MINING OF GENES ENCODING FOR DNA-MANIPULATING ENZYMES FROM HOT SPRINGS USING METAGENOMIC TECHNIQUES

A thesis submitted in fulfilment of the requirements for the degree of Masters of Technology, in the Department of Biotechnology, Faculty of Applied and Computer Sciences, at Vaal University of Technology.

MOKOENA MORENA INDIA (211125326)

SUPERVISOR: Dr NA Feto

CO-SUPERVISOR: Dr K Rashamuse

September 2019

OMICS

OMICS Research Group

Vaal University of Technology

Navigating our future through OMICS

DECLARATION

I, Mokoena Morena India, student no: 211125326, hereby attest that this thesis entitled “Mining of Genes Encoding for DNA Manipulating Enzymes from Hot Springs using Metagenomics Techniques” submitted in the fulfilment of the requirements for the degree of masters of technology in Biotechnology at Vaal University of Technology was composed solely by myself and the work contained herein is my own except explicitly stated otherwise. Being my original work, it has never been submitted for a degree in any other university or any professional qualification.

………………………………………

STATEMENT 1

This dissertation is being submitted in fulfilment of the requirements for the degree of Magister Technologiae Biotechnology.

Signed......................... Date ………………….

STATEMENT 2

This dissertation is the result of my own independent investigation, except where otherwise stated. Other sources are acknowledged by giving explicit references. A bibliography is appended.

Signed ………………. Date …………………

STATEMENT 3

I hereby give consent for my dissertation subject to approval by my supervisor, if accepted, to be available for photocopying and for interlibrary loan, and for title summary to be made available to outside organisations.

Signed ……………... Date…………………

ACKNOWLEDGEMENTS

I would firstly like to thank the most High God for giving me the opportunity to study, for the strength, wisdom and patience beyond measure. For the gift of life, relationships, and for keeping me throughout my studies. It wasn’t an easy one but I made through YOU. To Dr NA Feto, thank you very much sir, this research would not have been possible if it was not for your supervision, mentorship, emotional support and guidance throughout the journey. Your criticism and love really shaped me to be a better somebody. I am very grateful for every assistance and for always pushing me towards greatness. I would also want to appreciate my co-supervisor Dr K. Rashamuse for the all the contributions and support towards completion of this study. I thank Dr. A Mitema, a postdoc working under Dr. Feto’s supervision, for his review of my thesis and for all his contributions to my work. God bless you sirs!!

Thanks to the Hub and Spokes scholarship, TIA-SABDI and VUT award for their financial support. Thanks to the Vaal University of Technology and Biotechnology Department for opening their laboratories for me to carry out my research. To the OMICS research group, thank you for your warmth and love always. Special thanks to my friends and fellow post-graduate students, Lesego, Grace, Judith, Gladys, Nthabiseng, Dikole, Falatsi, Bakoena and Ms Baburum. You guys have all been a blessing, thank you so much. Another special thanks to Mr TS Motloi, Lab Technician at OMICS Research facility for his support towards completion of this work.

To my loving and caring parents, Mr. DM Mokoena and Mrs. RM Mokoena (Mama le Papa), I am deeply humbled by your constant love, care, support and your prayers rendered to God on my behalf. Thank you for always believing in my dreams no matter what and for never giving up on me. Thank you for always having my back and those phone calls (lol). From the deepest parts of my heart I thank you and hope to always make you proud of me. To my every sibling, Mike, Ntsekiseng, Kedibone, Paena and Palesa thank you for everything guys.

Mining of Genes Encoding for DNA-Manipulating Enzymes from Hot Springs using Metagenomic Techniques

General Abstract

The use of conventional culture-based approach results in vast majority of microbes (90 - 99%) unaccounted for. However, over the past years, the use of metagenomics, which is a culture-independent comprehensive approach has enabled researchers to access nearly 100% of the microbiome. In this study, three hot springs (44 – 70 ºC) in Limpopo province of South Africa were investigated as potential sources of genes encoding for DNA-manipulating enzymes (DNA polymerase, DNA ligase and endonuclease), which are central in genetic engineering. They are usually grouped into four broad classes (nucleases, ligases, polymerases and modifying enzymes) depending on the type of the reaction they catalyze. Accordingly, hot spring metagenomic DNA was successfully extracted using modified SDS-CTAB method involving gel purification and electroelution. Consequently, a portion of the extracted metagenomic DNA was used for sequencing and another for fosmid library construction. Sequencing was done using Illumina MiSeq next generation sequencing platform and sequence data analyzed and *de novo*-assembled using CLC Genomic Workbench, which resulted in 5 681 662 reads and 7 338 contigs. A metagenome expression fosmid library of approximately 2.16 x 10^3 clones was also constructed using CopyControl™ HTP Fosmid Library Production Kit with pCC2FOS™ Vector. A BLAST algorithm in NCBI revealed 57 distinct genes for DNA polymerase, 29 genes for DNA ligase and more than 100 genes for endonuclease II enzymes. Hence, three genes related to thermophiles representing genes for DNA polymerase, DNA ligase and endonuclease II were selected. Accordingly, the three genes were codon-optimized, synthesized and successfully cloned into pET- 30a (+) and overexpressed in *Escherichia coli* BL21 (DE3) by inducing with 0.5 mM IPTG and incubating overnight at 16ºC. The cells were lysed using B-PER Reagent, protein extracted and purified using AKTA start protein purification system and purity of 85- 95 % was achieved. From this study, it can be concluded that metagenomics as an approach, can be used to mine for putative DNA-manipulating enzymes from hot spring metagenome. Besides, further study should be conducted to formulate the developed DNA-manipulating enzymes and study the practical application and chart way for commercialization. Moreover, the constructed fosmid library could also be screened for potentially novel thermo-stable biomolecules of industrial and therapeutic importance.

Key words: Hot spring, metagenomic library, DNA-manipulating enzymes, Sequencing, Cloning, Expression.

TABLE OF CONTENTS

Chapter 1: A review on DNA manipulation enzymes and metagenomics as a strategy for mining for biomolecules from hot spring.

Chapter 2: DNA Extraction, Sequencing and Metagenomic Fosmid Library Construction

Chapter 3: Expression and Purification of DNA Manipulating Enzymes

List of Figures **Page**

List of Tables **Page**

ABBREVIATIONS

aa	Amino acid
APS	Ammonium percolate
BACs	Bacterial artificial chromosome
bp	Base pair
BSA	Bovine serum albumin
BLAST	Basic local alignment search tool
º C	Degrees Celsius
Cfu	Colony forming units
ATP	Adenosine Triphosphate
C-terminus	Carboxy terminus
ddH2O	Deionized distilled water
DNA	Deoxyribonucleic acid
dNTP	Deoxynucleotide triphosphate
E. coli	*Escherichia coli*
EDTA	Ethylene diamine tetra-acetic acid
g	Grams
His	Histidine
NADH	Nicotinamide adenine dinucleotide
h	Hour(s)
IMAC	Immobilized metal affinity chromatography
IPTG	Isopropyl β-D-thiogalactosidase
	Catalytic turnover number
kDa	Kilo Dalton
LB	Luria-Bertani
µg	Microgram
µL	Microliter
mL	Milliliter
MW	Molecular weight

N-terminus	Amino-terminus
OD600	Optical density at 600 nanometers
ORF	Open reading frame
PCR	Polymerase chain reaction
rpm	Revolutions per minute
SDS-PAGE	Sodium dodecyl sulphate polyacrylamide gel electrophoresis
TE	Tris/EDTA
TBE	Tris-borate-EDTA
Tris-HCl	Tris (hydroxymethyl) methylamine hydrochloride
v/v	Volume to volume
w/v	Weight to volume

Chapter 1: A review on DNA manipulating enzymes and metagenomics as a strategy for mining enzymes.

Table of Contents

CHAPTER ONE: A review on DNA manipulating enzymes and metagenomics as a strategy for mining of such enzymes from hot spring .

1.1. INTRODUCTION

A hot spring can be defined as a release of naturally hot groundwater that appears at the surface as a stream of flowing water (Todd 1980). This water body is typically heated by subterranean volcanic activity or from rainwater that infiltrates into the underground permeable rocks or via faults, then get heated up by hot temperatures in the earth surface (Olivier et al. 2008). The temperature in hot springs usually excludes eukaryotic life (near to 60 °C), which limits the microbial life to bacteria, archaea and viruses (Lopez'- Lopez' 2013). Hot spring like many extreme environments harbours potentially essential and untapped microbial communities for use as resources for biotechnological processes and products (Olivier et al. 2011). They are good sources of potentially unexploited thermostable enzymes of medical, pharmaceutical, industrial and biotechnological relevance, such as DNA manipulating enzymes.

In recent decades, microbial community studies including functional diversity and phylogeny relied mostly upon culture-based techniques, which might exclude about 90 - 99% of the bacterial community in most environmental samples (Fuhrman 2012). This drawback can be linked to the methods of plating, the composition of the cultivation medium, and growth conditions (Fuhrman 2012). Consequently, a better approach that allows for studying and exploitation of environmental samples to discover, isolate and characterize functional genes is necessary. Metagenomics, which can be defined as the culture-independent genomic analysis of microbial communities can be used as a better alternative. This approach of genomics can, in theory, identify genes of any sequence and function from environmental samples (Schloss and Handelsman 2003). It also allows for construction of DNA libraries that can be screened to discover genes of interest, which can be expressed in a host for a specific trait (Schloss and Handelsman 2003). Therefore, the major aim of the study was to deploy metagenomics approach as a tool to mine genes encoding for DNA-manipulating enzymes from a hot spring. We also cloned the genes of interest in pET- 30a (+), expressed and purified the recombinant protein.

1.2. DISCUSION

1.2.1. Hot spring

A spring can be defined as a concentrated release of ground-water that appears at the surface as a stream of flowing water (Todd 1980). When the temperature of the spring is above that of ordinary groundwater, it is referred to as a thermal spring. It is a natural geological phenomenon found in all continents (Olivier et al. 2011). Hot springs occur as a result of either recent volcanic activity or from rainwater that infiltrates into the underground permeable rocks or via faults, or fractures of less porous rocks (Olivier et al. 2008). The springs are all of meteoric origin and range from warm to boiling in temperature. The mineral composition of the hot spring waters reflects the geological formations that occur at a depth of the source of the spring water rather than surface geology, since nearby springs have unique chemical properties (Olivier et al. 2011). This implies that springs located in the same area may not have the same development potential. Therefore, one hot spring is different from others with regards to chemical composition, temperature and its gradients of light (Lopez'- Lopez' 2013).

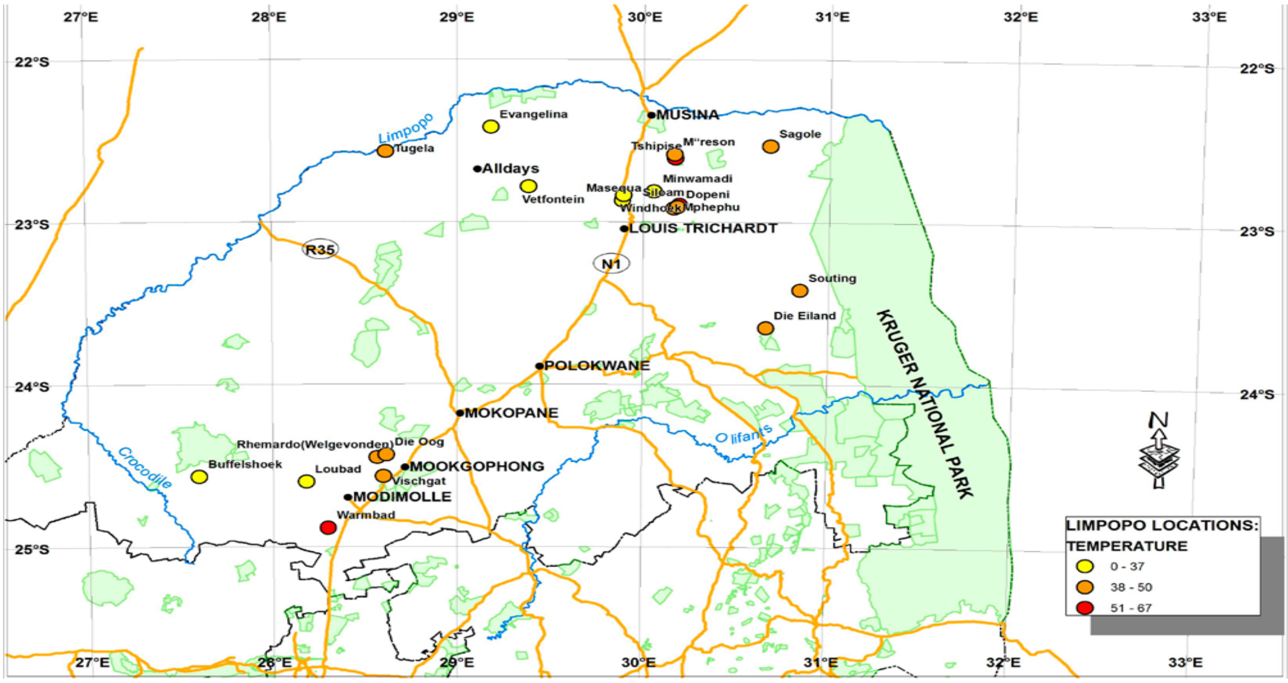


Figure 1.1. Location of hot springs in Limpopo Province, South Africa (Olivier et al. 2011).

