were extracted from colonies with positive PCR results using PureYield™ Plasmid Miniprep System (Promega, USA).

Several sequencing reactions were prepared using BigDye® Terminator v3.1 Cycle Sequencing Kit. Primers used for sequencing were ORF5\_F (Ta 51°C), ORF5\_R (Ta 59°C), also Sp6 reverse primer (5'-TAATACGACTCACTATAGGG-3' of Ta 59°C) and T7 forward primer (5'-GA TTT AGGTGACACTATAG-3' of Ta 51°C) promoter primers of pGEM®-T Easy cloning vector. Each Sequencing reaction comprised big dye terminator, 5X Sequencing Buffer, primer (forward or reverse), purified cloned p-GEM®-T Easy vectors sample and distilled sterilized water completed to the desired volume. Sequencing was carried out in the Applied Biosystems 3730x DNA Analyzer. Thermo cycler conditions were initial denaturation 96°C for 5 minutes, 30 cycles (denaturation at 96°C for 30 seconds, annealing at 55°C for 30 seconds, extension at 72°C for 30 seconds) and a final extension step for 7 minutes. The resulting sequences were analyzed using BioEdit v7.2.3 sequence analysis program.

## **5. Construction of LigATII expression systems:**

### **5.1 Expression of LigATII gene using pET SUMO® Expression System:**

In order to clone LigATII into Champion™ pET SUMO expression plasmid (Invitrogen™) [N-terminal Histidine-tagged and SUMO fusion protein], the LigATII ORF was amplified from purified recombinant p-GEM-T® verified to contain the insert with correct sequence. The amplification was performed using ORF5\_F and ORF5\_R. The purified amplicon was ligated into Champion™ pET SUMO plasmid using the pET SUMO TA Cloning® reagents. The plasmid was transformed into electrocompetent *E. coli* BL21 DE3 by the MicroPulser™ Electroporation Device (Bio-Rad) according to manufacturer's instructions

(Transformation in *E.coli* Top10 was done with same procedure and kept as glycerol stocks). Transformants were plated on LB-agar plates containing 50µg/ml kanamycin. The transformants having the insert with the correct orientation were identified by colony PCR using SUMO Forward sequencing primer (5'-AGATTCTTGTACGACGGTATTAG-3') and the ORF5\_R primer. The sequence fidelity of the cloned pET SUMO vector was confirmed by sequencing using the pET SUMO forward and T7 reverse primers.

## **5.2 Expression of LigATII gene using Champion™ pET100 Directional TOPO® Expression System:**

In order to clone LigATII into Champion™ pET100 Directional TOPO® expression plasmid [N-terminal Histidine-tagged], the LigATII ORF was amplified from purified recombinant pET SUMO plasmid verified to contain the insert with correct sequence. The amplification was performed using pet100\_F (5'-CACCATGGTCAAGACTATCGAACCTAT-3') and ORF5\_R. The forward primer included CACC nucleotide sequence upstream the ATG codon in order to facilitate the directional cloning of LigATII gene into the pET100 Directional TOPO® expression vector. The PCR reaction was carried out using Phusion High Fidelity DNA polymerase (Thermo scientific). Additionally a control reaction that involves producing a control PCR product of size (~ 750 bp) using the reagents included in the kit was done and this product was used directly in a TOPO® cloning reaction to give the control insert recombinant vector. Also the purified LigATII amplicon was ligated into Champion™ pET100 Directional TOPO® expression system using the kit reagents.

The plasmid was transformed into electrocompetent *E. coli* BL21 DE3 by the MicroPulser™ Electroporation Device (Bio-Rad) according to manufacturer's instructions (Transformation in *E.coli* Top10 was done with same procedure and kept as glycerol stocks). Also transformation into the electrocompetent *E. coli* BL21Star (DE3) by the MicroPulser™ Electroporation Device (Bio-Rad)

according to manufacturer's instructions was done for better induction of the expression of the LigATII.

Transformants were plated on LB-agar plates containing 200µg/ml ampicillin. In order to verify the sequence of the LigATII insert, extracted recombinant plasmids were subjected to chain termination sequencing using the Applied Biosystems 3730xl DNA Analyzer and the BigDye® Terminator v3.1 Cycle Sequencing Kit (Life Technologies, USA). Sequencing reactions were carried out for each recombinant plasmid using pet100\_F, ORF5\_R, also the T7 promoter (5'-TAATACGACTCACTATAGGG-3') and the T7 reverse (5'-TAGTTATTGCTCAGCGGTGG-3') primers of pET100 vector. The resulting sequences were analyzed using BioEdit v7.2.3 sequence analysis program.

### 5.3 Induction of expression of LigATII:

The induction of LigATII protein expression in both expression systems was attempted as follows: freshly prepared culture from an overnight growth of *E.coli* BL21 (DE3) harboring either of the expression systems was cultured in LB broth containing antibiotic as required at 37°C with shaking at 250 rpm, to an optical density at 600 nm (OD<sub>600</sub>) of (0.4 - 0.6). In case of pET SUMO expression system, a 1 ml aliquot of the culture was saved before the addition of IPTG as uninduced control. In case of recombinant pET100 expression vector, 1 hour interval samples were taken from the culture before reaching the required OD for analysis.

The induction conditions ranged from 0.5 mM IPTG at 37°C for 2 hours to 0.1 mM IPTG at 30°C or 25°C for a time interval of O/N & 16 hours, respectively in case of recombinant pET SUMO expression system. In the case of the recombinant pET100 Directional TOPO® expression system induction condition was 0.5 mM IPTG at 37°C for 2 hours, with a control uninduced sample separated before the IPTG addition and treated for the same incubation conditions as the induced samples. Also a trial with the addition of the 1% glucose to the initial culture in order to suppress the basal induction from the expression vector was

done. This glucose was removed from the culture prior to the addition of IPTG by centrifugation at 4,629 x g (6,000 rpm), decantation of the supernatant media and addition of fresh media with the appropriate amount of antibiotic to resuspend the pellet and proceed to induction steps.

The culture was harvested by centrifugation at 4,629 x g (6,000 rpm), 4°C. The pellet was thawed at 42°C and frozen at -80°C interchangeably for 3 successive times then resuspended in 1000µl of lysis buffer comprising 20mM NaH<sub>2</sub> PO<sub>4</sub>, 0.5M NaCl, 20mM imidazole (pH 7.4), 1mM phenylmethylsulfonyl fluoride (PMSF) and 1mg/ml Lysozyme. After 30 minutes of ice incubation, the cells were disrupted with sonication on ice by the SONIFIER<sup>®</sup> 150, Branson (15 seconds sonication with 15 seconds pause in between for 6 times). The crude cell lysate was separated from the cell debris by centrifugation for 25 minutes, 15,557 x g (11,000 rpm) at 4°C. The supernatant and pellet were analyzed by 12% sodium dodecyl sulfate - polyacrylamide gel electrophoresis (SDS-PAGE). SDS-PAGE gel preparation and buffers used was carried out according to Laemmli's standard protocol.

### 6. Functional Identification of the LigATII in the *E. coli* GR501:

Transformation of extracted recombinant pET100 Directional TOPO<sup>®</sup> expression vector and the control insert recombinant vector into electrocompetent *E. coli* GR501 temperature sensitive strain by the MicroPulser<sup>™</sup> Electroporation Device (Bio-Rad) was done according to manufacturer's instructions. This transformation was done to examine the functional activity of the LigATII protein expression *in vivo* by complementation of the deficient DNA ligase gene in the *E. coli* GR501 strain that leads to temperature sensitivity & lethality of the strain at the non permissive temperature. Transformants were plated on LB-agar plates containing 200µg/ml ampicillin. Colonies were collected and stock cultures containing 15 % glycerol (v/v) were stored at -80 °C and streaked onto fresh LB agar plates as required. At the start of each experiment, cultures containing *E. coli* GR501 were confirmed to be temperature sensitive.

### **6.1 Assay for LigATII complementation of temperature sensitive defect of *E. coli* GR501 on agar plates:**

To assay for complementation of the temperature sensitive defect, Glycerol stocks of transformations with the required vectors, maintained at -80°C, streaked onto fresh LB agar plates containing 200 µg/ml ampicillin and grown overnight at the permissive temperature 30°C. single colonies were then streaked onto two fresh LB / Amp agar plates: the plate streaked first was incubated at 43 °C and the one streaked second was incubated at 30 °C. Note that complementation of the LigATII protein expressed from pET100 Directional TOPO® could be achieved without addition of IPTG, indicating that expression from the strong *lac*-derived promoter of this plasmid was not completely inhibited in *E. coli* GR501.

### **6.2 Assay for strain viability of the *E. coli* GR501 by the LigATII complementation of the temperature sensitive defect:**

For analysis of the viability of the *E. coli* GR501 strain harbouring the recombinant vectors, the appropriate cultures were grown in LB/Amp at 30 °C overnight. Viable cell counts were determined by plating 200 µl of a 10<sup>-6</sup> dilution of the overnight culture onto LB/Amp agar plates and counting the colonies after aerobic incubation at 30 °C or 43 °C for 24 hours.

### **6.3 Assay for LigATII complementation of temperature sensitive defect of *E. coli* GR501 in liquid cultures:**

For liquid-culture growth, single colonies from plates grown at 30 °C were inoculated into 5 ml liquid medium and grown overnight at 30 °C. These cultures were diluted 100-fold into fresh LB medium, containing 200µg/ml ampicillin and incubated at the required temperatures of 30 °C or 43 °C. Growth of bacteria was detected by monitoring OD<sub>600</sub> every 30 min for the first 2 hours and subsequently every 15 min for the remainder of the incubation period, which is 5 hours.



# Chapter 3: Results & Discussion

---

## 1. Search for DNA ligase in the ATII-LCL Metagenomic Database:

Through a search with the keyword “DNA ligase” in the pfam database, four accessions were selected: PF01068 {ATP dependent DNA ligase domain}, PF04679 {ATP dependent DNA ligase C terminal region}, PF04675 {DNA ligase N terminus} and PF14743 {DNA ligase OB-like domain} to curate and build a concatenated HMM model from the raw HMM of each domain. The model was used in an HMM search against the 454 assembled ORFs metagenomic database for improving the quality of annotation. Simultaneously, the full length sequences of the alignment for each domain were concatenated and used for a tblastn against the 454 assembled ORFs metagenomic database. The resulted reads of both searches were extracted and filtered for unique reads then subjected to assembly.

Following the assembly of reads, 91 contigs were obtained and ORFs were detected in all contigs by artemis tool and annotated using BLASTx. The result revealed a “contig 5” of size 1514 bp generated from 270 reads that contained within an ORF of 918 bp (305 amino acids) encoding a putative DNA ligase (assigned hereafter the name “LigATII”). The ORF displays similarities to the modular architecture of two distinct domains (the adenylation domain of LigD and the oligonucleotide binding (OB)-fold domain) that are conserved to ATP-dependent DNA ligases.