To date, there are about 90 identified thermal hot springs in South Africa (Olivier et al. 2011). Limpopo province has more thermal springs than any other areas with the springs such as, Mphephu, Siloam, Sagole, Tshipise and many others (Tekere et al. 2015). Physical and chemical characteristics of Mphephu, Tshipise and Siloam are compared in **Table 1.1.** Thermal springs in many countries are utilised for different reasons including power generation, space heating, and industrial purposes among others (Jonker et al. 2013). Several handfuls of hot springs in South Africa have been developed mainly for recreation, tourism and some are bottled and sold for their therapeutic purposes (Olivier et al. 2011). Unlike other hot springs, South African thermal springs are some of the under-utilised and under-researched natural resources. However, due to increasing interest in the value of hot springs worldwide, there is a belief that local interest will also be developed (Olivier et al. 2011).

The first research made on hot springs focused mainly on their physicochemical and geological characteristics as opposed to their biological properties (Tekere et al. 2015). Hot springs like many other natural environments harbour a large number of microbial diversity most of which have not yet been characterised (Olivier et al. 2011). The temperature in hot springs is usually near 60 °C which excludes eukaryotes and limits the microbial life to bacteria, archaea and viruses (Lopez'- Lopez' 2013). These microbial populations are collectively referred to as thermophilic microorganisms.

1.2.2. Thermophiles and Thermostable biomolecules

Thermophiles are heat-loving microorganisms that can grow optimally between 55 and 80 °C, while hyperthermophiles grow above 80 °C, but these are only approximate figures (Zahoor et al. 2012). Hyperthermophiles were first isolated in the 1960s from hot springs situated in Yellowstone National Park (Zahoor et al. 2012). Some hyperthermophiles have been recorded to require temperatures of at least 90°C for survival (Zierenberg et al. 2000). Because thermophiles can grow or tolerate high temperatures, it is logical to expect that they must contain some metabolites that can function at high temperatures. These thermo-enzymes are usually not only thermostable but also active at high salinity and extreme pH (Gomez and Steiner 2004).

Table 1.0. The physical and chemical characteristics of water at Tshipise, Siloam and Mphephu hot springs*

	Hot Spring		
Parameter	**Mphephu**	**Tshipise**	**Siloam**
Temperature (º C)	43	58	45
DO (%)	65.3	34.7	9.9
pH	8.08	8.85	9.70
TDS (ppm)	199.36	460.56	203.76
Conduct. (mS/m)	44.00	81.00	39.00
Sodium (mg/l)	44.37	156.31	65.15
Potassium (mg/l)	1.14	4.25	1.10
Calcium (mg/l)	13.73	5.58	1.31
Magnesium (mg/l)	11.25	0.17	0.07
Fluoride (mg/l)	3.16	5.63	1.01
Nitrate (mg/l)	2.12	0.61	0.00
Chloride (mg/l)	39.38	168.97	47.85
Sulphate (mg/l)	9.26	53.17	18.20
Phosphate (mg/l)	0.00	0.00	0.00
Carbonate (mg/l)	0.00	6.00	18.00
Bicarbonate (mg/l)	151.28	126.88	102.48

* Source: Tekere et al. (2012).

The biomolecules recovered from some thermophiles have proven to be of great use in the modern fields of biotechnology (e.g. heat stable DNA polymerases that are utilized in polymerase chain reaction), medicine, industry and in surfactants because they can function under conditions that would denature enzymes taken from most organisms (Zahoor et al. 2012). The need for such thermostable biomolecules in the growing field of biotechnology has encouraged research into organisms that are capable of growth at high temperatures. In the last two decades, many scientific studies have described the isolation of novel thermophilic microorganisms from both archaea and bacteria. Bacterial thermophiles have received attention for their potential in the conversion of substrates of plant origin to end products such as ethanol, fuels and compounds with potential for the production of bulk chemicals (Zahoor et al. 2012).

1.2.3. Strategies for Gene Discovery

Microorganisms are ubiquitous in nature; they exist almost everywhere even where it seems like there is no evidence of life. Microbes are essential to the functioning of all the earth's ecosystems and are dominant in most biogeochemical cycles (Fuhrman 2012), so it is crucial to understand how they exist and function in nature. A vast diversity of microorganisms offer a significant amount of genetic pool that can be exploited to isolate novel genes (Culligan et al. 2014). These genes recovered can also be used further to characterise new biomolecules (Ferrer et al. 2008). The ability to capture the whole genetic pool from a natural environment can help researchers to understand the microbial diversity better, as well as make screening for biomolecules of interest and other products practical. Two methods can be used to access genes from the environment; culture dependent and culture independent.

1.2.3.1.Culture-dependent methods

Conventional cultivation methods rely on growing cultures on the plate, for isolation, identification and characterisation of microorganisms from environmental samples (Dias et al. 2014). In a culture based method, novel genes are isolated from animal tissues, plants and organisms. Microorganisms are cultured under their suitable growth conditions to obtain desired traits. Previous studies which relied upon the cultivation of microbes are limiting as they missed about 90-99% of microorganisms in most environments, with rare exceptions (Fuhrman 2012). These limitations can be as a results plating techniques, the composition of growth medium, and growth conditions (Singh et al. 2009; Dias et al. 2014). Studies have been conducted to improve cultivation techniques of microorganisms that are incapable of growth in cultivation conditions that simulate their natural environment, but still, the proportion of uncultivated to cultivated bacteria remain high (Neelakanta and Sultana 2013).

1.2.3.2. Culture-independent methods

The limitations of culture-based techniques led to the development of culture independent method called metagenomics amongst others. Metagenomics is defined as the culture-independent genomic analysis of microbial communities. This approach allows access to desired proteins or biomolecules directly from environmental samples without a need of cultivation. The term is built from statistical concept of meta-analysis (the process of combining statistically separate analysis) and the word genomics (the full breakdown of an organism's genetic material) (Rondon et al. 2000). The term was first used in the year 1988 in a study by Handelsman et al. (1988). Metagenomics has since been applied in microbial ecology studies and extraction of the whole microbial genome in a complex environment.

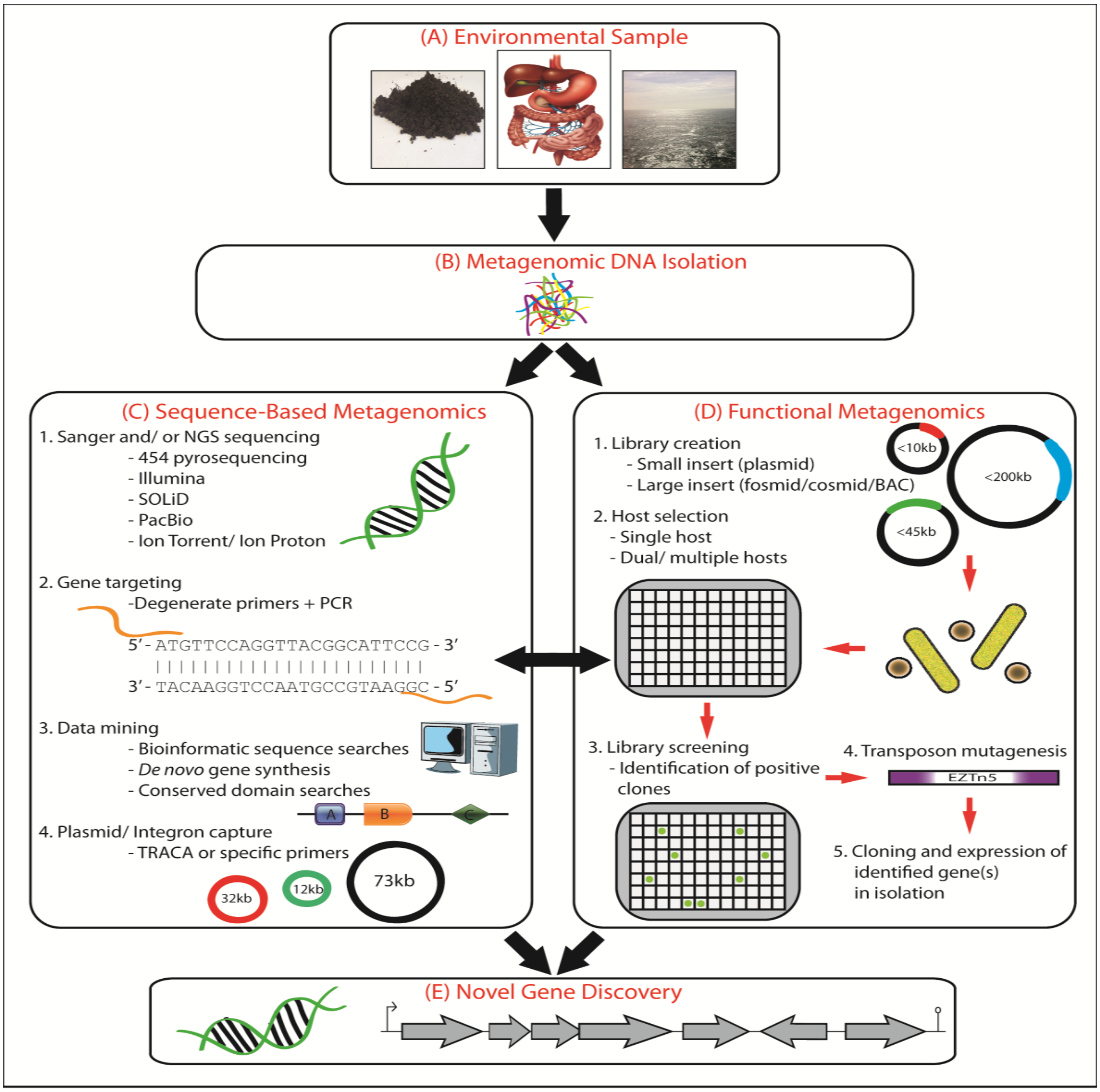

(A) Environmental Sample
(B) Metagenomic DNA Isolation
(C) Sequence-Based Metagenomics
1. Sanger and/ or NGS sequencing
- 454 pyrosequencing
- Illumina
- SOLiD
- PacBio
- Ion Torrent/ Ion Proton
2. Gene targeting
-Degenerate primers + PCR
5'- ATGTTCCAGGTTACGGCATTCCG - 3'
3'- TACAAGGTCCAATGCCGTAAGGC - 5'
3. Data mining
- Bioinformatic sequence searches
- De novo gene synthesis
- Conserved domain searches
A
B
C
4. Plasmid/ Integron capture
- TRACA or specific primers
32kb
12kb
73kb
(D) Functional Metagenomics
1. Library creation
- Small insert (plasmid)
- Large insert (fosmid/cosmid/BAC)
<10kb
<200kb
2. Host selection
- Single host
- Dual/ multiple hosts
<45kb
3. Library screening
- Identification of positive clones
4. Transposon mutagenesis
EZTn5
5. Cloning and expression of identified gene(s) in isolation
(E) Novel Gene Discovery

Figure 1.2. An outline of the primary methods of novel gene discovery using metagenomics (Culligan et al. 2014).

The method of metagenomics is developed upon current developments in the polymerase chain reaction (PCR) amplification, microbial genomics, cloning of genes that share similar sequences (e.g.16S rRNA) directly from the environmental samples (Schloss and Handelsman 2003). This approach of genomics can, in theory, identify genes of any sequence and function from environmental samples (Schloss and Handelsman 2003). The study or analysis of metagenomic DNA is a quite challenging task since the total DNA is a combination of genomes from many diverse organisms from the environment (Belkova et al. 2007). The core of applied metagenomics is to be able to express the discovered genes in a cultivable surrogate host (Bashir et al. 2014). One of the critical goals of metagenomics analysis is to be able to reconstruct genomes of un-culturable organisms via genomic library studies and to be able to ultimately build each chromosome (Schloss and Handelsman 2003). An overview of the metagenomics approach is outlined in **Figure 1.2.**

1.3. Metagenomic methods

In metagenomics, whole genomic DNA can be extracted either directly or indirectly from the environment. Metagenome extraction must be carried out correctly to get access to the entire community within the sample and to ensure that high quantity and quality is obtained (Culligan et al. 2014). In direct DNA extraction, cells are lysed while they are within the matrix of the environmental sample, while indirect DNA extraction, cells are first isolated from the environmental sample followed by lysis and purification (Fan et al. 2012). According to Desai et al. (2008), direct DNA extraction is a widely used method, and it is most suitable for small DNA fragments and construction of large DNA inserts is not much favoured by direct extraction. In direct extraction, DNA is accessed by using mechanical, enzymatic, and chemical lysis (Dias et al. 2014). The primary task in obtaining genetic material from environmental samples involves extracting the intact DNA or RNA in quality and quantity sufficient for downstream analysis, regardless of the relative concentrations of the species present (Dias et al. 2014). Thus, it is vital to maintain microbial diversity, avoid contamination and shearing of DNA and also to consider the origin of the sample and physicochemical conditions (Singh et al. 2013).

In environmental samples, a significant proportion of microorganisms are nonculturable by known culture methods, which leads to inaccurate interpretations of the microbial diversity and gene functions. Thus, the set of DNA extracted by different extraction techniques must be a full representation of the microbiota present in the sample and also generating significant amounts of high-quality genetic material. So the methods used must be carefully chosen as they define the quality of the analysis. Thus, according to Dias et al. (2014) the choice of the optimal method for nucleic acid extracting of a sample relies on the purpose of the analysis, and should take into consideration factors such as the following:

- **Extraction yield**: Minimal amounts of DNA are required for detection of specific DNA fragments. Thus, when extracting DNA, it is crucial to prioritise obtaining substantial concentrations of genetic material, and this should include those from microorganisms that are present in low levels.
- **Maintenance of DNA integrity:** The quality of metagenomic analysis depends on whether the extracted nucleic acid is intact or fragmented during the extraction processes. Efficient DNA extraction methods ensure increased efficiency and reliability of the metagenomic studies, allowing DNA belonging to species present in low density be detected.
- **The purity of the extracted material:** Most downstream applications require that DNA extracted be of high purity. The detection and isolation of prokaryotic genes with specific functions need minute quantities of eukaryotic material contaminants that increase the number of clones to be prepared and screened. According to Gabor et al. (2003) a sample containing only 0.1% of eukaryotic cells, extracted DNA can consist of up to 91% of eukaryotic DNA. Thus, the lesser the concentration of eukaryotic DNA in material obtained, the more efficient is the screening for genes of interest. Also, when dealing with environmental samples especially of soil or sediment origin, the primary limitation in DNA isolation is contamination due to humic acid and phenolic compounds. These compounds affect downstream processes such as restriction digestion, cloning and transformation of the isolated DNA (Liles et al. 2008).

Metagenomic libraries can be generated by cloning DNA directly isolated from an environmental sample into a suitable vector (Carola and Rolf 2011). Different types of vectors are used for metagenomic library construction, and the choice depends on the type and size of the library as well as the type of method screening that will be employed (Handelsman 2004). Selecting an appropriate vector is crucial for the maintenance and expression of the cloned genes (Kakirde et al. 2010). Several vectors used in metagenomic studies include small insert vectors, phage-based vectors, fosmids, cosmids and bacterial artificial chromosomes (BACs). High molecular weight DNA is usually cloned in cosmids (25 to 45 kb), fosmid (15 to 40 kb) and BACs (100-200 kb). If the target is for small genes, the DNA inserts range of between 2 and 10 kb can be constructed into plasmid vectors (Li et al. 2009). Vectors used metagenomic library construction should be compatible with the host microorganism selected for screening as there are already predetermined host for specific vectors. Steps involved in metagenomic library construction are presented in **Figure 1.3** below as adapted from Mirete et al. (2016). The use of heterologous gene expression has some disadvantages. Primarily, for the host to be able to express the genes of the cloned DNA, it must have a compatible expression system. Otherwise, the discovered activities would be low, thus requiring high-throughput screening procedures (Handelsman 2004). Secondly, due to the challenges of extracting high-quality DNA from the environmental sample, sometimes there is the uniform distribution of microorganisms in the sample which affects the full picture of microbial diversity that representative of the sample. Many metagenomic researches employ *Escherichia coli* as a preferred cloning host. The preference is usually because *E. coli* has high transformation efficiency and genetic manipulation as compared to other cloning hosts. It is also deficient in restriction-modification systems and lacks genes of homologous recombination (Casali 2003). The screening of metagenomic libraries for desired biomolecules, novel genes, enzymes and other products involves two metagenomic approaches, function-based screening of expression libraries and sequence-based analysis (Lorenz et al. 2002). The choice of screening method depends on the type of constructed library, functional activity of interest, and the time and resources available to characterise the library.

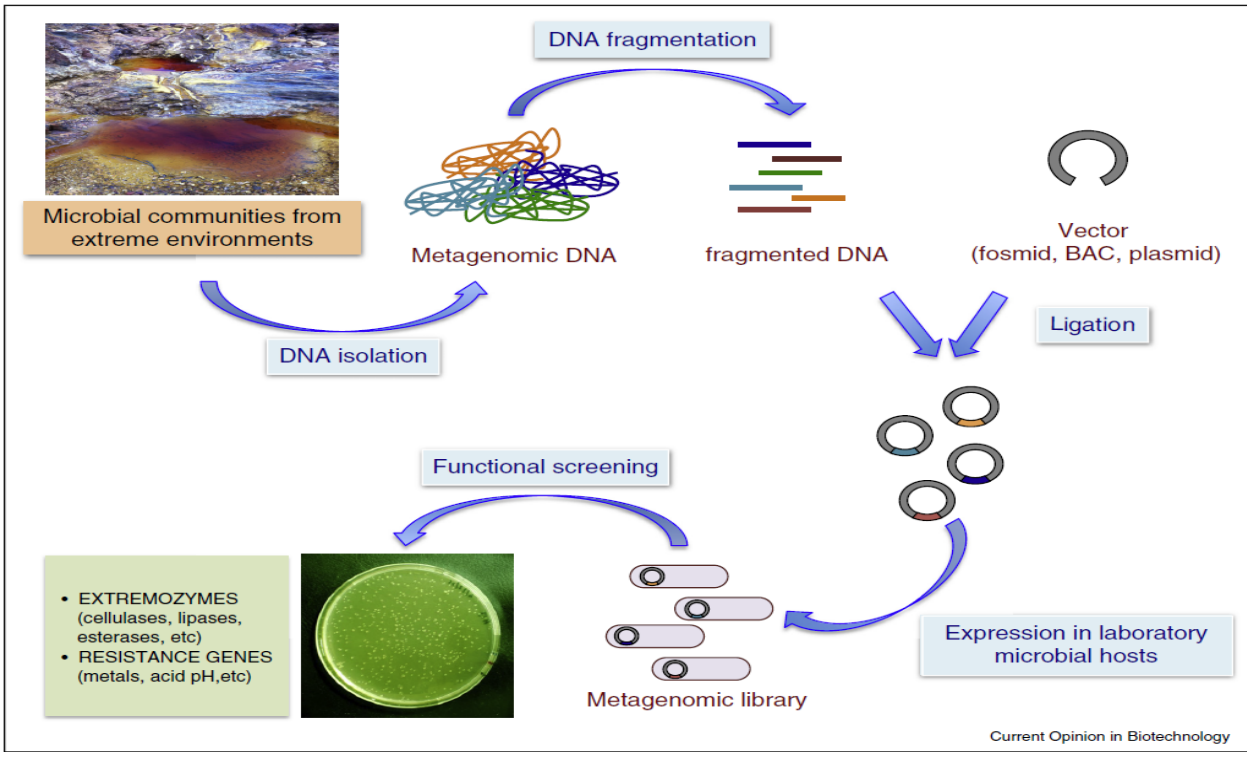

DNA fragmentation
Microbial communities from extreme environments
Metagenomic DNA
fragmented DNA
Vector (fosmid, BAC, plasmid)
Ligation
DNA isolation
Functional screening
• EXTREMOZYMES (cellulases, lipases, esterases, etc)
• RESISTANCE GENES (metals, acid pH,etc)
Metagenomic library
Expression in laboratory microbial hosts
Current Opinion in Biotechnology

Figure 1.3. Schematic diagram outlining the steps involved in the construction of a metagenomic library (Mirete et al. 2016).

1.3.1. Function-based metgenomics

Extracted DNA in function-driven metagenomics is also captured and cloned in a host; then the fragments are screened for a specific function (Handelsman 2009). Genes are accessed by this approach without any prior knowledge of the targeted gene sequence information. This allows the discovery of unknown gene products. The function-driven analysis is started by identifying clones that express specific desired attributes, followed by the classification of the active clones by sequence and biochemical analysis (Schloss and Handelsman 2003). It rapidly identifies clones of importance in agriculture, medicine or industry; focusing on proteins and natural products of useful functions (Schloss and Handelsman 2003). This approach has successfully identified both novel and traditional antibiotics such as lipases, chitinases, antibiotic resistance genes, membrane proteins, enzymes encoding genes and many more. (Entcheva et al. 2001). It is important to realise that in function-driven metagenomics the function of interest must be absent from the surrogate host so that the acquired role after cloning can be solely attributed to metagenomics DNA (Handelsman 2009). Even though function-driven analysis can identify genes by their function rather than a sequence, the draw-back is that most metagenomics DNA cannot be expressed in a surrogate host like *E. coli* (Handelsman 2009). Metagenomic libraries are constructed and subsequently screened for the target enzyme or compound (Li et al. 2009). Activity screening in the function based approach is accomplished by high throughput screening of library clones on indicator media. It also uses mutants of host strains that need heterologous complementation for growth under selective media or growth conditions. The transformed clones that contain the gene of interest will grow under specific growth conditions (Simon and Daniel 2009).

1.3.2. Sequenced-based metagenomics

Sequence-driven analysis, on the other hand, depends upon the employment of conserved DNA sequences to design PCR primers or hybridisation probes to screen for clones with desired sequences from metagenomics library (Schloss and Handelsman 2003). Discoveries of novel natural products and proteins have also resulted from the random sequencing of metagenomic clones (Schloss and Handelsman 2003). The genes are classified according to predicted functions, and types of proteins responsible for respective roles can be evaluated (Belkova et al. 2007). In random sequencing, the DNA is fragmented into pieces of a few thousand bases long, cloned, sequenced and then assembled using computational analysis (Fuhrman 2012). The derived sequences are then compared to other publicly available sequence databases such as GENBANK (Handelsman 2009). The advantage of sequence-based metagenomics is that it is independent of gene expression of the target genes (Lorenz et al. 2002).

The sequence-driven approach is limited to available knowledge of existing sequences; meaning that if the metagenomic gene is not similar to a known function in the databases, not much can be learned from the gene and its products (Handelsman 2009). Limitations of this approach is that acquisition of full-length gene that is required for the production of the desired product is not guaranteed. However, without the assistance of this approach some genes may have not yet been discovered (Tuffin et al. 2009). Therefore, bearing in mind the limitations of both metagenomics approaches, it is essential to learn that these methods are complementary and should be carried out in parallel (Handelsman 2009). The recent developments in advanced sequencing technologies have made access to genetic diversity from environmental samples easy (Kakirde et al. 2010). Conventionally, Sanger sequencing was used the only sequencing technique used to sequence large metagenomes from a complex environment. However, Sanger sequencing methods have limitations, mainly when the environmental sample consists of a complex community which requires more sequencing procedures (Tyson et al. 2004). The limitations of Sanger sequencing and the need to sequence large metagenomes have resulted in next-generation sequencing, which is a high throughput screening method (Fakruddin et al. 2012).

1.3.2.1. Next Generation Sequencing

Traditionally sequence determination is usually carried out using di-deoxy chain technology, which is commonly known as Sanger sequencing (Fakruddin et al. 2012). This method has been widely used for the past 30 years. Next-generation sequencing refers to the non-Sanger based high throughput DNA sequencing technology. Millions or billions of DNA strands can be sequenced in parallel, yielding more throughputs and minimising the need for fragment cloning methods that are often used in Sanger sequencing (Fakruddin et al. 2012). The next generation of metagenomics DNA, which involves no cloning steps is being further developed (Fuhrman 2012). The recent developments of next-generation sequencing technologies allow for quicker metagenomic library construction and sequencing and the result, the sequencing of 16S rRNA genes or other genes can be easily analysed (Bashir et al. 2014).

More recently, a new sequencing method called pyrosequencing has emerged, and many laboratories around the world have made attempts to develop another alternative of DNA sequencing (Fakruddin et al. 2012). Pyrosequencing technology was developed at Royal Institute of Technology (KTH) as the first alternative to Sanger sequencing. This technology depends upon luminometric detection of pyrophosphate released in nucleotide incorporation directed by DNA polymerase (Fakruddin et al. 2012). It is a method of DNA sequencing based on the 'sequencing by synthesis' principle. It is suited for sequencing of up to a hundred bases and has many unique advantages (Fakruddin et al. 2012). Sequencing by pyrosequencing avoids the need for labelled primers and nucleotides as well as gel electrophoresis (Fakruddin et al. 2012).

The low cost of next-generation sequencing has exponentially accelerated the growth of sequence-based metagenomics (Torsten et al. 2012). It is believed that as time progresses the use of metagenomics techniques for sequencing will be used similarly as 16S rRNA gene finger-printing is used currently (Fakruddin et al. 2012). Metagenomic short-gun sequencing speeding has shifted from traditional Sanger sequencing technology to next-generation sequencing (Fakruddin et al. 2012). The next generation sequencing methods that are widely implemented are Illumina system, GS-FLX 454 pyrosequencer, SOLID system, PacBio RS II and Helicos system (Mardis 2008). Further development in such technologies will decrease the limitations associated with sequence-based screening. The choice of specific next-generation sequencing (NGS) platform is made concerning varying features like the length of the read, the degree of automation, quality of data, throughput per run, simplicity in data analysis and the cost per run as compared in **Table 1.2** below.