## 2. Identification of LigATII:

### 2.1 Sequence Analysis of LigATII using BLASTx server:

Sequence similarity search against NCBI non-redundant protein database with BLASTx retrieved the highest similarity hit with “DNA ligase” protein from *Erysipelotrichaceae bacterium 5\_2\_54FAA* (Accession: WP\_0089783540.1) of the phylum Firmicutes. The maximum identity was 39%, positives 54%, query

coverage of 95% and an E-value of  $2e^{-51}$  (figure 7) such results predict some novel aspects recognized in LigATII.

Most of the hits were related to the phylum Firmicutes and the phylum Proteobacteria, where the later have been reported previously to be the most abundant phylum in the ATII-LCL<sup>97</sup>.

DNA ligase [Erysipelotrichaceae bacterium 5_2_54FAA]					
Sequence ID: <a href="#">reflWP_008978354.1</a> Length: 311 Number of Matches: 1					
<a href="#">▶ See 1 more title(s)</a>					
Range 1: 4 to 292 <a href="#">GenPept</a> <a href="#">Graphics</a>			▼ Next Match ▲ Previous Match		
Score	Expect	Method	Identities	Positives	Gaps
182 bits(463)	2e-51	Compositional matrix adjust.	117/300(39%)	163/300(54%)	21/300(7%)
Query 7	PMLAETLDPSDIDKLDWASFISEIKLDGCRVAVVNDGKVEL-RGRDQNLTPKFPPELSF-	64			
	P+ L ++ D ++I EIK+DG R +AY+ VEL R L KFPPEL				
Sbjct 4	PLYDAMLIGTEQPPFDDDAYIYEIKMDGVRCLAYLYKDHVELINKRHLKLSKFPPELKSL	63			
Query 65	--SVRKPCVIDGEI---TSADMSFEGIQHRVHKTKPMDIRIASKRYPVVIYWAFDILNLNG	119			
	+KPCVIDGE+ F IQ R + P IR SK++P ++ AFDILNL+G				
Sbjct 64	YKQAKKPCVIDGELHVFQDGKSDFFAIQRRTLTSDPFRIRQHSKKFPVFTAFDILNLGD	123			
Query 120	QDLTKKPLIERKEMLWENLIGN--CGV-RYLAHGQDGVSLFEKVRLGLEGIMAKRKKSK	176			
	+DL +KPL+ERK+L +N+ + C + RY+ Q+G +LF K+ LEGI+AKRK+S				
Sbjct 124	EDLCQKPLMERKKLLEKNIKESPICNISRYIE--QEGKALFALTQQQLEGIVAKRKESL	181			
Query 177	YQAGKRSDDLKIKTFEEGYTLIVGVTEGEGDRENTFGSLILAKETENGLAYVGNVSGSF	236			
	YQ GKR+ +W+K K E + +VG +E+ SL+LA L Y G+V G				
Sbjct 182	YQPGKRTKEWIKCKHLLEADFAVVGYP----KEHAMLSLVLAAYDNQKLRVCGHVTMGV	237			
Query 237	NKELLETLTYVLGLREYPCPFATEPDVGREVRFWTEPVFYCEVKHLGYGSDGLLRFVFEK	296			
	+K+ L + PCPF+ PD G E W P VK + Y G +R PVFEK				
Sbjct 238	SKDYL----FAHIKDTIPCFPSVLPD-GNEDATWISPFCLGTVKFMEYTQSGGMRQPVFEK	292			

**Figure (7): the best hit of blastx aligned with LigATII**

LigATII query aligned with the best similar hit of DNA ligase protein from *Erysipelotrichaceae bacterium 5\_2\_54FAA* that displays the highest maximum identity of 39%, positives 54% and E value of  $2e^{-51}$ .

## 2.2 Prediction of the upstream regulatory regions and ribosomal binding site of the LigATII:

Detection of the promoter regulatory regions (-35 and -10) upstream sequence of the ORF encoding the LigATII using Bprom software, based on the highest Linear discriminant function (LDF) score, was obtained. The -35 region (ctaccg) and the -10 region (agctatgct) were predicted at positions -164 bp & -135 bp upstream of the start codon residue (M), respectively. RBS “Shine-Dalgarno” was found 7 nucleotides upstream of the translation initiation site as the purine-rich region (aggagg) (figure 8).

```

agacagcggtaaaccacaagcgggaagcacttgagacagcggtaaaccacgagggtaccgatctcacagggcg
gtagctatgctgtgggtagctctgaacagcgataacttcacctagcgatagcacaatggcaaaagatatgcgttc
aggcaagcaggctggaggatgccgtagagaggttcgagaaaaacaagcctcagagctttcaggagggggaaa
1 - ATGGTCAAGACTATCGAACCTATGTTAGCCGAAACATTAGACCCAAGCGATATTGATAAG - 60
1 - M V K T I E P M L A E T L D P S D I D K - 20

61 - CTGGATTGGGCATCCTTTATCAGCGAGATTAAACTGGACGGATGCCGAGCAGTAGCCTAT - 120
21 - L D W A S F I S E I K L D G C R A V A Y - 40

121 - GTAAACGATGGCAAGGTAGAGCTTCGGGGTAGGGATCAAAACCTTACCCCGAAGTTCCCC - 180
41 - V N D G K V E L R G R D Q N L T P K F P - 60

181 - GAACTGTCTTTCAGCGTCAGGAAACCTTGCCTTATTGACGGGGAGATAACATCAGCGGAT - 240
61 - E L S F S V R K P C V I D G E I T S A D - 80

241 - ATGAGCTTCGAGGGAATACAGCACAGGGTTCACAAGACAAAGCCGATGGACATAAGGATC - 300
81 - M S F E G I Q H R V H K T K P M D I R I - 100

301 - GCTTCAAAGAGATACCCCGTGATTTACTGGGCTTTTGACATCCTGAACCTGAACGGGCAG - 360
101 - A S K R Y P V I Y W A F D I L N L N G Q - 120

361 - GATTTAACCAAAAAGCCTCTTATCAGAGAAAGGAAATGCTCTGGGAAAACCTGATAGGG - 420
121 - D L T K K P L I E R K E M L W E N L I G - 140

421 - AACTGCGGAGTGAGGTATCTTGCTCAGGGGACGGAGTAAGTCTTTTTGAGAAGGTT - 480
141 - N C G V R Y L A H G Q D G V S L F E K V - 160

481 - AAAAGACTGGGGCTTGAGGGGATAATGGCGAAAAGGAAAAAGTCAAAATACCAAGCTGGA - 540
161 - K R L G L E G I M A K R K K S K Y Q A G - 180

541 - AAGAGATCGGATGACTGGCTGAAAATAAAGACATTCGAGGAAGGAACTTACCTGATAGTC - 600
181 - K R S D D W L K I K T F E E G T Y L I V - 200

601 - GGGGTTACTGAGGGAGAGGGAGACAGGGAGAACACCTTTGGAAGCCTCATACTTGCCAAA - 660
201 - G V T E G E G D R E N T F G S L I L A K - 220

661 - GAAACGGAGAACGGGCTTGCCTATGTGGGAAATGTAGGTCGGGTTTCAATAAAGAAGT - 720
221 - E T E N G L A Y V G N V G S G F N K E L - 240




721 - CTGGAGACATTGACCTATGTTCTGGGGCTTCGGGAGTATCCCTGTCCGTTTGCCACAGAG - 780
241 - L E T L T Y V L G L R E Y P C P F A T E - 260

781 - CCAGATGTGGGCAGAGAGGTAAGGTTCTGGACTGAGCCTGTGTTCTACTGCGAGGTAAAG - 840
261 - P D V G R E V R F W T E P V F Y C E V K - 280

841 - CACTTGGGGTACGGGAGTGATGGGTTACTGAGGTTTCCCGTGTTCAAAAGATTGAAGGGG - 900
281 - H L G Y G S D G L L R F P V F K R L K G - 300

901 - GCAATAAGAAGTGGATAA - 918
301 - A I R S G * X - 320

```

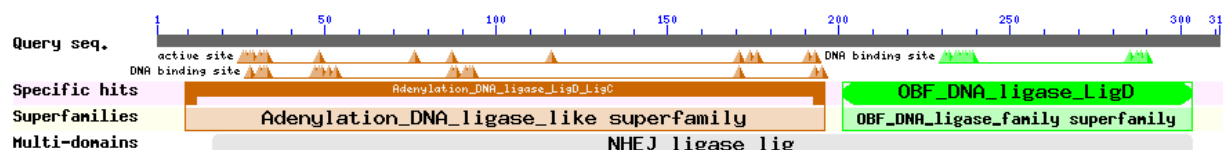
 -35     
 -10     
 RBS "Shine-Dalgarno"

**Figure (8): Nucleotide sequence and predicted amino acid sequence of the LigATII ORF**

The nucleotide sequence of the LigATII is shown with the upstream predicted regulatory region (-35 / -10 / 'Shine-Dalgarno') highlighted and the predicted amino acid sequence is given from the start codon residue (M) to the stop codon (\*).

## 2.3 Search for conserved domains within LigATII:

The search for conserved domains within the ORF of LigATII using the CDD tool detected an adenylatin or nucleotidyltransferase (NTase) N-terminal domain of LigD {cd07898} between amino acid residues 12 and 192. This domain contains the active site residue that binds ATP and is essential for catalysis. An oligonucleotide binding (OB)-fold C-terminal domain was also detected {cd07971} between residues 197 and 305 which contacts the nicked DNA substrates. Together these two domains comprise the catalytic core (ligase domain) of the LigD that is found in members of the ATP-dependent DNA ligase family DNA. This Family of DNA ligase is involved in repair of double-stranded breaks by non-homologous end joining (NHEJ) that is present in a minority of prokaryotes (figure 9).