Table 1.1. Comparison of distinct features of NGS platforms commonly applied in metagenomics research*

Sequencer	**Roche /454 GS FLX Titanium**	**HiSeg 2000**	**SOLiDv4**
NGS chemistry	Pyrosequencing	Sequencing by synthesis	Sequencing by ligation and exact call chemistry
Library/template preparation	Emulsion PCR (emPCR)	Solid phase amplification	Emulsion PCR for fragment/ mate- pair end sequencing
Average read length	230-310bp (highest among the NGS platforms) Now approaching 400-500 (titanium) pyroreads	Initially it was 36, now approaching 150	35
Run time (days)	24 hours (fastest of all)	4 days (fragment run) 9 days (mate pair run)	7 days (fragment run) 14 days (mate pair run)
Output data/run	0.7 Gb	600 Gb (over 1Tb with Illumina's Hiseq X Ten)	120 Gb
Advantage	Longer reads Least time for one run Amendable to multiplexing allowing many samples in single run	High throughput Most widely used platform	High accuracy due to ECC (exact call chemistry)
Limitations	High error rate in homopolymer region High cost of reagents Low in throughput Artificial replicate sequences during emPCR	Short read length Low multiplexing capability of samples Single base error with GGC motifs High error rate at tail end reads	Long run time Short read length

* Source: Kumar et al. (2015).

1.4. DNA Manipulating enzymes

1.4.1. DNA ligase

DNA ligase is one of the essential discoveries in molecular biology and biotechnology due to the role it plays in the molecular cloning of important genes (Al- Manasra and Al- Razen 2012). It was first discovered in 1967 and 1968 by the work of several laboratories (Al- Manasra and Al- Razen 2012). Polynucleotide ligases role is to join and seal the breaks by a phosphodiester bond between 5’ PO_4 and 3’ OH ends in DNA and RNA molecules; using a multistep ligation reaction that involves the use of an AMP molecule to be covalently joined to both ligase and polynucleotide substrate, respectively (Pascal 2008). This process allows the joining of similar and foreign DNA sequences. DNA ligase is ubiquitous in almost all living organisms, and it is required for survival functions and maintaining the integrity of the DNA backbone structure (Al- Manasra and Al- Razen 2012). They are housekeeping enzymes that are essential for survival roles and cellular process linked to breaks filling in the nucleic acid backbone structure by joining the 3' hydroxyl, and 5' phosphoryl group ends and forming phosphodiester bonds. They have essential roles in DNA replication, repair and recombination (Rossi et al. 1997).

Polynucleotide ligases are divided or classified into two types according to what is the source of their cofactor; they are either ATP- or NAD+-dependent enzymes (Pascal 2008). Archaea, viruses and eukarya utilise ATP as a cofactor for DNA ligase, while eubacteria depend on NAD+ to perform ligation mechanism (Wilkinson et al. 2001). ATP and NAD+ -dependent DNA ligases rely on different accessory domains to carry out the formation of the first step of the ligation reaction (ligase-AMP) intermediate (Pascal 2008). ATP-dependent DNA ligases employ Oligomer-Binding (OB) domain that is situated near the C- terminus of NTase, while NAD+-dependent DNA ligases utilise *Ia* domain which is an N-terminal extension of the NTase to execute this function. ATP- and NAD+- dependent DNA ligases are illustrated in **Figure 1.4** (**A**). The ATP-dependent DNA ligase from bacteriophage T7 is a two-domain ligase: the nucleotide-binding domain (green) binds ATP and is connected to an OB-fold domain (yellow) by a flexible linker. (**B**) The NAD+-dependent DNA ligase from *Thermus filiformis* is a multidomain ligase. The basic folds for the nucleotide-binding domain and the OB-fold domain are similar to that found in T7 DNA ligase. Additionally, a zinc finger domain, a helix-hairpin-helix domain, and a BRCT domain extending from the C-terminus of the OB-fold domain. Domain Ia (grey), which helps in the step 1 reaction, is N-terminal to the nucleotide-binding domain and is exclusive to the bacterial ligases (Shuman 2009).

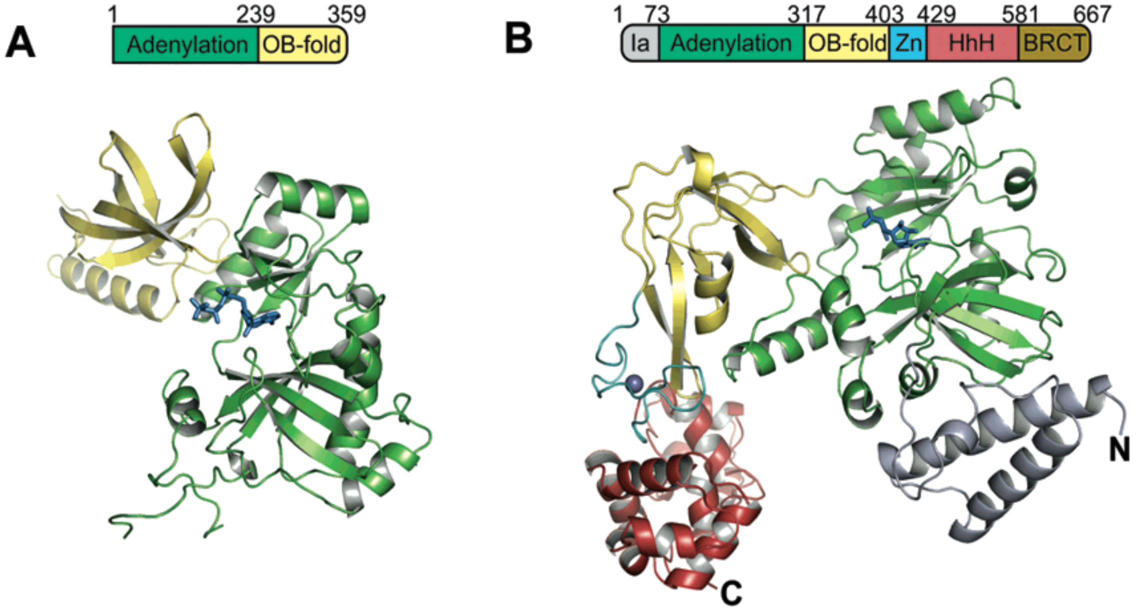


Figure 1.4. Structural differences in ATP- and NAD+-dependent DNA ligases (Shuman 2009).

Although both polynucleotide RNA and DNA ligases have many differences in their structural domains, they usually utilise multi-domain construction to carry out the multi 2 step ligation reaction mechanism. All polynucleotide ligases have a standard feature and critical building block in their structural construction which is the nucleotidyltransferase domain (NTase). This domain is located in a specific manner with N- and C-terminal appendages to maintain the overall ligation reaction and to provide substrate specificity according to unique DNA/RNA binding properties (Pascal 2008).

The DNA ligase from bacteriophage T4 is a commonly used enzyme in the molecular biology research laboratory. It can ligate either cohesive or blunt ends of DNA, oligonucleotides, as well as RNA and RNA- DNA hybrids, but not single- stranded nucleic acids. It can also ligate blunt- ended DNA with much higher efficiency than E.coli DNA ligase. Unlike E. coli DNA ligase, T4 DNA ligase cannot utilize NAD+, and it has an absolute requirement for ATP as a cofactor (Pascal 2008).

More lately, a method for the isothermal assembly of substantial DNA fragments commonly known as Gibson assembly (Gibson et al. 2009), which employs thermostable *Taq* DNA ligase, has been described and broadly accepted. DNA ligases can also be used in projects involving gene synthesis (Bang & Church 2008). They are essential in most next-generation sequencing (NGS) platforms, either during sample preparation or for adapter ligation (e.g. Ion Torrent sequencing), or for the sequencing reaction itself (e.g. SOLiD sequencing). The most commonly utilised DNA ligase in these projects is the ATP-dependent enzyme from bacteriophage T4, which is also one of the first to be discovered (Wilson et al. 2013). Consistent with its physiological role, in vitro it is highly effective at sealing single-stranded nicks in duplex DNA (Wilson et al. 2013).

While T4 DNA ligase has evolved to be a nick-sealing enzyme, it can also join double-stranded DNA fragments that have complementary, overhanging, single-stranded ends (**Figure 1.5**). Moreover, it is the only commercially available ligase that can join blunt-ended DNA duplexes in the absence of macromolecular enhancers such as polyethylene glycol (Miller et al. 2003). It is the ligation of cohesive or blunt-ended dsDNA fragments that are most commonly needed in molecular biology protocols. It is also inefficient at ligating pieces with single base overhangs (Lohman et al. 2011). For molecular biologists, the poor kinetic parameters of T4 DNA ligase typically manifest as failed cloning experiments, or sub-optimal libraries for Illumina and 454 sequencing runs (Wilson et al. 2013).

1.4.1.1. Ligation mechanism

Ligation mechanism of DNA ligase is a three steps reaction. During the initial step of the three ligation reaction, a phosphoamide bond (P-N) forms between the R-amino group of an active site lysine and the 5' phosphate of AMP for DNA and RNA ligases, or GMP for mRNA capping enzymes (Wilkinson et al. 2001). The activated enzyme-NMP adduct is produced in the absence of a nucleic acid substrate (Pascal 2008). The ligation reaction is an energy-dependent process and involves three successive steps, which comprises two covalent reaction intermediates (**Figure 1.6)**. In the first step; ligase is activated by covalent attachment between α-phosphate of AMP molecule and the enzyme forming a ligase–AMP intermediate and releasing inorganic pyrophosphate, whereas nicotinamide mononucleotide (NMN) is released by NAD+-dependent ligases. In the subsequent step; the AMP group is transferred from ligase to the 5' end phosphate group of the DNA molecule forming an AMP-DNA intermediate. In the third step; the hydroxyl group on the 3' end of the break in the substrate attacks the phosphate on the 5' end of the opposing nucleic acid strand creating uninterrupted DNA molecule backbone structure and releasing a free AMP and covalently join the DNA strands (Wilkinson et al. 2001; Pascal 2008).

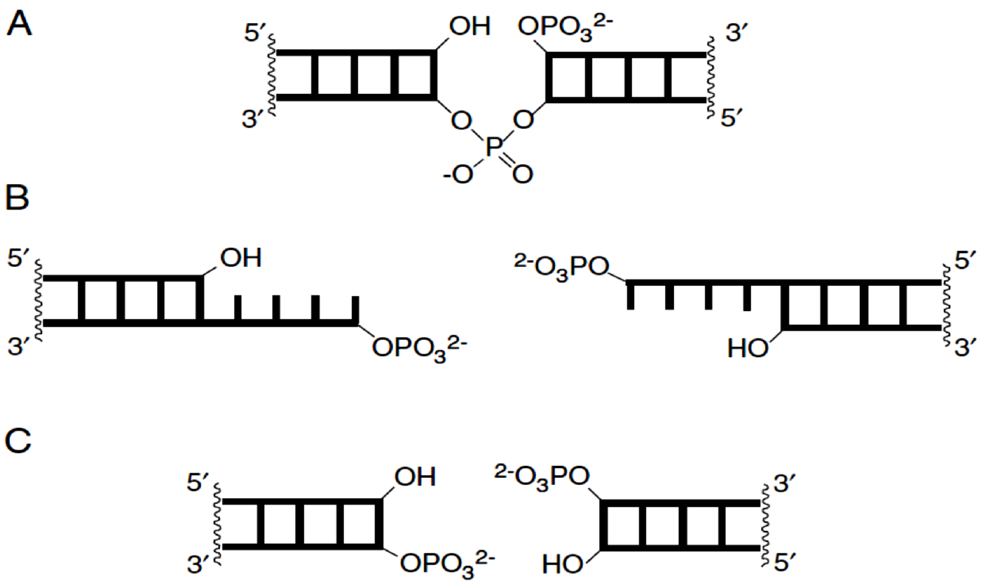


Figure 1.5. Ligase substrates; (A) Nicked dsDNA, (B) cohesive ends, and (C) Blunt ends (Lohman et al. 2011).

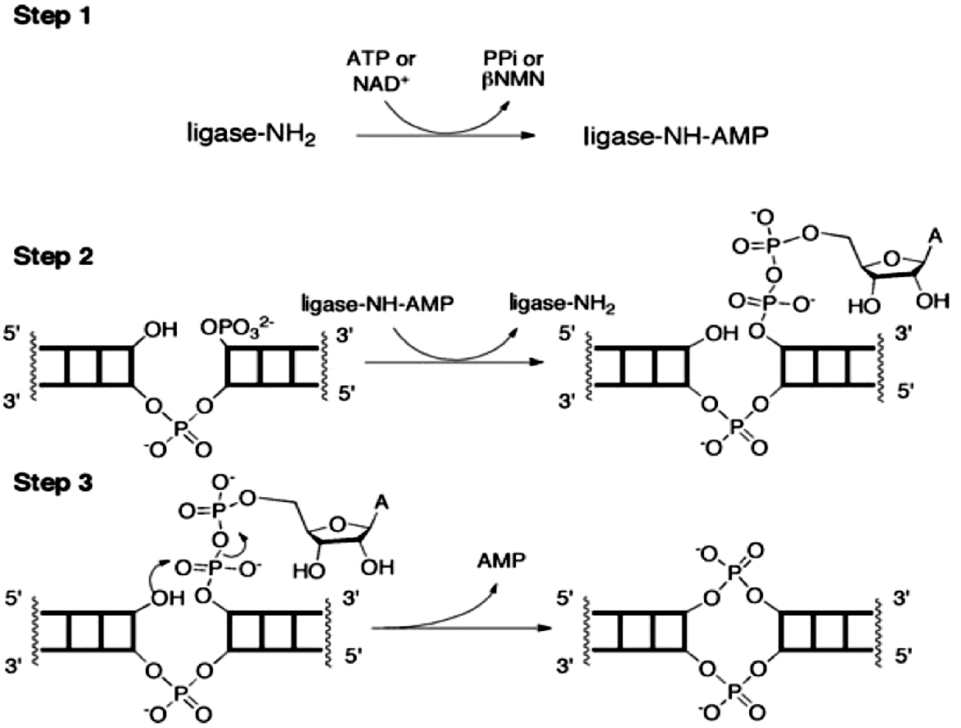


Figure 1.6. The steps involved in the DNA ligation mechanism (Lohman et al. 2011).