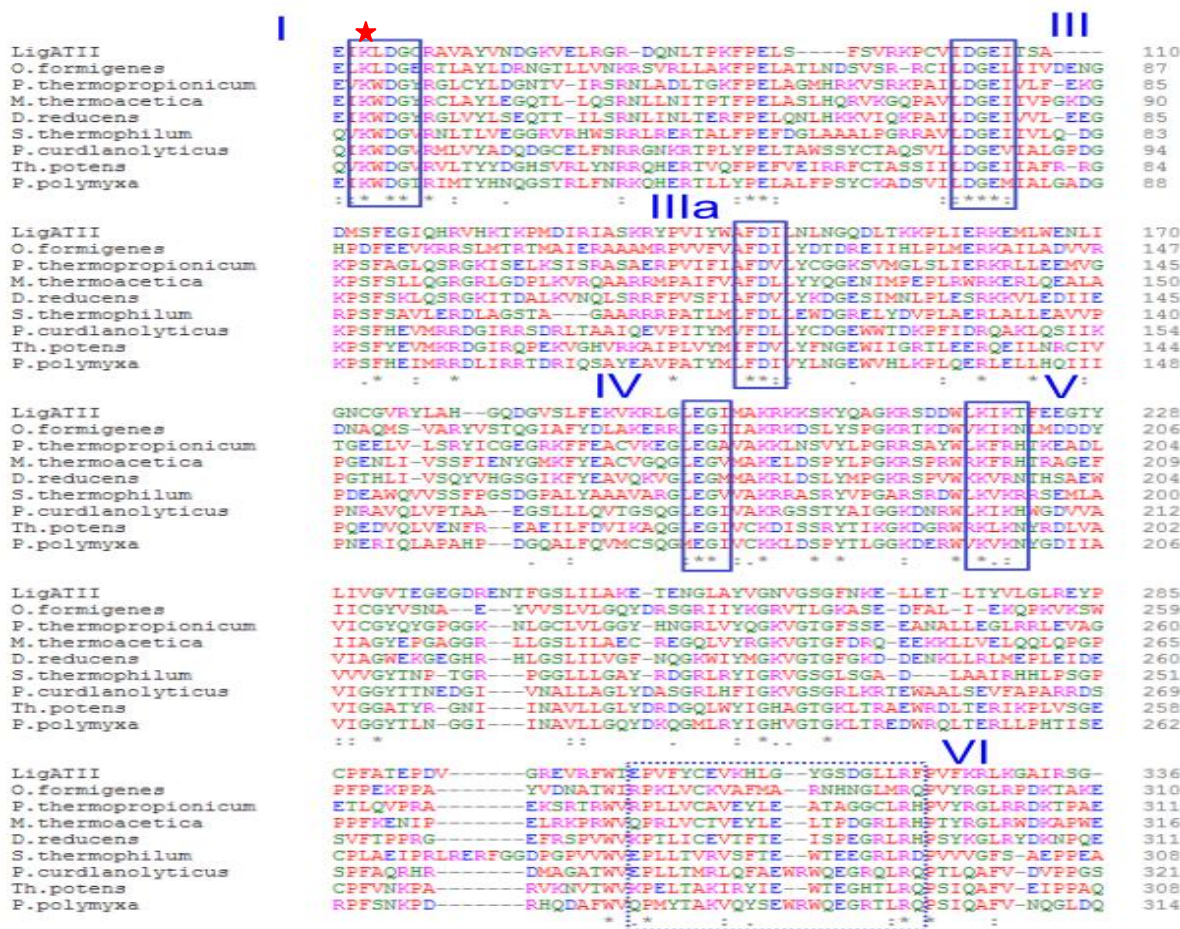
**Figure (9): The conserved domain identified within the LigATII**

The two characteristic conserved domains to the members of ATP dependent ligases family were detected in the LigATII by the CD search. The adenylation domain {cd07898} and the OB fold domain {cd07971} forms together the conserved catalytic core essential for activity.

## 2.4 Multiple sequence alignment of LigATII:

The presence of the catalytic core that consists of six sequence motifs (I, III, IIIa, IV, V and VI) conserved among bacterial ATP dependent DNA ligases was detected in LigATII from the multiple sequence using Clustal Omega. These motifs form the sides of the groove between the two domains of the catalytic core by clustering around the ATP. The ATP is attached to a lysine residue in the active site to form the ATP-enzyme intermediate which was identified in LigATII as Lys31, a part of the conserved motif I. A glutamate residue in motif III that forms the hydrogen bond with the ATP sugar ribose was identified in LigATII as

Glu75. On the other hand a lysine residue in motif V contacts the phosphate group and was identified in LigATII as Lys90<sup>56</sup> (figure 10).



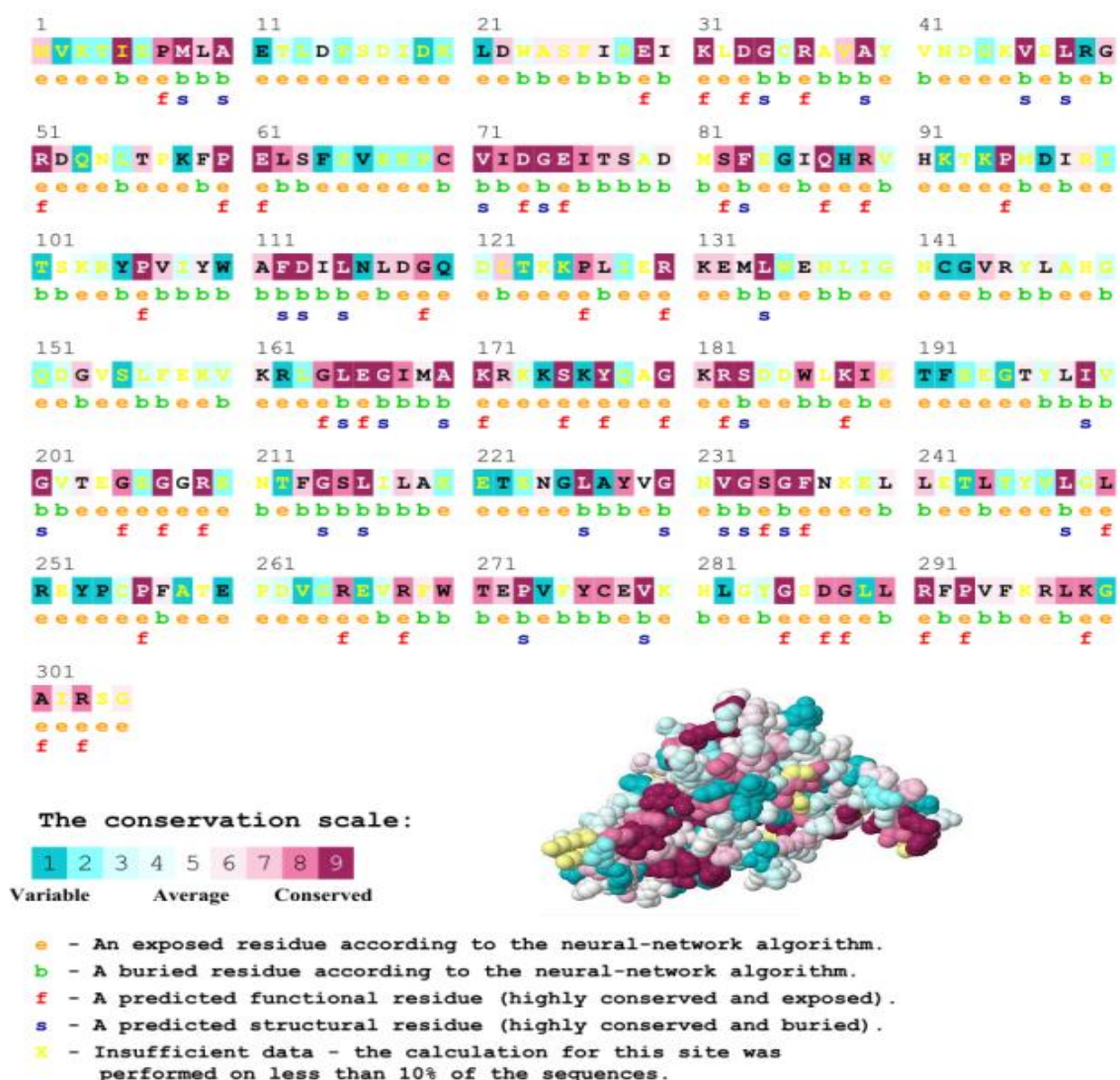
**Figure (10): Multiple sequence alignment of LigATII**

Multiple sequence alignment of LigATII with conserved sequence elements among Bacterial ATP dependent DNA ligases (Lig D) from *O. formigenes*, *Oxalobacter formigenes* OXCC13 (gi/229379681); *P. thermopropionicum*, *Pelotomaculum thermopropionicum* SI (gi/147677578); *M. thermoacetica*, *Moorella thermoacetica* ATCC 39073 (gi/83573245); *D. reducens*, *Desulfotomaculum reducens* MI-1 (gi/134052549); *S. thermophilum*, *Symbiobacterium thermophilum* IAM 14863 (gi/51892935); *P. curdianolyticus*, *Paenibacillus curdianolyticus* YK9 (gi/304346063); *Th. potens*, *Thermincola potens* JR (gi/296031639); *P. polymyxa*, *Paenibacillus polymyxa* E681 (gi/305856429). Six conserved motifs (motifs I, III, IIIa, IV, V and VI) were identified (blue boxes). The active-site motif of DNA ligases that contain the adenylated lysine (KXDG) is highlighted (the first blue box) that contain within the asterisked Lysine residue that binds ATP.



## 2.5 Identification of functional regions in LigATII using ConSurf:

The ConSurf server, based on multiple alignments, predicted that almost all of the Aspartic acid (D) and Glutamic acid (E) residues are functional residues that are highly conserved and exposed to the outside of the LigATII protein structure (figure 11). From this information we can imply that LigATII possesses certain degree of halophilicity referring to previously characterized proteins that are known to share the same properties<sup>98</sup>.