The ligation reaction is known to proceed in three steps: (1) reaction of the enzyme with a cofactor to form an enzyme-AMP covalent intermediate; (2) transfer of the AMP to a 5' phosphorylated DNA terminus; and (3) joining of a 3' hydroxyl DNA terminus with the adenylated 5' strand accompanied by the release of AMP (Shuman 2009).

1.4.2. DNA polymerase

DNA polymerases are of key importance in DNA replication and repair. DNA polymerases are enzymes that copy or make DNA molecules from deoxyribonucleotides (dNTPs), the building blocks of DNA. These enzymes are crucial for DNA replication and they usually function in pairs to synthesise two identical DNA strands from a single DNA molecule (Garcia- Diaz and Bebenek 2007). During DNA synthesis, DNA polymerase reads the existing DNA strands to form two new strands that match the existing original ones. DNA polymerase Pol I from *E. coli* was first discovered in 1958 by A. Kornberg and colleagues (Lehman et al. 1958; Rothwell and Waksman 2005; Garcia- Diaz and Bebenek 2007). The discovery of many other polymerase enzymes soon followed, and it was understood that they had significantly different characteristics. However, it was not until recently when sequence information became available through sequencing that the details behind those biochemical differences could be understood

(Rothwell and Waksman 2005). It then became clearer that polymerases, although evolutionarily related, were, in fact, divergent, and the difference in features of their primary sequence brought about their classification into different families that are still current (families A, B, C, X and Y). For example, DNA polymerase A belongs to the A family and has three different domains: 5'- 3' exonuclease domain on the N- terminus, a central proofreading 3'- 5' exonuclease and also a polymerase domain at C terminus of the enzyme (Simon et al. 2009).

DNA polymerases have revolutionised molecular biology with their ability to amplify small amounts of DNA *in vitro* (Kaguni 2018). Over the last 20 years their use in the Polymerase chain reaction (PCR) has overcome a major limiting actor in daily Medicine i.e. the quantitative problem of the small amounts of DNA available for testing. These small amounts of DNA can be a single gene, or just part thereof (Drouin 2007). According to "future market insights" website, the global DNA polymerase market value is projected to grow at 6.5% during the ten years period (2017-2027) and reach a market value of US$ 389.4 Million by the end of 2027. The main factors contributing to the increase in demand for DNA polymerase include high spending in life sciences R&D laboratories, molecular diagnostics tools for diseases diagnosis, development of novel point-of-care, as well as the high demands in epigenetics research. Additionally, factors such as increasing use of high-fidelity DNA polymerases for crude samples, as well as an enhanced sales network and distribution agreements by various vendors are driving revenue growth of the global DNA polymerase market. However, multiple factors such as limited reproducibility of research studies, availability of alternatives and a high cost of products, especially those used in molecular diagnostics are expected to limit market growth; hence a need to further produce low cost and high fidelity DNA polymerases (https://www.futuremarketinsights.com/reports/dna-polymerase-market).

1.4.2.1. DNA polymerase families

The recent development in high throughput sequencing projects brought about a revolution in the polymerase studies (Garcia- Diaz and Bebenek 2007). Within a short space of time, numerous novel DNA polymerase genes were identified (Goodman and Tippin 2000). According to Ohmori et al. (2001), one of the first breakthroughs was the identification of a novel family of DNA polymerases, family Y, which is widely believed to carry out the synthesis of opposite template lesions in a process known as translesion synthesis (Prakash et al. 2005). Thus DNA polymerases are generally categorised into five different groups or families, i.e. A, B, C, X and Y. Structures of polymerases families are shown in **Figure 1.7.** The proteins are represented as ribbon configurations. The fingers (coloured gold), palm (coloured red), and thumb subdomains (green). **Figure 1.7** (**A**) is a structure of apo Klentaq1 of family A. The 3'-5' vestigial exonuclease domain is indicated in silver, (**B**) is a structure of apo RB69 DNA polymerase of family B. The 3'-5' exonuclease domain and the N-terminal domain are illustrated in grey and silver, respectively. **Figure 1.7** (**C**) is the structure of apo pol b DNA polymerase of family X. The lyase domain is indicated grey, (**D**) is a structure of the Dpo4 DNA polymerase of family Y. The little finger subdomain is shown in silver. Finally, designated as (**E**) is the structure of the p66 subunit of reverse transcriptase (RT family). The RNAseH is indicated in grey whereas subdomains as silver respectively (Prakash et al. 2005).

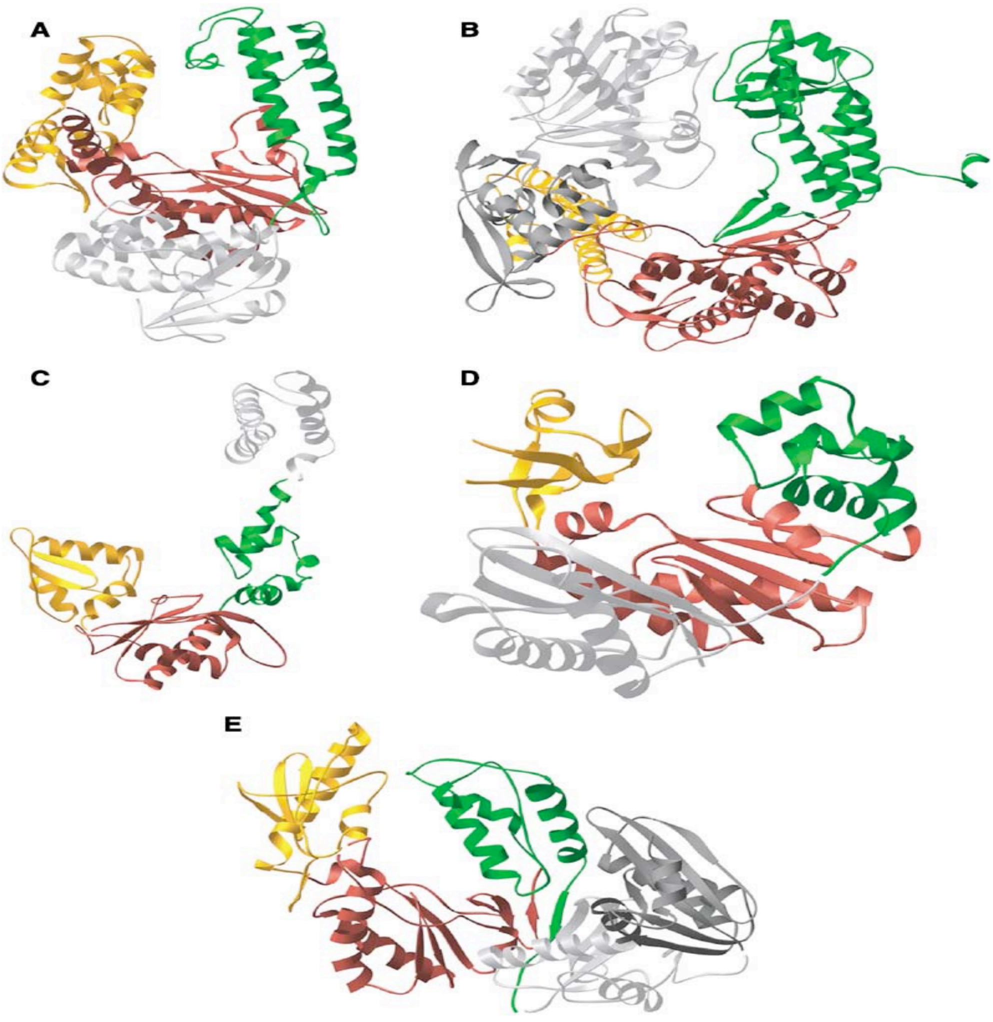


Figure 1.7. Structural differences between family A, B, X, Y, and RT polymerases (Garcia-Diaz and Bebenek 2007).

1.4.2.1.1. Family A

One of the members of this family is prokaryotic DNA polymerase I (Pol I) from *E. coli* was discovered about 50 years ago (Lehman et al.1958; Garcia- Diaz and Bebenek 2007). It is the first DNA polymerase to be isolated, whose structure was thoroughly studied and determined (Garcia- Diaz and Bebenek 2007). It was previously thought to be the main replicative polymerase in bacteria, but it was later understood to play a crucial role in DNA repair and maturation of Okazaki fragments (Garcia- Diaz and Bebenek 2007). Besides DNA polymerisation, *E. coli* Pol I possesses two additional activities; a 3'-5' and a 5'-3' exonuclease. Of these two, the 3'-5' exonuclease activity is common in quite a few other members of the family. This exonuclease property is referred to as a proofreading activity because it can cut nucleotides that are incorrectly inserted by the polymerase (Rothwell and Waksman 2005). DNA polymerase I is common amongst prokaryotes. It is the most abundant polymerase in *E. coli* and accounts for more than 95% of polymerase activity. However, it has been discovered that cells that lack pol I activity can be substituted either pol I, Pol II, Pol III, Pol IV or pol V. During replication pol I add ~15 to 20 nucleotides per second, suggestive of poor processivity. To circumvent this, pol I add nucleotides at the origin of replication (Ori) and 400bp downstream, another polymerase (pol III) takes over replication at a much higher speed. Though the bacterial polymerase of this category only plays a small role in replication, members of this family from other organisms do perform some genomic replication (Garcia-Diaz and Bebenek 2007).

1.4.2.1.2. Family B

Like most family A enzyme, most family B enzymes contain an associated 3′-5′ exonuclease activity (Banach-Orlowska et al. 2005). However, unlike members of other families, family B polymerases are multisubunit enzymes. Prokaryotic DNA polymerase II is a product of the pol B gene, which plays the vital role in DNA repair as well as in replication restart to avoid lesions. It is believed to direct the activity polymerase at the replication fork and also offers support to stick a Pol III avert terminal mismatches. Its presence in the cell can range between ~30- 50 copies per cell in un-induced cells and about ~200- 300 copies in SOS induction (Banach-Orlowska et al. 2005).

1.4.2.1.3. Family C

Polymerases belonging to this family are the main replicative polymerases in bacteria. DNA polymerase III also known as holoenzyme is the primary family C polymerases playing a pivotal role in DNA replication. It is made up of the clamp-loading complex, the pol III core and, the beta sliding clamp processivity factor (Garcia- Diaz and Bebenek 2007). The core includes three subunits; α subunit which is known to be the polymerase activity centre, the δ subunit which functions as the exonucleolytic proofreader, and the θ subunit which may stabilise δ. The core and the beta sliding clamp exist in duplicate, to enable the processing of both the leading and lagging DNA strands (Banach-Orlowska et al. 2005).

1.4.2.1.4. Family X

These are small, monomeric polymerases that play a role in short gaps filling during DNA repair (Ramadan et al. 2004). A common characteristic of most members in this group is the presence of an N-terminal 8 kDa DNA binding domain, which aids binding to gapped substrates (Beard and Wilson 2006). They exist in different organisms, from individual viruses and bacteria to yeast and mammals (Garcia-Diaz et al. 2005). This ubiquity would seem to be related to their ability to carry out gap-filling. Under this category, the most researched of these enzymes is Pol β that is involved in repair of base damage through the BER process (Wilson et al. 2000). Other family X polymerases include three enzymes that participate in the V (D) J recombination process: Pol λ, Pol μ (Garcia- Diaz and Bebenek 2007) and the template-independent terminal deoxynucleotidyl transferase (Bertocci et al. 2006).

1.4.2.1.5. Family Y

Polymerases from this category have numerous common features or characteristics. They do not possess an exonuclease activity, and they have a domain referred to as PAD; wrist or little-fingers domain (Ling et al. 2001; Garcia- Diaz and Bebenek 2007) appears to control substrate specificity. They have a high specificity of synthesis on damaged DNA (Kunkel 2004). According to Ling et al. (2001) members of this family, unlike other families, possess a loose DNA binding pocket for the nascent base pair. These family Y enzymes can accommodate damaged or distorted DNA structures in their active site, making polymerisation of damaged DNA possible. In fact, the primary role of these enzymes appears to be in DNA lesion tolerance pathways if the cell fails to repair DNA lesions that can interfere with the replication process. And these lesions are encountered by the replication fork; family Y polymerases can bypass those lesions by polymerising across the damaged site, in a process that has been termed translesion synthesis (Prakash et al. 2005). In *E. coli*, DNA polymerase IV is one such an example of family Y enzymes. It is an error-prone DNA polymerase which participates in non-targeted mutagenesis. Pol IV is a Family Y polymerase expressed by the dinB gene that is switched on via SOS induction caused by stalled polymerases at the replication fork. During SOS induction, Pol IV production is increased tenfold and one of the functions during this time is to interfere with Pol III holoenzyme processivity (Garcia- Diaz and Bebenek 2007). This creates a checkpoint, stops replication, and allows time to repair DNA lesions via the appropriate repair pathway (Jarosz et al. 2007). Another function of Pol IV is to perform translesion synthesis at the stalled replication fork. Cells lacking dinB gene have a higher rate of mutagenesis caused by DNA damaging agents (Nakamura et al. 2012).

DNA polymerase V is also a family Y DNA polymerase that plays a crucial role in SOS response and translesion synthesis DNA repair mechanisms (Jarosz et al. 2007). Transcription of Pol V through the umuDC genes is highly controlled to produce only Pol V when impaired DNA is present in the cell generating an SOS response. Stalled polymerases cause RecA to bind to the ssDNA, which induces the LexA protein to self-digest (Raychaudhury and Basu 2011). The same RecA-ssDNA nucleoprotein post-translationally alters the UmuD protein to UmuD' protein. The UmuD and UmuD' form a heterodimer that interacts with UmuC, which in turn triggers umuC's polymerase catalytic activity on damaged DNA (Raychaudhury and Basu 2011). Pol IV catalyses both insertion and extension with high activity, whereas Pol V is considered the primary SOS TLS polymerase. One example is the bypass of intrastrand guanine thymine cross-link where it is shown by the variation in the mutational signatures of the two polymerases, that is, Pol IV and V contend for TLS in the intra-strand crosslink (Raychaudhury and Basu 2011).

1.4.3. Restriction enzymes

The term restriction enzyme was first used in the studies of phage lambda, a virus that infects bacteria and it refers to the phenomenon of host- controlled restriction and modification of such bacteriophage (Winnacker 1987). Salvador Luria and Giuseppe Bertani were the first people to identify the phenomenon in the early 1950s (Bertani and Weigle 1953). It was discovered that, for a bacteriophage λ that can grow well in one strain of *Escherichia coli* K, when it is grown in another strain *E. coli*, its yields dropped remarkably, by as much as 3-5 orders of magnitude. The host bacteria, in this example *E. coli* K, is referred to as the restricting host and it can reduce the biological activity of the phage λ. If a phage infects one strain, the ability of that phage to proliferate also becomes limited in other strains. In research done in the laboratories of Werner Arber and Matthew Meselson in the 1960s, it was shown that the restriction is caused by an enzymatic manipulation of the phage DNA, and the enzyme involved was therefore termed a restriction enzyme (Dussoix 1962; Arber and Linn 1967; Meselson and Yuan 1968). The restriction enzymes identified by Arber and Meselson were type I restriction enzymes, which digest DNA randomly away from the recognition site (Arber and Linn 1967; Meselson and Yuan 1968).

In 1970, Hamilton O. Smith and colleagues isolated and characterised the first type II restriction enzyme, *Hin*dII, from the bacterium *Haemophilus influenzae* (Smith and Wilcox 1970). Type II restriction enzymes are more useful for laboratory work as they cut DNA at the position of their recognition sequence. Subsequently, in work done by Daniel Nathans and Kathleen Danna, it was showed that cleavage of simian virus 40 DNA by restriction enzymes produces specific fragments that can be separated using polyacrylamide gel electrophoresis, thus revealing that restriction enzymes can also be used for mapping DNA. Therefore, restriction enzymes allow DNA to be manipulated, leading to the development of recombinant DNA technology which entails joining together of DNA molecules from various species that are introduced into a host organism to synthesise new genetic combinations that are of significance to science, agriculture, medicine, and industry (Luria and Human 1952; Villa-Komaroff et al. 1978).

Restriction enzymes are found in archaea and bacteria, and they are used by these organisms as a defence mechanism against viruses. These enzymes excise foreign DNA in the process called restriction; meanwhile, host DNA is protected by a modification enzyme (a methyltransferase) that modifies the prokaryotic DNA and blocks cleavage. Collectively, these two processes make up the restriction modification system (Kobayashi 2001). To date, more than 3000 restriction endonucleases have been identified and studied in detail. Over 600 of these enzymes are commercially available and are commonly used for DNA modifications in molecular biology laboratories as vital tools in molecular cloning (Roberts 1976).

According to Kessler and Manta (1990), the function of restriction enzymes is to recognise a particular sequence of nucleotides to create a double-stranded cut in the DNA. The number of bases found in recognition site, usually 4- 8 bases is typically used to classify recognition sequences. Furthermore, the number of bases in the DNA sequence determines how many times the recognition site will appear by chance in any given genome. Many of the recognition sequences are palindromic, which means the sequence of bases reads the same backwards and forwards (Pingoud and Jeltsch 2001). Theoretically, two types of palindromic sequences are possible in DNA, i.e. Mirror- like palindrome and inverted repeat palindrome. In mirror- like a palindrome, the sequence reads the same forward and backwards on one strand of DNA, whereas in an inverted repeat palindrome, the sequence reads the same forward and backwards, but these sequences are found in complementary strands of a double-stranded DNA. Among two palindromes, inverted repeat palindromes are the most common, and they have greater biological importance (Pingoud and Jeltsch 2001).

Restriction enzymes are commonly categorized into four classes (Types I, II III, and IV); according to their structure or whether they cleave their DNA substrate at their recognition site or if cleavage position and recognition site are separate from one another, or they cut once through each strand of a DNA double helix (Yuan 1981; Bickle and Kruger 1993). However, DNA sequence analyses of restriction enzymes show numerous variations, indicating that there are more than four types.

1.4.3.1. Classes of restriction enzymes

All types of restriction enzymes recognise specific short DNA sequences and carry out the endonucleolytic cutting of DNA to give particular fragments with terminal 5'-phosphates. Below is a summary of different classes of restriction endonucleases as outlined in Williams (2003) and Sistla and Rao (2004).

- Type I enzymes cut at sites remote from a recognition site, and they require both ATP and S-adenosyl-L-methionine to carry out its function. They are a multifunctional protein with both restriction and methylase activities.
- Type II enzyme cut within or at short, specific distances from a recognition site, with most of which require magnesium. They only have restriction modification function, independent of methylase.
- Type III enzymes cleave at sites a few bases from a recognition site. They require ATP, and the presence of S-adenosyl-L-methionine stimulates the reaction but is not required. In nature, they exist as part of a complex with a modification methylase.
- Type IV enzymes target modified DNA, e.g. methylated, hydroxymethylated and glucosyl-hydroxymethylated DNA.

1.4.3.1.1. Type l Restriction endonucleases

Type I restriction endonucleases were the first among restriction enzymes to be identified and were isolated from two different strains of *E. coli*, i.e. K-12 and B. There are random cutters, and they cleave at 1000bp away from their restriction site (Murray 2000). The process of DNA translocation directs cleavage at these random these random sites, making type I restriction enzymes molecular motors. The recognition site is asymmetrical and is made of two portions; the first one contains 3-4 nucleotides, while the second one comprises 4-5 nucleotides. They possess both restriction and modification activities subject to the methylation status of the targeted DNA. To achieve their full activity, they require the presence of factors S-Adenosylmethionine (AdoMet), hydrolysed adenosine triphosphate (ATP) as well as magnesium ions (Murray 2000). They possess three subunits namely; HsdR, HsdM and HsdS. HsdR subunit is needed for restriction activity while HsdM is essential for the addition of methyl groups to host DNA during methyltransferase activity and HsdS is vital for recognition site specificity in addition to both restriction and modification activity (Bickle and Kruger 1993; Murray 2000).

1.4.3.1.2. Type II Restriction endonucleases

These restriction enzymes are different from the type I restriction enzymes in many different ways. Firstly, they form homodimers, and they possess recognition sites that are uninterrupted, palindromic and are 4- 8 nucleotides in length. Their recognition site and their cleavage sites are the same, they do not require ATP or AdoMet for their activity, and they usually need magnesium (Mg^{2+}) ions as a cofactor (Pingould and Jeltsch 2001). They cut the phosphodiester linkages of a double helix DNA. To achieve blunt ends, it cleaves the centre of a strand, or it can cut DNA at a staged position to yield the overhangs called sticky ends (Ninfa et al. 2010). They are the most utilized and commercialized restriction enzymes. Between the 1990s and early 2000s new type II enzymes were identified that did not follow all the typical characteristics of this enzyme class, and a new naming system was introduced to categorise this large family into subfamilies based on known characteristics of type II enzymes (Pingoud and Jeltsch 2001). A summary of different subfamilies of type II enzymes, their recognition sequence and examples is given in **Table 1.3**.

A letter suffix was introduced to identify subgroups of type II restriction enzymes as outlined below.

- Type IIB restriction endonucleases are multimers, possessing more than one subunit (Pingoud and Jeltsch 2001).They cut DNA on both sides of their recognition to cut out the recognition site. They need both Mg^{2+}cofactors and AdoMet.
- Type IIE restriction endonucleases cut DNA following interaction with two copies of their recognition sequence (Pingoud and Jeltsch 2001). One recognition site serves as the target for cleavage, while the other site acts as an allosteric effector that improves the effectiveness of enzyme cleavage.
- Like type IIE enzymes, type IIF restriction enzymes interact with two copies of their recognition sequence but cut both sequences at the same time (Pingoud and Jeltsch 2001).

- Type IIG restriction endonucleases only have a single subunit, like most Type II restriction enzymes, but it is activated by the presence of AdoMet cofactor (Pingoud and Jeltsch 2001).
- Type IIM restriction endonucleases, e.g. *Dpn*I, can recognise and cleave methylated DNA (Pingoud and Jeltsch 2001).
- Type IIS restriction endonucleases cut DNA at a specific distance from their non-palindromic asymmetric recognition sites. This characteristic feature is mostly used to carry out in-vitro cloning techniques. These enzymes may also function as dimers.
- Lastly, type IIT restriction enzymes possess two different subunits. Some recognise palindromic sequences while others have asymmetric recognition sites (Pingoud and Jeltsch 2001).

1.4.3.1.2.1. Type III Restriction endonucleases

Type III enzymes are the beta-subfamily of N6 adenine methyltransferases, comprising the nine motifs that characterise this family, including motif I, the AdoMet binding pocket (FXGXG), and motif IV, the catalytic region (S/D/N (PP) Y/F) (Bourniquel and Bickle 2002; Sistla and Rao 2004). These enzymes recognise two distinct sequences that are inversely oriented and non- palindromic. They cleave DNA between 20- 30 base pairs after the recognition site (Dryden et al. 2001). They possess more than one subunit, and they require both cofactors AdMet and ATP for DNA methylation and restriction, respectively. They form part of prokaryotic DNA restriction and modification mechanism that defends the organism from invading foreign DNA. Type III enzymes are multifunctional proteins that are made up of two subunits, Res and Mod. The Mod subunit plays a role in recognising the DNA sequence that is specific for the modification methyltransferase, which by function, it is equivalent to the S and M subunits of type I restriction enzymes. The Res subunit is essential for restriction, although it is devoid of enzymatic activity of its own. Type III enzymes function by methylating only one DNA strand, at position N- 6 of adenosyl residues, meaning that the newly replicated DNA will have only one strand methylated, which is enough to protect the organism's DNA against restriction from foreign invading DNA (Dryden et al. 2001).

Table 1.2. The different subfamilies type II enzymes, their recognition sequence and examples*

Subtype	Characteristic feature	Example	Recognition sequence
Orthodox	Palindromic recognition site, which is recognized by a homodimeric enzyme, cleavage occurs within or adjacent to recognition site.	*Eco*RI	G |AATTG CTTAA|C
		*Eco*RV	GAT|ATC CTA |TAG
		*Bgl*I	GCCNNNN|NGGC CGGN|NNNCCG
Type IIS	Asymmetric recognition site with cleavage occurring at defined distance.	*Fok*I	GGATGN,|NNNN CCTACN,NNNN|
Type IIE	Two sites required for cleavage, one serving as allosteric effector.	*Nae*I	GCG|CGC CGC|GCG
Type IIF	Two sites required for cleavage, both sites are cleaved in a concerted reaction by homotetrameric enzyme.	*Ngo*MIV	G|CCGGC CGGCC|G
Type IIT	Different subunits with restriction and modification activity.	*Bpu* 101	CC|TNAGC GGANT|CG
		*Bsl*I	CCNNNNN|NNGG GGNN|NNNNNCC
Type IIG	One polypeptide chain with restriction and modification activity.	*Eco* 571	CTGAAGN$_{14}$NN| GACTTCN$_{14}$|NN
Type IIB	Cleavage on both sides of the recognition site.	*Bcg*I	NN|N$_{10}$CGAN$_{6}$TGCN$_{10}$NN| NNN$_{10}$GCTN$_{6}$ACGN$_{10}$|NN
		*Bpd*I	NN$_{4}$|N$_{8}$GAGN$_{5}$CTCN$_{8}$N $_{4}$N| |NN$_{4}$N$_{8}$CTCN$_{5}$GAGN$_{8}$ |N$_{4}$N
Type IIM	Methylated recognition site.	*Dpn*I	G ^{m}A|TC C T|$^{\mu}$AG

* Source: Pingoud and Jeltsch (2001).