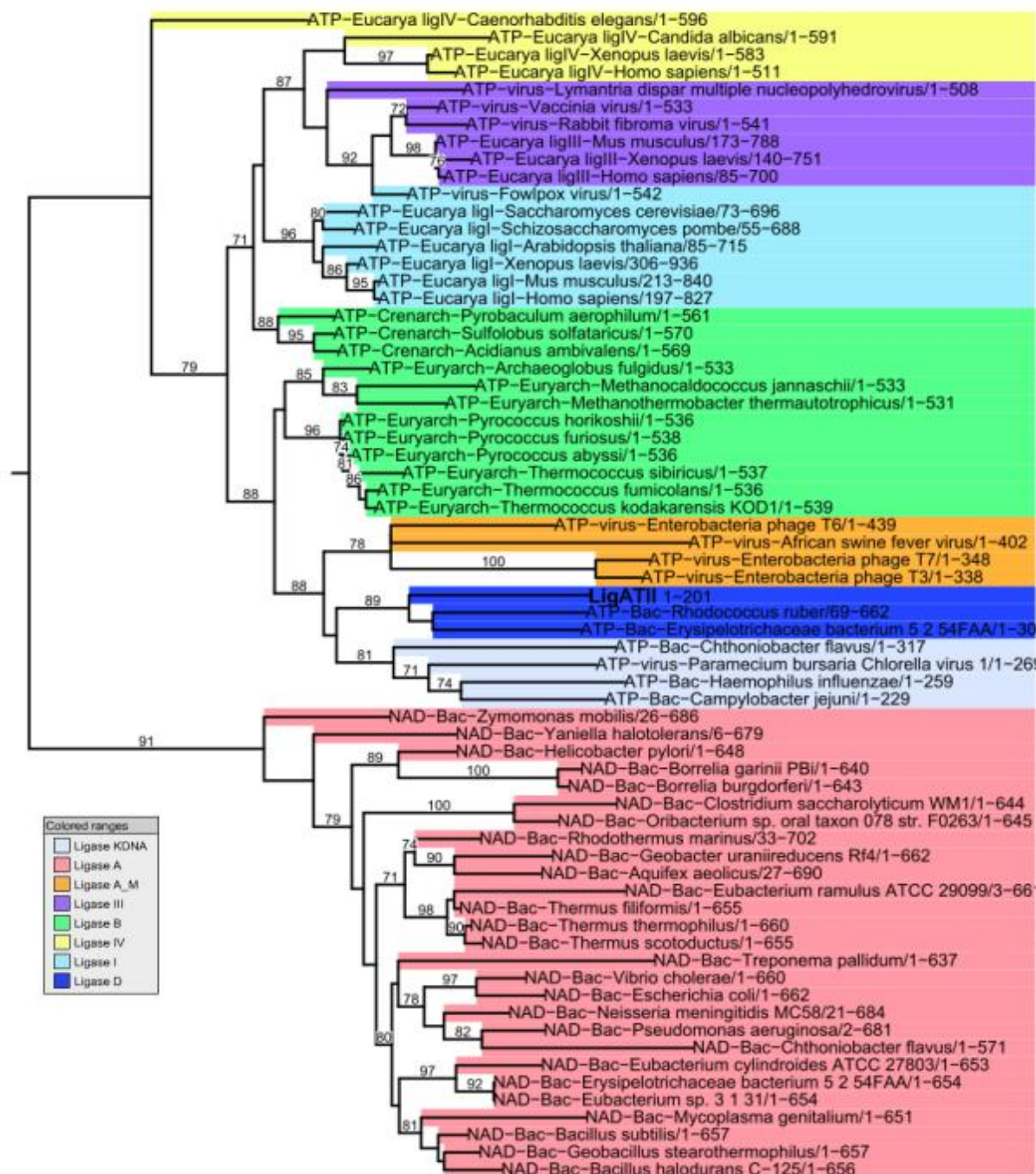


**Figure (11): Identification of functional regions in LigATII on the ConSurf server**  
 Aspartic acid (D) and Glutamic acid (E) residues were predicted to be functional, highly conserved and exposed to the surface of the protein.

## 2.6 Phylogenetic analysis of LigATII:

Since the long-established conception that bacteria have only  $\text{NAD}^+$  - dependent DNA ligase, was proven wrong by the identification and characterization of ATP-dependent ligases in many bacterial species<sup>48</sup>, Bacterial Ligases enzymes are now classified into two families ATP-dependent ligases and  $\text{NAD}^+$  -dependent ligases based on the cofactor specificity required for the ligase-adenylate formation<sup>42,48</sup>. In order to determine whether LigATII classifies as a member of one of these two families and to gain further insights about the evolutionary relations between bacterial, archeal, viral and eukaryotic ligases, a multiple sequence alignment of LigATII together with sequences of DNA ligases available in NCBI database with representatives from bacteria, virus, euryarcheota, crenarcheota and eukaryotes was performed. A phylogenetic tree was constructed by maximum likelihood estimation and LigATII grouped with members of ATP-dependent DNA ligases family (LigD) (figure 12).

The tree topology is classified into a clade of NAD-dependent sequences found exclusively in bacteria. This clade appears to be grouped closely together and independent from the ATP-dependent DNA ligases, because despite the fact that they catalyze the same reaction through similar mechanisms they are different in structure and binding systems<sup>64</sup>. In contrast, the other clade is confined to ATP-dependent ligases that is diverse among eukaryotes and prokaryotes where Bacterial ATP-dependent DNA ligases are to be found most closely related to proteins from Archaea and viruses<sup>64</sup>. Bacterial ATP-dependent ligases can be grouped into three categories based on primary sequence analysis: LigB, LigC, and LigD and the presence of these proteins within vast bacterial genomes imply specialization and a division of labor in their cellular roles, as occurs with eukaryotic ligases<sup>42</sup>. The eukaryotic ATP-dependent ligases are divided into three types (DNA ligase I, III and IV) that appear to be descended from a common ancestor<sup>55</sup>.



**Figure (12): Phylogenetic analysis tree for LigATII**

Sequences of Both: ATP-dependent and NAD-dependent DNA ligase from bacteria, archaea, viruses and eukaryotes were used to construct the tree. Branch numbering indicate bootstrap values, only above 60% values are showing. Each clade is given a certain color that corresponds to the color label box.



The eukaryotic DNA Ligase IV was initially identified to be involved in the NHEJ (non homologous end joining) pathway in eukaryotes that is strictly required for double stranded (DSB) repair and function through a complex with Ku70/80 heterodimer in a template independent manner for re-synthesis of the duplex broken strands <sup>42</sup>.

Recently, homologues of the proteins involved in NHEJ have been identified in prokaryotes <sup>49</sup>. And remarkably, these prokaryotic homologues genes (Ku homodimer protein) typically reside in operons containing a conserved ATP-dependent DNA ligase <sup>43</sup>. The phylogenetic distribution of these ligases shows that their genes are most widely distributed in Proteobacteria ( $\alpha$ ,  $\beta$ ,  $\delta$ , and  $\epsilon$  families), Actinobacteria and Firmicutes <sup>49</sup>, where the later was defined to be the phylum to which the best hit organism of LigATII belongs.

The evolutionary origin of NHEJ enzymes among bacteria is yet considered ambiguous. However the hypothesis is that NHEJ systems may most commonly be acquired by horizontal gene-transfer events since there is no obvious phylogenetic pattern between the diverse bacterial species that possess those genes <sup>43</sup>. Such functions could include specialized forms of genome evolution under conditions that lead to the accumulation of DSBs <sup>49</sup>, or it may promote genetic diversity as a structurally dependent strategy specifically adopted by prokaryotes particularly under harsh growth conditions (such as under high salt stress) where specific selection pressures may be in place <sup>60</sup>. Also when faced with prolonged exposure to harsh environments, dormant bacterial spores must repair accumulated DSB caused by ionizing radiation, and desiccation encountered during exposure to space to ensure their survival and genome integrity and owing to the fact that they contain only single copy of the genome, the NHEJ appears to be is the key player for repair mechanisms in these stages <sup>42,50</sup>.

Other physiological roles of the bacterial NHEJ pathway were described previously in bacteria, such as *H. influenzae*, *N. meningitidis* and the O157:H7 strain of *E. coli*, as they are naturally competent for transformation and have the

capacity to take up DNA from their environment and integrate it into their chromosomes and this is where the role of NHEJ appears to facilitate genome circularization of bacteriophages<sup>42,99</sup>.

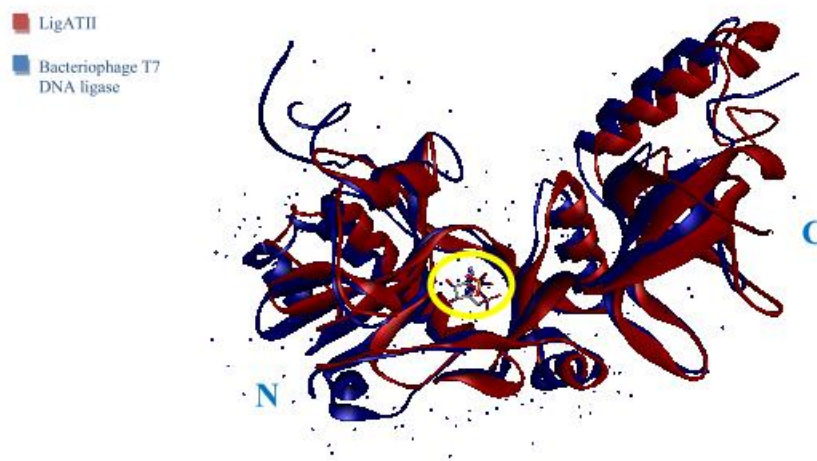
## **2.7 Comparative homology modeling of LigATII:**

To examine the structural basis and the possibility of substrate interaction within the active site pocket of LigATII the three dimensional model structure was generated. This was done by choosing the crystal structure of ATP-Dependent DNA Ligase from Bacteriophage T7 complexed with ATP adenosine Tri-phosphate (PDB ID: 1AO1|A). This structure was chosen from the list of best hit search done from BLASTP search against PDB database using LigATII as query. This template showed a sequence similarity of 40%, sequence identity of 26% and coverage of 46% of LigATII protein length. Despite the fact that there were other proteins showing higher sequence similarity and identity, they were not co-crystallized with ATP and therefore were not used to allow the prediction of the substrate interaction. And also for the reason that Bacteriophage T7 ATP dependent DNA ligase being the simplest DNA ligase that contains the catalytic core of the adenylation domain and the OB domain shared between DNA ligases<sup>56</sup>.

This three dimensional model of LigATII was built using MODELLER 9.13 through structural alignment. From the generated five models, the best model was selected based upon satisfying Discrete Optimized Protein Energy (DOPE). DOPE is the pairwise atomistic statistical potential that measures the energy of the predicted models according to the number of iterations in each generated model. Higher iteration numbers in one model ensures its reliability.

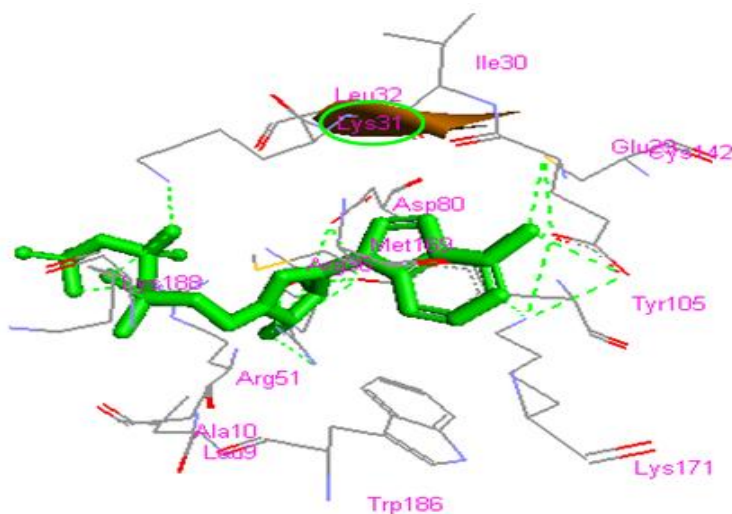
The model for LigATII was visualized by Discovery Studio® visualizer 4. The LigATII was well fitted with the ATP substrate and superimposed with the corresponding pocket of catalytic active site of the template Bacteriophage T7

(figure 13). The candidate active site interacting with ATP revealed the catalytic Lys31 involvement (figure 14).



**Figure (13): Structural superimposition of LigATII with Bacteriophage T7 DNA ligase**

The superimposition of predicted LigATII model with the template Bacteriophage T7 co-crystallized with ATP showing the antiparallel  $\beta$ -sheets in the binding pocket around the yellow circle that labels the binding ATP. The structural domains are indicated as (N) for N-terminal adenylation domain and (C) for C-terminal OB fold domain.



**Figure (14): Predicted binding site of LigATII with ATP**

In the interaction of the LigATII protein with the ATP, the green circle indicates the catalytic residue Lys31. While the green dots represent hydrogen bonds interacting with ATP.

To review the LigATII structure in light of the current crystallized structures of DNA ligases. We compared the generated model of the LigATII to the structure model of Bacteriophage T7 DNA ligase. The model structure of LigATII revealed the presence of N-terminal adenylation domain consisting mainly of anti parallel  $\beta$  sheets flanked by  $\alpha$  helices and this domain contain the ATP binding pocket beneath one of the  $\beta$  sheets. This fold displays similarity with a number of other enzymes that are bound to ATP. Many of the residues in this pocket belong to the five motifs of the six conserved motifs described in the multiple sequence alignment and form the sides of the groove that binds ATP. The other domain is a C-terminal domain, OB fold domain, which mediates the oligonucleotide recognition and is found in a diverse range of protein families. This domain is involved in a conformational change that stimulates the activity of the adenylation domain through the positioning of the ATP for the direct in line attack by the active site lysine<sup>56</sup>.

This structural comparison strongly supports that LigATII is an ATP dependent DNA ligase.

### **2.8 Molecular weight and isoelectric point prediction:**

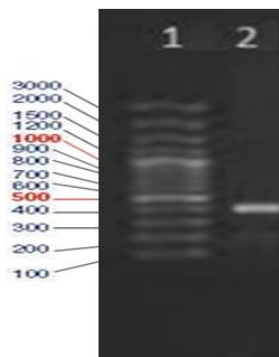
The ORF of LigATII was predicted by the Compute pI / Mw tool of the ExPASy to encode a protein of 305 amino acids with predicted molecular weight of 34 kDa and a theoretical isoelectric point of 7.58.

## **3. Isolation of the LigATII enzyme from the DNA of the Atlantis II brine pool:**

### **3.1 Screening the 1X amplified environmental DNA of the LCL of Atlantis II brine pool:**

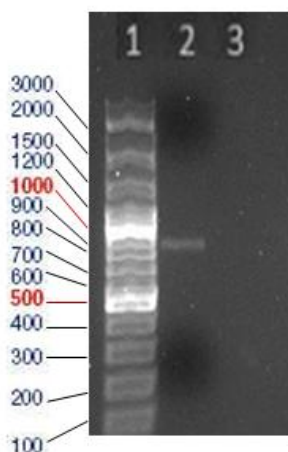
Amplification of the LigATII from the 1X whole genome amplification (WGA) DNA of the Atlantis brine pool LCL was achieved through PCR with the previously described conditions using two sets of primer pairs. The first pair of primers (contg5\_F & contg5\_R), was designed on a read in the middle of the gene

and showed amplification at expected size (~ 400 bp) (figure 15). Then a second pair of primers(ORF5\_F & ORF5\_R), was designed in order to amplify the full-length of the ORF of LigATII and showed amplification product at the expected size (~ 900 bp) (figure 16).



**Figure (15): PCR product of a single read from LigATII amplified from the 1 X WGA DNA of Atlantis II LCL**

This PCR product was generated using contg5\_F & contg5\_R primers, lane 1: GeneRuler™ 100bp plus DNA ladder (Thermo Scientific), lane 2: 400 bp amplicon.



**Figure (16): PCR product of LigATII ORF amplified from the WGA DNA**

This PCR product was generated using primer designed on ORF of LigATII, lane 1: GeneRuler™ 100bp plus DNA ladder (Thermo Scientific), lane 2: 900 bp positive amplicon using ORF5\_F & ORF5\_R primers, lane 3: negative control.

**3.2 Cloning and sequencing of the PCR product:**

The amplicon produced using the second set of primers (~ 900 bp) was cloned into pGEM<sup>®</sup>-T Easy cloning vector and transformed into electrocompetent *E.coli* Top10. The presence of the insert in the selected white colonies was confirmed by colony PCR followed by plasmid extraction from the positive clones.

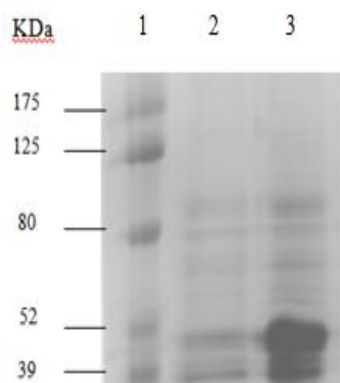
Sequencing of the extracted plasmids using pGEM<sup>®</sup>-T Easy vector promoter primers (T7 (forward) & SP6 (reverse)) to insure coverage of the ORF, in addition to ORF5\_F & ORF5\_R primers, revealed a clone that harbors the aspired gene (LigATII) identified in the database through computational analysis.

**3.3 Expression of LigATII Gene in Champion<sup>™</sup> pET SUMO<sup>®</sup> expression system:**

This expression system was used for expression of the heterologous gene LigATII in *E. coli* as a fusion to SUMO protein to take the advantage of increased expression levels as well as enhanced solubility of the recombinant protein. LigATII ORF was amplified through PCR using ORF5\_F and ORF5\_R primers which produced a PCR product of size ~900 bp with A-overhangs ligated to the IPTG-inducible Champion<sup>™</sup> pET SUMO<sup>®</sup>. This expression plasmid was used to express LigATII as N-terminal-6-Histidine-tagged-SUMO fusion recombinant protein. The protein product was therefore of predicted size 47 KDa due to the extra 11kDa SUMO fused protein and 2kDa 6-Histidine tag. Ligation products were successfully transformed into *E. coli* BL21 (DE3) cells. Colony PCR that was carried out using the SUMO Forward and ORF5\_R primers showed positive clones from which recombinant plasmids were extracted and the sequence of the inserts within were confirmed using the SUMO Forward and the T7 reverse primers of the vector.

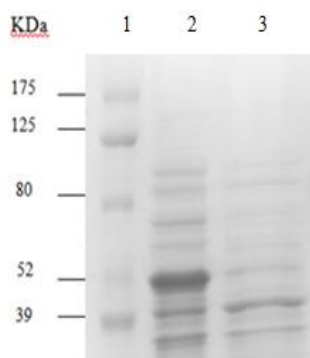
Expression was induced from the positive clone 5, where the time interval of 2 hours at 37°C with 0.5 mM IPTG concentrations showed good induction pattern (figure 17). However, these conditions showed high yield of protein trapped in

inclusion bodies of cell debris after lysis (figure 18). Various induction conditions were used from lower IPTG concentrations of 0.1mM and temperatures of 30°C & 25°C for different time intervals (O/N & 16 hours respectively) but still after cell lysis, the gel analysis revealed the presence of LigATII in inclusion bodies (Data not shown). And therefore another expression system was used to avoid the large sized SUMO fusion which is the Champion <sup>™</sup> pET100 Directional TOPO <sup>®</sup> described following.



**Figure (17): SDS-PAGE for the analysis of the expression level of LigATII-pET SUMO after induction for 2 hours at 37°C**

lane 1: ProSieve <sup>®</sup> Colour Protein Marker (Lonza), lane 2: cell lysate before addition of IPTG, lane 3: 0.5mM IPTG induction at 37°C for 2 hours.

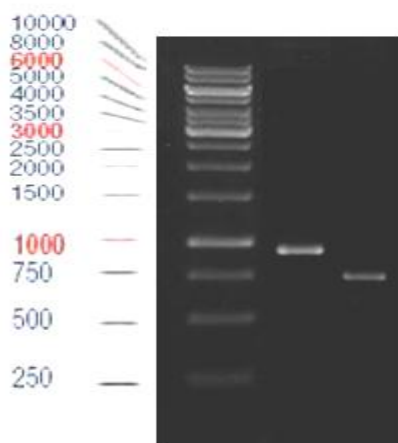


**Figure (18): SDS-PAGE for the analysis of the induced culture of LigATII-pET SUMO after cell lysis**

The culture that was induced under conditions of 0.5mM IPTG for 2 hours at 37°C was subjected to separation after cell lysis and analyzed on 12 % SDS-PAGE lane 1: ProSieve <sup>®</sup> Colour Protein Marker (Lonza), lane 2: cell debris (pellet) and lane 3: Supernatant.

### 3.4 Expression of LigATII Gene in Champion™ pET100 Directional TOPO® expression systems:

The protein LigATII was amplified using the primers pet100\_F and ORF5\_R with a proof-reading DNA polymerase. Additionally a control reaction was done which involved producing a control PCR product of size (~ 750 bp) using the reagents included in the kit and using this product directly in a TOPO® cloning reaction. Note that the pet100\_F primer is designed to contain a sequence, CACC, at the 5' end of the primer before the ATG start codon. These 4 nucleotides, CACC, base pair with the overhang sequence, GTGG, in each pET100 Directional TOPO® vector stabilizing the PCR product in the correct orientation for the directional cloning of the obtained PCR product (size ~900 bp) (figure 19).



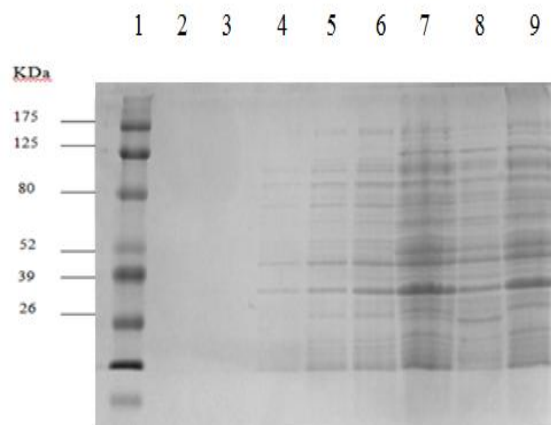
**Figure (19): PCR product from the amplification of ORF of LigATII**

PCR was performed using primers designed to amplify ORF of LigATII to be ligated into the pET100 Directional TOPO® expression vector, lane 1: GeneRuler™ 1 Kb DNA ladder (Thermo Scientific), lane 2: 900 bp positive amplicon using pet100\_F and ORF5\_R primers, lane 3: 750 bp control amplicon using primers supplied in the kit.

As a result, when this vector was transformed into *E. coli* BL21™ (DE3), the LigATII protein overexpressed from this vector contained a 6-Histag and Xpress™ epitope within an extra 4 kDa at the N-terminus accounting for a protein with a total predicted theoretical molecular weight of 38 KDa. Unfortunately, the induction of LigATII from this expression system under



conditions of IPTG concentration of 0.5 mM at 37 °C and for time interval of 2 hours did not reveal good pattern of induction. Rather it appears as no expression was induced when compared to the control sample separated before induction and treated under the same incubation conditions as the induced sample (figure 20).



**Figure (20): SDS-PAGE for the analysis of the expression level of LigATII-pET 100 Directional TOPO®**

lane 1: ProSieve ® Colour Protein Marker (Lonza), lane 2: sample from cell lysate at zero time of starting culture, lane 3: cell lysate after 2 hour, lane 4: cell lysate after 3 hours, lane 5: cell lysate after 4 hours, lane 6: cell lysate reached O.D.<sub>600nm</sub> 0.6, lane 7: sample after 0.5mM IPTG induction at 37°C for 2 hours interval , lane 8: control sample separated from the cell lysate culture at O.D.<sub>600nm</sub> 0.6, lane 9: control sample separated from the cell lysate culture at O.D.<sub>600nm</sub> 0.6 after incubation for 2 hours at 37°C. Basal level of Expression of the protein is noticed to build up even prior to addition of the inducer and after induction shows the same pattern to the cell lysate control sample left to grow for the same time interval.