1.4.3.1.2.2. Type IV and V Restriction endonuclease

Type IV enzymes typically methylate DNA and are exemplified by the McrBC and Mrr systems of *E. coli* (Barrangou et al. 2007). Type V restriction enzymes, e.g. the cas9-gRNA complex from CRISPRs employs guide RNAs to target specific non-palindromic sequences found on invading organisms (Barrangou et al. 2007). They can cut DNA of variable length, provided that a suitable guide RNA is present. The simplicity and flexibility of use of type V restriction enzymes make them promising for future genetic engineering applications (Horvath and Barrangou 2010).

1.4.3.2. Application of Restriction endonucleases

They are used in molecular biology laboratory to aid in insertion of genes into a plasmid vectors in the process of gene cloning and for protein production projects (Zhang et al. 2005). To afford the plasmids optimal use, they are usually modified to possess a short polylinker sequence called multiple cloning sites (MCS); abundant in sequences for restriction enzymes recognition (Russel 2001). This enables flexible insertion of gene fragments into the vector of choice. It is important to note that naturally available restriction sites within the genes to be inserted influence the selection of restriction enzymes for digesting DNA since it is essential not to cut wanted DNA (Zhang et al. 2005). To carry out gene cloning both gene fragment and plasmid must be cleaved with the same restriction enzymes, and then joined together with the help of DNA ligase (Russel 2001).

They are also used contrast between gene alleles by recognizing single base changes referred to as single nucleotide polymorphism (SNPs) (Russel 2001; Zhang et al. 2005). This application is however only possible if the SNPs modifies the restriction site available in the allele. In this technique, restriction endonuclease is used for genotyping a DNA sample without the requirement of expensive gene sequencing (Russel 2001). The method is carried out as follows; DNA sample is first digested with the restriction enzyme to produce DNA fragments, and these fragments are subjected to gel electrophoresis to separate the pieces according to size. In principle, the only alleles that will provide two visible bands of DNA are those with the correct restriction site, while those with modified restriction sites will not be cut and will only produce one band. Moreover, DNA map by restriction digest can be created that reveal relative positions of the genes (Zhang et al. 2005). Restriction digest can also generate a particular pattern of bands after gel electrophoresis that can be used for DNA fingerprinting. These enzymes are particularly very important in molecular laboratories, thus they do offer commercial benefits.

1.5. Cloning, expression and purification of recombinant protein

The gene of interest is usually cloned, and the protein is amplified into suitable expression host systems such as bacteria, yeasts, insects and plants (Demain and Vaishnav 2009). The choice of the expression system is decided on based on the protein quality, the function, yield and speed of production. One of the mostly used expression systems is *E. coli* and continues to be the system of choice for protein expression and production (Chen 2012). The benefits of using *E. coli* as an expression system may include rapid growth and expression, ease of culture and high productivity (Swartz 1996). *E. coli* recombinant expression systems have also been developed mainly to facilitate maximum protein recovery, soluble protein production and ease in protein purification. It has been used for the cost-effective production of many commercially available proteins (Jonasson et al. 2002). Escherichia coli expression systems are widely used to produce recombinant proteins. For high-level protein production, BL21 (DE3) is the most suitable E. coli strain. It has the benefit of being deficient in both lon and ompT proteases, and it is compatible with the T7 lac O promoter system (Peti and Page 2007). The parallel use of affinity tags with recombinant DNA techniques, allows the facile modification of proteins of interest leading to efficient identification, production and isolation from the host system (Structural Genomics Consortium et al. 2008).

The most traditionally used expression systems are based on pET vectors which facilitate expression of a target gene under the control of the lac operator and T7 RNA polymerase promoter (Structural Genomics Consortium et al. 2008). The vectors are built for use in λDE3 lysogen strains of *E. coli*, which carry a genomic copy of the lac repressor controlled T7 RNA polymerase gene. When the conditions are repressive, T7 RNA polymerase is not produced, and transcription of the target gene is very insignificant. After induction, most of the cellular protein synthesis machinery is devoted to the production of the target protein (Studier et al. 1990). When using T7 systems, protein expression can be induced either by manipulating the carbon sources during E. coli growth or with the chemical inducer isopropyl-β-d-thiogalactoside (IPTG) (Studier 2005). Historically, the most commonly used antibiotic-selection marker has been ampicillin, but it has recently been replaced by carbenicillin. Vectors encoding resistance to chloramphenicol and kanamycin are now widely used as well (Studier et al. 1990). Recent advances in genomics, proteomics, and bioinformatics have facilitated the use of recombinant DNA technology to evaluate any protein of interest, without prior knowledge of the protein's cellular location or function. The parallel use of affinity tags with recombinant DNA techniques allows the simple modification of proteins of interest leading to efficient identification, production, and isolation from the host system (Structural Genomics Consortium et al. 2008).

Purification of recombinant proteins is most accomplished using a purification tag which can be located either at the N- or C-terminus of a protein of interest. Recombinant proteins produced in E. coli systems can also be purified using conventional chromatographic methods based ion exchange, on size exclusion, and hydrophobic interaction that separate proteins according to size, charge, and hydrophobicity, respectively (Rosano and Ceccarelli 2014). However, it is very challenging to find the proper combination of components and conditions (expression vectors, tags, linkers, growth conditions, expression cells, inducer concentration, and many more) to obtain a highly expressed and soluble recombinant protein that can be purified in large amounts (Bernier et al. 2018). The choice of the purification tag is particularly important in the design of the fusion protein. Still, there is no single tag which satisfies all requirements. However, many strategies have been developed that improve solubility and purity of the targeted recombinant proteins. These strategies include the addition of fusion tags, and some expression vector systems allow the expression of the protein of interest as a fusion partner to improve both solubility and purification (Esposito and Chatterjee 2006).

Purification of recombinant proteins is mostly accomplished using a purification tag which can be located either at the N- or C-terminus of a protein of interest. Recombinant proteins produced in *E. coli* systems can also be purified using conventional chromatographic methods based ion exchange, on size exclusion, and hydrophobic interaction that separate proteins according to size, charge and hydrophobicity, respectively (Rosano and Ceccarelli 2014). However, it is very challenging to find the proper combination of components and conditions (expression vectors, tags, linkers, growth conditions, expression cells, inducer concentration and many others) to obtain a highly expressed and soluble recombinant protein that can be purified in large amounts (Bernier et al. 2018). The choice of the purification tag is particularly important in the design of the fusion protein. Still, there is no single tag which satisfies all requirements. However, many strategies have been developed that improve solubility and purity of the targeted recombinant proteins. Some of these strategies include the addition of fusion tags, some expression vector systems allow the expression of the protein of interest as a fusion partner to improve both solubility and purification (Esposito and Chatterjee 2006).

As a chromatographic procedure, IMAC has the benefits of having robust, mild elution conditions, specific binding and the ability to regulate selectivity by using chromatography buffers with low imidazole concentrations. There is a wide array of common resins with slightly different binding capacities and binding strengths, but all tolerate harsh cleaning procedures (Structural Genomics Consortium et al. 2008). Most purification steps can be integrated by high-performance liquid chromatography; the most commonly used devices are the ÄKTA systems from GE Healthcare (Structural Genomics Consortium et al. 2008).

The purity of the recombinant protein can be improved by controlling the amount of recombinant protein to the size of the column; lower-affinity contaminants can be avoided with a relative excess of the histidine-tagged recombinant protein. After protein purification, samples are visualized by SDS-PAGE. If the protein is stained with a dye such as Coomassie brilliant blue, the intensity of the bands will usually be proportional to the amount of protein. This allows the purity of the sample to be estimated and whether the purified protein is of the expected size (Bradford 1976).

1.6.Problem statement

Hot springs are potential sources of thermostable enzymes and DNA manipulating enzymes; however, they have not yet been explored widely using metagenomic approaches. DNA manipulating enzymes are usually scarce locally as they are mostly distributed by overseas suppliers. The shortage in these enzymes limits the amount of research done locally. So producing enough of these enzymes will promote research and decrease the operational costs of molecular biology laboratories. Previous studies relied mostly on conventional method for gene mining and protein production, however, as it is understood today, only about 1- 10 % of microorganisms can be grown in the laboratory. Thus, there is a demand to explore the remaining 90- 99 % of bacteria which could not be captured using conventional culture-based techniques.

1.6. Rationale

Hot springs are a source of theoretically untapped thermostable enzymes of medical, pharmaceutical and industrial relevance including the DNA manipulating ones. Even though South Africa is endowed with some of these hot springs, such resources have not yet been comprehensively explored through conventional as well as metagenomics techniques. Thus, there is an absolute need to examine such resources and uncover potentially novel or isoforms of DNA-manipulating thermostable enzymes through the metagenomic approach. Through the expansion of the already known DNA manipulating enzymes, genes with novel properties are needed for the discovery of novel or improved molecular techniques and tools (Simon et al. 2009). Many of these enzymes (e.g. Thermostable DNA Polymerases) have been described and commercialised, and they add economic value. The need for new DNA Polymerases that combine the practical advantages of bacterial enzymes with improved thermostability has motivated the on-going screening of genes producing these enzymes (Moser et al. 2012).

1.7. Aim

To mine genes encoding for DNA manipulating enzymes such as DNA polymerase, DNA ligase, endonucleases from hot springs using metagenomic techniques and construct fosmid library.

1.8. Objectives

1. To extract metagenomic DNA from the hot springs
2. To sequence metagenomic DNA using Illumina MiSeq next-generation sequencing platform
3. To carry out *de novo* assembly and *in silico* sequence analysis using CLC Genomic Workbench
4. To construct a fosmid library using CopyControl™ HTP Fosmid Library Production Kit with pCC2FOS™ Vector
5. To express and purify the potential DNA manipulating enzymes

1.9. References

AL- MANASRA, A. & AL- RAZEN, F. 2012. Cloning and Expression of a new bacteriophage (SHP*h*) DNA ligase isolated from sewage. *Journal of Genetic Engineering and Biotechnology*, 10:177- 184.

ARBER, W. & LINN, S. 1969. DNA modification and restriction. *Annual Review of Biochemistry*. 38:467- 500.

BANACH- ORLOWSKA, M., FIJALKOWSKA, I. J., SCHOAPER, R. M. & JONCZYK, P. 2005. DNA polymerase II as a fidelity factor in chromosomal DNA synthesis in *Escherichia coli*. *Mol Microbiol.* 58; 1: 16- 70.

BANG, D. & CHURCH, G. M. 2008. Gene synthesis by circular assembly amplification. *Nat. Methods*. 5: 37- 39.

BARRANGOU, R., FREMAUX, C., DEVEAU, H., RICHARDS, M., BOYAVAL, P., MOINEAU, S., ROMERO, D.A. & HORVATH, P. 2007. CRISPR provides acquired resistance against viruses in prokaryotes. *Science*. 315; 5819:170912.

BASHIR, Y., SINGH, P.S. & KONWAR, K.B. 2014. Metagenomics; An application based perspective. *Chinese Journal of Biology*, 1460360. 7-15.

BEARD, W. A. & WILSON, S.H. 2006. Structure and mechanism of DNA polymerase Beta. *Chem. Rev.* 106; 2: 361- 82.

BELKOVA, N., TAZAKI, K., ZAKHOROVA, J.R. & PARFENOVA, R. 2007. Activity of bacteria in water of hot springs from southern and central Kamchatskay geothermal provinces, Kamchatka Peninsula, Russia. *Microbiological research*. 162:99-107.

BERNIER, C. S., CANTINI, L. & SALESSE, C. 2018. Systematic analysis of expression, solubility and purification of a passenger protein in fusion with different tags. 152: 92-106.

BERTANI G. & WEIGLE, J.J.1953. Host controlled variation in bacterial viruses. *Journal of Bacteriology*. 65; 2: 113–21.

BERTOCCI, B., DE SMET, A., WEILL, J.C. & REYNAUD, C.A. 2006. Nonoverlapping functions of polymerases μ, λ and terminal deoxynucleotidyltransferase during immunoglobulin V (D) J recombination in vivo. *Immunity*. 25; 31- 41.

BICKLE, T. A. & KRÜGER, D. H. 1993. Biology of DNA restriction. *Microbiological Reviews*. 57; 2: 434–50.

BOURNIQUEL, A.A. & BICKLE, T. A. 2002. Complex restriction enzymes: NTP- driven molecular motors. *Biochimie.* 84; 11: 1047-57.

BRADFORD, M.M. 1976. A rapid sensitive method for the quantification of microgram quantities of protein utilizing the principle of protein- dye binding. *Anal Biochem*. 72:248-254.

CASALI, N. 2003. *Escherichia coli* host strains, *E. coli* Plasmid Vectors. *Springer*. 27-48.

CHEN, R. 2012. Bacterial expression systems for recombinant protein production: *E. coli* and beyond. *Biotechnology Advances*. 30: 1102-1107.

COLORA, S. & ROLF, D. 2011. Metagenomic analysis: Past and future trends. *Applied and Environmental Microbiology*. 77; 4:1153- 1161.

CULLIGAN, P.E., SLEATOR, R.D., MARCHESI, J.R. & HILL, C. 2014. Metagenomics and novel gene discovery. *Virulence*. 5:3; 399- 412.

DEMAIN, A. L. AND VAISHNAV. P. 2009. Production of recombinant proteins by microbes and higher organisms. *Biotechnology Advances*. 27: 297-306.