This unusual pattern was thoroughly explained in the pET vectors manual. In the ampicillin resistance bearing plasmids, the direction of transcription of the drug resistance gene (*bla*) is in the same orientation and downstream as the T7 promoter. And although a T7 transcription terminator is located before the *bla*, it is only approximately 70% effective, allowing T7 RNA polymerase read-through to produce a small amount of  $\beta$ -lactamase RNA in addition to the target RNA which results in the accumulation of  $\beta$ -lactamase enzyme in induced cultures along with

the target protein as a level of basal expression of T7 RNA polymerase from the *lac* promoter even in the absence of inducer.

Also ampicillin selection tends to be lost in cultures because secreted  $\beta$ -lactamase and the drop in pH that accompanies bacterial fermentation will degrade the drug. If the inoculums already have had a fraction of cells lacking plasmid, by the time the subculture has grown to a density where expression of the target gene is to be induced, possibly only a minor fraction of the cells will contain the target plasmid. As a conclusion, target genes are poorly expressed due to the fact that only a small fraction of cells in the cultures contained plasmid. Although the manual recommended precautions to avoid this disadvantage by the addition of 1% glucose in the culture medium to delay these effects, nevertheless the induction pattern did not change as a result of this addition (Data not shown).

Other recommendation in the manual of pET100 Directional TOPO<sup>®</sup> was the use of another strain for transformation that is BL21 Star<sup>®</sup> (DE3). This strain contains in addition to the *lac* repressor, a mutation for the lack of two key proteases which reduces degradation of heterologous proteins expressed in the strain and enhances the expression capabilities. Unfortunately, this increase in expression yielded the gene toxic to the strain before reaching the saturation level of the culture.

Concluded is that the very high levels of protein expressed in these systems seem to be inappropriate for obtaining soluble and good yield of the recombinant protein LigATII. This might be attributed to the leakiness of the strong promoter T7 even in the repressed state. Recommended for use are the weaker promoter plasmids with lower plasmid copy number. Along with shifting to a more stringent host example the BL21 Star<sup>™</sup> (DE3) pLysS strain. This strain contains the pLysS plasmid, which produces T7 lysozyme that binds to T7 RNA polymerase and inhibits transcription, leading to a reduced basal level of

expression of T7-driven heterologous genes and accordingly more control on the expression system.

However since the selection of this specific pET expression systems was based upon studies reporting successful identification of active ligase proteins utilizing pET vectors<sup>39,100</sup>. This particular obscure basal level of expression were to be taken as an advantage in the functional identification of LigATII activity *in vivo* that is explained as follows.

#### **4. Functional Identification of the LigATII in the temperature sensitive mutant *E. coli* (GR501):**

On account of the vital involvement of DNA ligases in replication, its inactivation clearly leads to the non-viability of most bacteria that dependent on either one of the ligases family. Taking the example of *E.coli*, It has been identified that *E.coli* contains only NAD-dependent DNA ligase (*ligA*). Therefore, several temperature-sensitive (ts) strains of *E. coli* were isolated during the 1970s and shown to have mutations in *ligA*<sup>101</sup>. *E. coli* GR501 is one of these ligase-deficient strains that due to a mutation in *ligA*, a reduction in DNA replication is followed by DNA degradation and cell death occur at high temperatures. This strain has been useful for the analysis of functional DNA ligases that are found not only in bacteria as NAD-dependent DNA ligases, but also the ts mutation can be complemented by ATP-dependent DNA ligases that are used for replication in other systems, including human DNA ligase I<sup>39</sup> and T7 DNA ligase<sup>100</sup>.

In Doherty *et al.* study (1996), the tested clones were able to complement *E. coli* GR501 ts mutation at the high non permissive temperature of 43 °C and achieve viability for the strain. This occurred in spite of the fact that the T7 ligase gene is under the control of a T7 RNA polymerase promoter, which should not be recognized in this cell line because they lack a copy of T7 RNA polymerase in their genome. The interpretation would be that the low level of basal expression

of the T7 ligase gene from the leaky plasmid used must have been sufficient to complement the lig-deficient strain, as a result of read-through activity from the  $\beta$ -lactamase promoter on these plasmids which was discussed earlier.

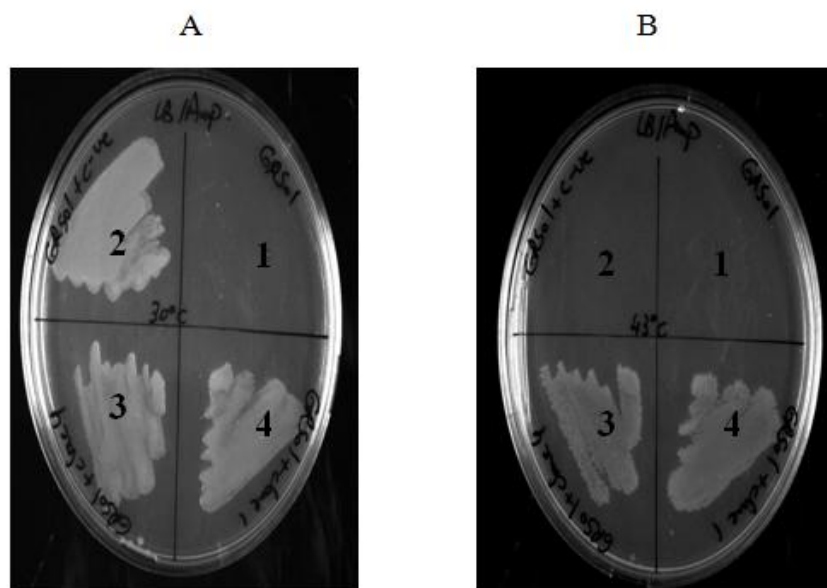
And thus following this argument the LigATII cloned in pET100 Directional TOPO<sup>®</sup> expression system was transformed, along with the control insert recombinant vector, into the *E. coli* GR501 to examine the ability of the LigATII to complement the DNA ligase deficiency in this strain and attain viability at the non permissive temperature 43 °C demonstrated in three different experiments.

Each experiment result is the average of three independent experiments and involve the use of two negative controls that were performed to confirm that complementation of growth at non-permissive temperatures was due to the expression of the functional LigATII in *E. coli* GR501. First control is the thermosensitivity of empty cells that are not harboring any vectors. Second control, to confirm that the complementation was not due to non-specific expression from the vector itself, involves the thermosensitivity from cells harbouring the control insert recombinant vector clone.

### **4.1 Assay for LigATII complementation of temperature sensitive defect of *E. coli* GR501 on agar plates:**

In the first experiment *E. coli* GR501 strains harboring the previously described control insert recombinant vector, in addition to the 5 clones of the strain harboring the LigATII/ pET 100 Directional TOPO<sup>®</sup>, were grown on LB/ Amp plates at the permissive temperature of 30 °C and the non permissive temperature of 43 °C. This assay was carried out without IPTG, as it had been shown that the vector expression system allowed a level of basal expression of the protein even in the absence of the inducer.

The result was growth observation at 30 °C for all, except cells lacking the vectors that did not have ampicillin resistance gene. In contrast, only plasmids expressing LigATII represented as clone 1 LigATII and clone 4 LigATII complemented the temperature-sensitive mutation and allowed *E. coli* GR501 to grow well on plates at 43°C. Note that these observations confirm that mutations in the NAD<sup>+</sup>-dependent DNA ligase of *E. coli* GR501 can be complemented by expression of an ATP-DNA ligase (figure 21).



**Figure (21): Growth complementation of *E. coli* GR501 streaking on LB/ Amp plates**

*E. coli* GR501 transformed with vector pET100 Directional TOPO ® expressing LigATII and control proteins was streaked onto LB agar plates containing ampicillin and grown at 30 °C (A) or 43°C (B), 1. Control 1: cells that are not harboring any vectors, 2. Control 2: cells harboring pET100 Directional TOPO ® vector cloned with control insert, 3. Cells harboring pET100 Directional TOPO ® vector cloned with clone 4 LigATII, 4. Cells harboring pET100 Directional TOPO ® vector cloned with clone 1 LigATII.

## 4.2 Assay for strain viability of the *E. coli* GR501 by the LigATII complementation of the temperature sensitive defect:

To further assess the efficiency of complementation of *E. coli* GR501 by the LigATII, we compared strain viability of the strain harboring the control and the LigATII gene cloned to the vector. Viable cell counts were determined by plating 200µl of the  $10^{-6}$  dilution of the overnight culture onto LB or LB/Amp (as required) and counting the colonies after aerobic incubation of plates at 30 °C and 43 °C O/N (Table 1). Viable counts at 43°C revealed the non viability of *E. coli* GR501 alone or with the expression of the control insert recombinant vector together with the good viability of the strain when encoding the LigATII clone 1 or clone 4. This effect was consistent in three independent experiments and show that expression of LigATII restores the viability of *E. coli* GR501.

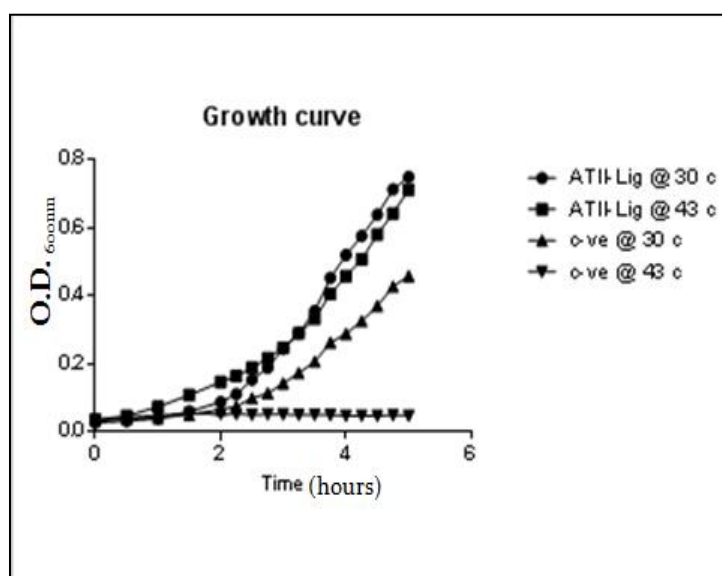
Strain/clone	viability at 30 °C CFU	Viability at 43 °C CFU
GR501 non transformed	194	0
GR501 transformed with control insert vector	124	0
GR501 transformed with LigATII clone 1 vector	294	128
GR501 transformed with LigATII clone 4 vector	250	112

**Table (1): *E. coli* GR501 strain viability at 30 °C and 43 °C**

The left-hand column provides the expressed product from the vector pET100 Directional TOPO<sup>®</sup>. Viability is expressed as the number of colony forming units per 200 µl from the  $10^{-6}$  dilution of the culture. Results represent the average of three independent experiments.

### 4.3 Assay for LigATII complementation of temperature sensitive defect of *E. coli* GR501 in liquid cultures:

To better determine the complementation of the *ts* mutation, growth of *E. coli* GR501 with or without expression of LigATII was followed in liquid culture at the permissive and non permissive temperatures. At 30°C, the growth of *E. coli* GR501 was similar in both of the strains harboring the control gene or harboring the LigATII clone 4. At 43°C, growth was dramatically reduced in the absence of the LigATII, in contrast to in its presence where the growth is normal (figure 22).



**Figure (22): Growth curves of liquid culture of *E. coli* GR501**

Growth of the *E. coli* GR501 in liquid culture with expression of LigATII cloned in pET 100 Directional TOPO™ or without its expression in the control (c-ve) expression in pET 100 Directional TOPO™ at 30 °C and 43°C. Data represent the average of three experiments under the growth conditions indicated.

The experiments on solid and in liquid media confirm that expression of LigATII is functional and can complement the temperature sensitivity mutation carried by *E. coli* GR501 strain.

## Chapter 4: Conclusion & Prospective

---

In conclusion, we here describe the identification of LigATII from the prokaryotic environmental DNA of the LCL of the Atlantis II brine pool using the metagenomic approach. The LigATII was identified to be functionally active *in vivo* by complementation of the defective ligase activity in the *E. coli* GR501 temperature sensitive strain. Sequence analysis, phylogenetic analysis and homology modeling revealed the bacterial origin and the interaction with the ATP at the active site cavity therefore provide evidence that LigATII protein is the ligase domain of a DNA ligase that is similar to members of the LigD ATP-dependent DNA ligase family.

The LigD family members of DNA ligases are involved in the NHEJ pathway for DSB repair specifically adopted by prokaryotes to repair genomes under harsh growth conditions. This characteristic of the LigATII reflect the conditions of the environment from which it was isolated where the Atlantis II brine pool is a unique high salt and high temperature stress extreme environment that is suitable for mining for biocatalysts having potential in industry. DNA ligases are one of those biocatalysts that are of great potential in biotechnological applications utilizing the Ligase chain reaction (LCR) for the diagnosis of various genetic and infectious diseases.

Attempts for expressing the isolated LigATII gene in heterologous hosts were done. However optimization through expression in a more stringent *E. coli* strain, such as BL21Star <sup>TM</sup> (DE3) pLysS strain that produces T7 lysozyme, is required to reduce the basal level of expression through the transcription inhibition of T7 RNA polymerase. Also further research focused on the *in vitro* enzymatic assay



of LigATII, in presence or absence of other NHEJ factors, is required to characterize the catalytic machinery in terms of substrate specificity, metal cofactor requirement and ligation fidelity versus mutagenesis.

Parallel efforts to investigate the effect of incubation of the LigATII protein in various ranges of temperatures and salt concentrations is required to explain the activity and stability of the enzyme and to better understand its potential applications in the biotechnology field. Additionally, an extensive phylogenetic study is required to gain insights on the uncovered evolutionary origin of the prokaryotic NHEJ as most of the bacterial ATP-dependent DNA ligases are yet not characterized.

Finally, this metagenomic study demonstrates that the Atlantis II Deep and consequently other similar extreme brine pool environments are unique mining sites for the discovery of novel biocatalysts that holds great potential in the biotechnological and industrial applications. Therefore mining for more DNA ligase enzymes, in addition to other biocatalysts, from these environments should be considered and is greatly encouraged.

## References:

1. Winckler, G. *et al.* Sub sea floor boiling of Red Sea Brines : New indication from noble gas data. *Elsevier* **64**, 1567–1575 (2000).
2. Siam, R. *et al.* Unique prokaryotic consortia in geochemically distinct sediments from Red Sea Atlantis II and discovery deep brine pools. *PloS one* **7**, e42872 (2012).
3. Winckler, G. *et al.* Constraints on origin and evolution of Red Sea brines from helium and argon isotopes. *Elsevier* **184**, 671–683 (2001).
4. Schmidt, M. *et al.* High-resolution methane profiles across anoxic brine–seawater boundaries in the Atlantis-II, Discovery, and Kebrit Deep (Red Sea). *Chemical Geology* **200**, 359–375 (2003).
5. Qian, P.-Y. *et al.* Vertical stratification of microbial communities in the Red Sea revealed by 16S rDNA pyrosequencing. *The ISME journal* **5**, 507–18 (2011).
6. Pierret, M.C. *et al.* Chemical and isotopic (  $^{87}\text{Sr} / ^{86}\text{Sr}$ ,  $\delta^{18}\text{O}$ ,  $\delta\text{D}$  ) constraints to the formation processes of Red-Sea brines. *Elsevier* **65**, 1259–1275 (2001).
7. Mapelli, F. *et al.* Microbial diversity in deep hypersaline anoxic basins. *Springer-Verlag/Wein* 21–36 (2012).
8. Research - KAUST Red Sea Expedition Spring 2010. at <<http://krse.kaust.edu.sa/spring-2010/research.html>>
9. Hartmann, M. *et al.* Hydrographic structure of brine-filled deeps in the Red Sea—new results from the Shaban, Kebrit, Atlantis II, and Discovery Deep. *Marine Geology* **144**, 311–330 (1998).
10. Swift, S. a. *et al.* Vertical, horizontal, and temporal changes in temperature in the Atlantis II and Discovery hot brine pools, Red Sea. *Deep Sea Research Part I: Oceanographic Research Papers* **64**, 118–128 (2012).
11. Gomes, J. *et al.* The Biocatalytic Potential of Extremophiles and Extremozymes. *Food Technol. Biotechnol.* **42**, 223–235 (2004).
12. Demirjian, D. C. *et al.* Enzymes from extremophiles. *Current opinion in microbiology* 144–151 (2001).
13. Egorova, K. *et al.* Industrial relevance of thermophilic Archaea. *Current opinion in microbiology* **8**, 649–55 (2005).
14. Niehaus, F. *et al.* Extremophiles as a source of novel enzymes for industrial application. *Applied microbiology and biotechnology* **51**, 711–29 (1999).

15. Thakur, N. L. *et al.* Marine molecular biology: an emerging field of biological sciences. *Biotechnology advances* **26**, 233–45 (2008).
16. Dupré, J. *et al.* Metagenomics and biological ontology. *Studies in history and philosophy of biological and biomedical sciences* **38**, 834–46 (2007).
17. Fox, G. E. *et al.* Comparative Cataloging of 16S Ribosomal Ribonucleic Acid: Molecular Approach to Procaryotic Systematics. *International Journal of Systematic Bacteriology* **27**, 44–57 (1977).
18. Amann, R. *et al.* Ribosomal RNA-targeted nucleic acid probes for studies in microbial ecology. *FEMS microbiology reviews* **24**, 555–65 (2000).
19. Amann, R. I. *et al.* Phylogenetic identification and in situ detection of individual microbial cells without cultivation . *Microbiological Reviews* **59**, 143–169 (1995).
20. Tress, M. L. *et al.* An analysis of the Sargasso Sea resource and the consequences for database composition. *BMC bioinformatics* **7**, 213 (2006).
21. Utierlinden, A. G. Profiling of complex microbial populations by denaturing gradient gel electrophoresis analysis of polymerase chain -Amplified Genes Coding for 16S rRNA. *Applied and Environmental Microbiology* **59**,695-700 (1993).
22. Liu, W. *et al.* Characterization of microbial diversity by determining terminal restriction fragment length polymorphisms of genes encoding 16S rRNA. *Applied and Environmental Microbiology* **63**, 4516–4522 (1997).
23. Handelsman, J. *et al.* Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. *Chemistry & biology* **5**, R245–9 (1998).
24. Kerkhof, L. J. *et al.* Ocean microbial metagenomics. *Deep Sea Research Part II: Topical Studies in Oceanography* **56**, 1824–1829 (2009).
25. Stein, J. L. *et al.* Characterization of uncultivated prokaryotes : isolation and analysis of a 40-kilobase-pair genome fragment from a planktonic marine archaeon . *Journal of Bacteriology* **178**,591–599 (1996).
26. Liu, H. *et al.* A BAC clone-based physical map of ovine major histocompatibility complex. *Genomics* **88**, 88–95 (2006).
27. Forde, B. M. *et al.* Next-generation sequencing technologies and their impact on microbial genomics. *Briefings in functional genomics* **12**, 440–53 (2013).
28. Sogin, M. L. *et al.* Microbial diversity in the deep sea and the underexplored “rare biosphere”. *Proceedings of the National Academy of Sciences of the United States of America* **103**, 12115–20 (2006).
29. Huse, S. M. *et al.* Exploring microbial diversity and taxonomy using SSU rRNA hypervariable tag sequencing. *PLoS genetics* **4**, e1000255 (2008).

30. Morozova, O. *et al.* Applications of next-generation sequencing technologies in functional genomics. *Genomics* **92**, 255–64 (2008).
31. Zhang, J. *et al.* The impact of next-generation sequencing on genomics. *Journal of genetics and genomics = Yi chuan xue bao* **38**, 95–109 (2011).
32. Delong, E. F. *et al.* Community genomics among stratified microbial assemblages in the ocean's interior. *JSTOR* **311**, 496–503 (2013).
33. Steele, H. L. *et al.* Advances in recovery of novel biocatalysts from metagenomes. *Journal of molecular microbiology and biotechnology* **16**, 25–37 (2009).
34. Schmeisser, C. *et al.* Metagenomics, biotechnology with non-culturable microbes. *Applied microbiology and biotechnology* **75**, 955–62 (2007).
35. Uria, A. R. *et al.* Novel molecular methods for discovery and engineering of biocatalysts from uncultured. *Journal of Coastal Development* **8**, 53–71 (2005).
36. Singh, J. *et al.* Metagenomics: Concept, methodology, ecological inference and recent advances. *Biotechnology journal* **4**, 480–94 (2009).
37. Lorenz, P. *et al.* Metagenomics and industrial applications. *Nature* **3**, 510–517 (2005).
38. Simon, C. *et al.* Achievements and new knowledge unraveled by metagenomic approaches. *Applied microbiology and biotechnology* **85**, 265–76 (2009).
39. Lavesa-Curto, M. *et al.* Characterization of a temperature-sensitive DNA ligase from *Escherichia coli*. *Microbiology (Reading, England)* **150**, 4171–80 (2004).
40. Uchiyama, T. *et al.* Functional metagenomics for enzyme discovery: challenges to efficient screening. *Current opinion in biotechnology* **20**, 616–22 (2009).
41. Yoon, S. H. *et al.* Secretory production of recombinant proteins in *Escherichia coli*. *Recent patents on biotechnology* **4**, 23–9 (2010).
42. Pitcher, R. S. *et al.* Nonhomologous End-Joining in Bacteria : A Microbial Perspective. *Annual Review of Microbiology* **61**, 259-82 (2007).
43. Brissett, N. C. *et al.* Repairing DNA double-strand breaks by the prokaryotic non-homologous end-joining pathway. *Biochemical Society* **37**, 539–545 (2009).
44. Nakatani, M. *et al.* A DNA ligase from a hyperthermophilic archaeon with unique cofactor specificity. *Journal of bacteriology* **182**, 6424–33 (2000).
45. Nakatani, M. *et al.* Substrate recognition and fidelity of strand joining by an archaeal DNA ligase. *European journal of biochemistry / FEBS* **269**, 650–6 (2002).
46. Kaczmarek, F. S. *et al.* Cloning and Functional Characterization of an NAD<sup>2</sup> -Dependent DNA Ligase from *Staphylococcus aureus*. *Journal of Bacteriology* **183**, 3016–3024 (2001).

47. Gong, C. *et al.* Biochemical and Genetic Analysis of the Four DNA Ligases of Mycobacteria. *The journal of biological chemistry* **279**,20594-20606 (2004).
48. Shuman, S. *et al.* Bacterial DNA repair by non-homologous end joining. *Nature* **5**, 852-861(2007).
49. Bowater, R *et al.*. Making Ends Meet : Repairing Breaks in Bacterial DNA by Non-Homologous. *Plos Genetics* **2**, 0093-0099 (2006).
50. Moeller, R. *et al.* Role of DNA Repair by Nonhomologous-End Joining in Bacillus subtilis Spore Resistance to Extreme Dryness , Mono- and Polychromatic UV , and Ionizing Radiation. *Journal of Bacteriology* **189**, 3306-3311 (2007).
51. Zhu, H. *et al.* Bacterial Nonhomologous End Joining Ligases Preferentially Seal Breaks with a 3' OH Monoribonucleotide. *The Journal of biological chemistry* **283**, 8331-8339 (2008).
52. Shuman, S. DNA ligases: progress and prospects. *The Journal of biological chemistry* **284**, 17365–9 (2009).
53. Tomkinson, A. E. *et al.* DNA ligases: structure, reaction mechanism, and function. *Chemical reviews* **106**, 687–99 (2006).
54. Akey, D. *et al.* Recombination : Crystal Structure and Nonhomologous End-joining Function of the Ligase Component of Mycobacterium DNA Ligase D. *The Journal of biological chemistry* **281**, 13412-13423 (2006).
55. Martin, I. V. *et al.* Protein family review ATP-dependent DNA ligases. *Genome biology* **3**, 1–7 (2002).
56. Doherty, a J. *et al.* Structural and mechanistic conservation in DNA ligases. *Nucleic acids research* **28**, 4051–8 (2000).
57. Nishida, H. *et al.* The closed structure of an archaeal DNA ligase from Pyrococcus furiosus. *Journal of molecular biology* **360**, 956–67 (2006).
58. Cheng, C. *et al.* Characterization of an ATP-dependent DNA ligase encoded by Haemophilus influenzae. *Nucleic acids research* **25**, 1369–74 (1997).
59. Nair, P. A. *et al.* Structure of bacterial LigD 3' -phosphoesterase unveils a DNA repair superfamily. *PNAS* **107**, 12822-12827 (2010).
60. Kobayashi, H. *et al.* Multiple Ku orthologues mediate DNA non-homologous end-joining in the free-living form and during chronic infection of Sinorhizobium meliloti. *Molecular Microbiology* **67**, 350–363 (2008).
61. Zhu, H. *et al.* Substrate Specificity and Structure-Function Analysis of the 3' Phosphoesterase Component of the Bacterial NHEJ Protein , DNA Ligase D. *The journal of biological chemistry* **281**, 13873-13881 (2006).

62. Stephanou, N. C. *et al.* Mycobacterial Nonhomologous End Joining Mediates Mutagenic Repair of Chromosomal Double-Strand DNA Breaks. *Journal of bacteriology* **189**, 5237- 5246 (2007).
63. Lai, X. *et al.* Biochemical characterization of an ATP-dependent DNA ligase from the hyperthermophilic crenarchaeon *Sulfolobus shibatae*. *Extremophiles : life under extreme conditions* **6**, 469–77 (2002).
64. Wilkinson, A. *et al.* MicroReview Bacterial DNA ligases. *Molecular Microbiology* **40**, 1241–1248 (2001).
65. Lim, J. H. *et al.* Molecular cloning and characterization of thermostable DNA ligase from *Aquifex pyrophilus*, a hyperthermophilic bacterium. *Extremophiles : life under extreme conditions* **5**, 161–8 (2001).
66. Pascal, J. M. *et al.* A flexible interface between DNA ligase and PCNA supports conformational switching and efficient ligation of DNA. *Molecular cell* **24**, 279–91 (2006).
67. Odell, M. *et al.* Crystal structure of eukaryotic DNA ligase-adenylate illuminates the mechanism of nick sensing and strand joining. *Molecular cell* **6**, 1183–93 (2000).
68. Barany, F. Genetic disease detection and DNA amplification using cloned thermostable ligase. *PNAS* **88**, 189–93 (1991).
69. Qi, X. *et al.* L-RCA (ligation-rolling circle amplification): a general method for genotyping of single nucleotide polymorphisms (SNPs). *Nucleic acids research* **29**, E116 (2001).
70. Barany, F. The ligase chain reaction in a PCR world. *Genome Research* **1**, 5–16 (1991).
71. Gerry, N. P. *et al.* Universal DNA microarray method for multiplex detection of low abundance point mutations. *Journal of molecular biology* **292**, 251–62 (1999).
72. Seo, M. S. *et al.* Cloning and expression of a DNA ligase from the hyperthermophilic archaeon *Staphylothermus marinus* and properties of the enzyme. *Journal of biotechnology* **128**, 519–30 (2007).
73. Wiedmann, M. *et al.* Ligase Chain Reaction (LCR) -Overview and Applications. *Cold Spring Harbor Laboratory Press* **94**, 1054-9805 (1994).
74. Khanna, M. *et al.* Ligase detection reaction for identification of low abundance mutations. *Clinical biochemistry* **32**, 287–90 (1999).
75. Lehman, T. A. *et al.* Detection of K- ras Oncogene Mutations by Polymerase Chain Reaction-Based Ligase Chain Reaction. *Analytical Biochemistry* **239**, 153–159 (1996).
76. Harden, S. V. Real-Time Gap Ligase Chain Reaction: A Rapid Semiquantitative Assay for Detecting p53 Mutation at Low Levels in Surgical Margins and Lymph Nodes from Resected Lung and Head and Neck Tumors. *Clinical Cancer Research* **10**, 2379–2385 (2004).

77. Stary, A. *et al.* Comparison of ligase chain reaction and culture for detection of *Neisseria gonorrhoeae* in genital and extragenital specimens. *Journal of clinical microbiology* **35**, 239–42 (1997).
78. Chernesky, M. A. *et al.* Diagnosis of *Chlamydia trachomatis* infections in men and women by testing first-void urine by ligase chain reaction. *Journal of clinical microbiology* **32**, 2682–5 (1994).
79. Lindbråthen, A. *et al.* Direct detection of *Mycobacterium tuberculosis* complex in clinical samples from patients in Norway by ligase chain reaction. *Journal of clinical microbiology* **35**, 3248–53 (1997).
80. Wiedmann, M. *et al.* Discrimination of *Listeria monocytogenes* from Other *Listeria* Species by Ligase Chain Reaction. *Applied and Environmental Microbiology* **58**, 3443–3447 (1992).
81. Rondini, S. *et al.* Development of multiplex PCR-ligase detection reaction assay for detection of West Nile virus. *Journal of clinical microbiology* **46**, 2269–79 (2008).
82. Osioy, C. Sensitive Detection of HBsAg Mutants by a Gap Ligase Chain Reaction Assay. *Journal of Clinical Microbiology* **40**, 2566–2571 (2002).
83. Marshall, R. L. *et al.* Detection of HCV RNA by the asymmetric gap ligase chain reaction. *Cold Spring Harbor Laboratory Press* **94**, 1054–9803 (1994).
84. Stewart, J. *et al.* A quantitative assay for assessing allelic proportions by iterative gap ligation. *Nucleic acids research* **26**, 961–6 (1998).
85. Tian, F. *et al.* A new single nucleotide polymorphism genotyping method based on gap ligase chain reaction and a microsphere detection assay. *Clinical chemistry and laboratory medicine : CCLM / FESCC* **46**, 486–9 (2008).
86. Psifidi, A. *et al.* Novel quantitative real-time LCR for the sensitive detection of SNP frequencies in pooled DNA: method development, evaluation and application. *PloS one* **6**, e14560 (2011).
87. Li, J. *et al.* A colorimetric method for point mutation detection using high-fidelity DNA ligase. *Nucleic acids research* **33**, e168 (2005).
88. Noguchi, H. *et al.* MetaGeneAnnotator : Detecting Species-Specific Patterns of Ribosomal Binding Site for Precise Gene Prediction in Anonymous Prokaryotic and Phage Genomes. *DNA Research* **15**, 387–396 (2008).
89. Punta, M. *et al.* The Pfam protein families database. *Nucleic acids research* **40**, D290–301 (2012).
90. Rutherford, K. *et al.* Artemis : sequence visualization and annotation. *Bioinformatics Applications Notes* **16**, 944–945 (2000).
91. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST : a new generation of protein database search programs. *Nucleic Acids Research* **25**, 3389–3402 (1997).

92. Gasteiger, E. ExPASy: the proteomics server for in-depth protein knowledge and analysis. *Nucleic Acids Research* **31**, 3784–3788 (2003).
93. Marchler-Bauer, A. *et al.* CDD: a conserved domain database for interactive domain family analysis. *Nucleic acids research* **35**, D237–40 (2007).
94. Sievers, F. *et al.* Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular systems biology* **7**, 539 (2011).
95. Glaser, F. *et al.* ConSurf: Identification of Functional Regions in Proteins by Surface-Mapping of Phylogenetic Information. *Bioinformatics Applications Notes* **19**, 163–164 (2003).
96. Dereeper, A. *et al.* Phylogeny.fr: robust phylogenetic analysis for the non-specialist. *Nucleic acids research* **36**, W465–9 (2008).
97. Bougouffa, S. *et al.* Distinctive microbial community structure in highly stratified deep-sea brine water columns. *Applied and environmental microbiology* **79**, 3425–37 (2013).
98. Sayed, A. *et al.* A novel mercuric reductase from the unique deep brine environment of Atlantis II in the Red Sea. *The Journal of biological chemistry* **289**, 1675–87 (2014).
99. Pitcher, R. S. *et al.* Mycobacteriophage exploit NHEJ to facilitate genome circularization. *Molecular cell* **23**, 743–8 (2006).
100. Doherty, A. J. *et al.* Bacteriophage T7 DNA Ligase. *The Journal of biological chemistry* **271**, 11083–11089 (1996).
101. Konrad, E. B. *et al.* Genetic and Enzymatic Characterization of a Conditional Lethal Mutant of *Escherichia coli* K12 with a Temperature-sensitive DNA Ligase. *Journal of Molecular Biology* **77**, 519–529 (1973).