DIAS, R.S., SILVA, L.C.F., ELLER, M.R., OLIVERIA, V.M., De PAULA, S.O.S. &

DRYDEN, D.T., MURRAY, N.E. & RAO, D.N. 2001. Nucleoside triphosphate-dependent restriction enzymes. *Nucleic Acids Research*. 29; 18: 3728–41.

DUSSOIX, D. & ARBER, W. 1962. Host specificity of DNA produced by Escherichia coli. II. Control over acceptance of DNA from infecting phage lambda. *Journal of Molecular Biology*. 5; 1: 37–49.

ENTCHEVA, P., LIELD, W., JOHANN, A., HARTSH, T. & STREIT, W.R. 2001. Direct cloning from enrichment cultures, a reliable strategy for isolation of complete operons and genes from microbial consortia. *App. Environ. Microbiol*. 67: 88-99.

ESPOSITO, D. AND CHATTERJEE, D. K. 2006. Enhancement of soluble protein expression through the use of fusion tags. *Current Opinion in Biotechnology*. 17: 353-358.

FAKRUDDIN, M.D., CHOWDHURY, A., NUR HOSSAIN, M.D., BIN MANNA, K.S. & MAZUMDAR, R.M. 2012. Pyrosequencing- principles and applications. *Internal Journal of Life Science & Pharma Research*. 2; 2:1-2.

FAN, X., LIU, X., HUANG, R. & LIU, Y. 2012. Identification and characterization of a novel thermostable pyrethroid-hydrolyzing enzyme isolated through metagenomic approach: *Microbial Cell Factories*. 11:33.

FERRER, M., BELOQUI, A., TIMMIS, K. N. & GOLYSHIN, P. N. 2008. Metagenomics for mining new genetic resources of microbial communities: *Journal of Molecular Microbiology and Biotechnology*. 16: 109-123.

FUHRMAN, J.A. 2012.Metagenomics and its connection to microbial community organisation. *F1000 Biology reports.* 4:1-5.

GABOR, E.M., de VRIES, J.E. & JANSSEN, D. B. 2013. Efficient recovery of environmental DNA for expression cloning by indirect extraction methods. *FEMS Microbiology Ecology.* 44: 153- 163.

GARCIA- DIAZ, M. & BEBENEK, K. 2007. Multiple function of DNA polymerase. *Critical Reviews in Plant Sciences*. 26; 2:105-122.

GARCIA- DIAZ, M., BEBENEK, K., PEDERSEN, L. C., LONDON, R.E. & KUNKEL, T. A. 2005. Structure- function studies of DNA polymerase λ. *DNA repair*. 4: 1358- 1367.

GIBSON, D.G., YOUNG, R.Y., VENTER, J.C., HUTCHISON, C.A. & SMITH, H. O. 2009. Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat. Methods.* 6; 5: 343-5.

GOMEZ, J. & STEINER, W. 2004. The biocatalytic potential of extremophiles and extremozymes. *Food Technol Biotechnol*. 42; 4: 223- 235.

GOODMAN, M.F. & TUPPIN, B. 2000. Sloppier DNA polymerases involved in genome repair. *Curr. Opin. Genet. Dev.* 10; 2: 162- 8.

HANDELSMAN, J. 2004. Metagenomics: application of genomics to uncultured microorganisms. *Microbiology and Molecular Biology Reviews*. 4: 669-685.

HANDELSMAN, J. 2009. Metagenomics and microbial communities. Encyclopaedia of life sciences University of Wisconsin- Madisa, Wisconsin, USA.

HANDELSMAN, J., RONDON, M. R., BRADY, S. F., CLARDY, J. & GOODMAN, R. M. 1998. Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products: *Chemistry and Biology.* 5:245-249.

HORVATH, P. & BARRANGOU, R. 2010. CRISPR/Cas, the immune system of bacteria and archaea. *Science*. 327;5962: 16770.

JAROSZ, D. F., BEUNING, J.P., COHEN, E.S. & WALKER, G.C. 2007. Y- Family DNA polymerases in *Escherichia coli*. *Trends in Microbiology*. 15:70-77.

JONASSON, P., LILJEQVIST, S., NYGREN, P.A. AND STÅHL, S. 2002. Genetic design for facilitated production and recovery of recombinant proteins in *Escherichia coli*. *Biotechnology and Applied Biochemistry*. 35: 91-105.

JONKER, C., GINKEL, C. & OLIVIER, J. 2013. Association between physical and geochemical characteristics of thermal hot springs and algal diversity in Limpopo province, South Africa. University of South Africa, Department of environmental sciences. 39; 1:95-104.

KAGUNI, M. J. 2018. The Macromolecular Machines that Duplicates the *Escherichia Coli* Chromosome as targets for Drug discovery. *Antibiotics*. 7;1: 23

KAKIRDE, K. S., PARSLEY, L. C. & LILES, M. R. 2010. Size does matter: application-driven approaches for soil metagenomics: *Soil Biology and Biochemistry*, 42.1911-1923.

KESSLER, C. & MANTA, V. 1990. Specificity of restriction endonucleases and DNA modification methyltransferases a review. 3rd Edition. *Gene*. 92; 1–2: 1–248.

KOBAYASHI, I. 2001. Behavior of restriction-modification systems as selfish mobile elements and their impact on genome evolution. *Nucleic Acids Research*. 29; 18: 3742–56.

KUMAR, S., KRISHMANI, K.K., BUSHAN, B. & BRAHMANE, M.P. 2015. Metagenomics; Retrospects and Prospects in High Throughput Age. *Biotech Res Int.* 2015: 121735.

KUNKEL, T.A. 2004. DNA replication fidelity. *J. Biol. Chem*. 17:16895- 8.

LEHMAN, I. R., BESSMAN, M. J., SIMMS, E.S. & KORNBERG, A. 1958. Enzymatic synthesis of deoxyribonucleic acid. I. Preparation of substrates and partial purification of enzyme from *Escherichia coli. J. Biol. Chem.* 223; 1:163-70.

LI, L. L., MCCORKLE, S. R., MONCHY, S., TAGHAVI, S. & van der LELIE, D. 2009. Bioprospecting metagenomes: glycosyl hydrolases for converting biomass: *Biotechnology for Biofuels*. 2.

LILES, M. R., WILLIAMSON, L. L., RODBUMRER, J., TORSVIK, V., GOODMAN, R. M. & HANDELSMAN, J. 2008. Recovery, purification, and cloning of high-molecular-weight DNA from soil microorganisms: *Applied and Environmental Microbiology*.74: 3302-3305.

LING, H., BOUDSOCQ, F., WOODGATE, R. & YANG, W. 2001. Crystal Structure of a Y- family DNA polymerase in action: A mechanism for Error- prone and lesion- bypass replication. *Cell Press.* 1: 91-102.

LOHMAN, G. J., TABOR, S., & NICHOLS, N. M. 2011. DNA ligases. *Curr. Protoc. Biol.* 3; 3: 14.

LOPEZ'-LOPEZ', O., CERDAN, M.E. & GONZALEZ-SISO, M.I. 2013. Hot springs metagenomics. *Life*, 2:308- 320.

LORENZ, P., LIEBETON, K., NIEHAUS, F. & ECK, J. 2002. Screening for novel enzymes for biocatalytic processes: accessing the metagenome as a resource of novel functional sequence space: *Current Opinion in Biotechnology*. 13: 572-577.

LURIA, S. E. & HUMAN, M.L. 1952. A nonhereditary, host-induced variation of bacterial viruses. *Journal of Bacteriology*. 64; 4: 557–69.

MARDIS, E.R. 2008. Next- generation DNA sequencing methods. *Ann Rev Genomics Hum Genet*. 9: 387-402.

MESELSON, M. & YUAN, R. 1968. DNA restriction enzyme from *E. coli*. *Nature*. 217; 5134:11104.

MILLER, E. S., HEIDELBERG, J. F., EISEN, J. A., NELSON, W. C., DURKIN, A.S., CIECKO, A., FELDLYM, T. V., WHITE, O. PAULSEN, I. T., NIEMAN, W. C., LEE, J., SZCYPINSKI, B. & FRASER, C. M. 2003. Complete genome sequence of the broad-host- range vibriophage KVP40: Comparative genomics of a T4- related bacteriophage. *J. Bacteriol.* 185: 5220- 5233.

MIRETE, S., MORGANTE, V. & GONZALEZ- PASTOR, J. E. 2016. Functional metagenomics of extreme environments. *Current Opin Biotechnol*. 38: 143-9.

MURRAY, N. E. 2000. Type I restriction systems: sophisticated molecular machines (a legacy of BERTANI and WEIGLE). *Microbiology and Molecular Biology Reviews*. 64; 2: 412–34.

NAKAMURA, T., ZHOO, Y., YAMAGATA, Y., HUA, Y. J. & YANG, W. 2012. Watching DNA polymerase n make a phosphodiester bond. *Nature*. 487; 7406: 196- 201.

NEELAKANTA, G. & SULTANA, H. 2013. The use of metagenomics approaches to analyse changes in microbial communities. *Microbiology insights*, 6:37- 48.

NINFA, J. A., BALOU, D. P. & BENORE, M. 2010. Fundamental Laboratory Approaches for Biochemistry and Biotechnology. Hoboken, N.J: *John Wiley & Sons*. p. 341. ISBN 978-0-470-08766-4

OHMORI, H., FRIEDBERG, E. C., FUCHS, R. P., GOODMAN, M. F., HANAOKA, F., HINKLE, D., KUNKEL, T.A., LAWRENCE, C.W., LIVNEH, Z., NOHMI, T., PRAKASH, S., TODO, T., WALKER, G.C., WANG, Z. & WOODGATE, R. 2001. The Y- family of DNA polymerases. *Mol. Cell.* 8; 1: 7-8.

OLIVIER, J., VAN NIEKERK, H. & VAN DER WALT, I. 2008. Physical and chemical characteristics of thermal springs in Waterberg area in Limpopo province, South Africa. *Water SA*. 2; 34:163- 174.

OLIVIER, J., VENTER, J.S. & JONKER, C.Z. 2011. Thermal and chemical characteristics of hot water spring in northern part of Limpopo province, South Africa. *Department of Environmental Sciences. Unisa*. 37; 4:427- 436.

PASCAL, J. M. 2008. DNA and RNA ligases: Structural variations and shared mechanisms. *Curr. Opin. Struct. Biol.* 18; 1: 96-105.

PETI, W. & PAGE, R. 2007. Strategies to maximize heterologous protein expression in *E. coli* with minimal cost. 51; 1:1-10.

PINGOUD, A. & JELTSCH, A. 2001. Structure and function of type II restriction endonucleases. *Nucleic Acids Research*. 29; 18: 3705–27.

PRAKASH, S., JOHNSON, R. E. & PRAKASH, L. 2005. Eukaryotic translession synthesis DNA polymerases: Specificity of structure and function. *Annu. Rev. Biochem.* 74: 317- 53.

RAMADAN, K., SHEVELEV, I. & HUBSCHER, U. 2004. DNA polymerase- X family: Controllers of quality? *Nature Reviews Molecular Cell Biology.* 5:1038- 1043.

RAYCHAUDHURY, P. & BASU, K. A. 2011. Genetic requirement for mutagenesis of G [8,5- Me] T cross- link in Escherichia coli: DNA polymerases IV and V compete for Error-Prone bypass. *Biochemistry.* 50; 12: 2330- 2338.

ROBERTS, R. J. 1976. Restriction endonucleases. *CRC Critical Reviews in Biochemistry*. 4; 2: 123–64.

RONDON, M. R., AUGUST, P. R., BETTERMANN, A. D., BRADY, S. F., GROSSMAN, T. H., LILES, M. R., LOIACONO, K. A., LYNCH, B. A., MACNEIL, I. A. & MINOR, C. 2000. Cloning the soil metagenome: a strategy for accessing the genetic and functional diversity of uncultured microorganisms: *Applied and Environmental Microbiology*. 66: 2541-2547.

ROSANO, G. L. AND CECCARELLI, E. A. 2014. Recombinant protein expression in microbial systems. *Frontiers in Microbiology*. 5: 1-2.

ROSSI, R., MONTECUCCO, A., CIARROCCHI, G. & BIAMONTI, G. 1997. Functional Characterization of the T4 DNA Ligase: A new insight into the mechanism of action. *Nucleic Acids Res.* 1; 11: 2106- 13.

ROTHWELL, P. J. & WALKSMAN, G. 2005. Structure and mechanism of DNA polymerases. *Adv. Protein Chem.* 71:401- 40.

RUSSELL, D.W. & SAMBROOK, J. 2001. Molecular cloning: a laboratory manual. Cold Spring Harbor, N.Y: Cold Spring Harbor Laboratory. ISBN 0-87969-576-5.

SCHLOSS, D. P. & HANDELSMAN, J. 2003. Biotechnological prospects from metagenomics. *Current opinions in Biotechnology.* 14:303-310.

SEO, M.S., YUN, M.S., JEONG, J.K., CHOI, J., LEE, S. M., KIM, J.H., LEE, J. H. &

SHUMAN, D.G., YOUNG, L., CHUANG, R.Y., VENTER, J. C., HUTCHISON, C. A. & SMITH, H.O. 2009. Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat. Methods.* 6; 5: 343- 5.

SILVA, C.C. 2014. Metagenomics: Library Construction and Screening. *Methods*. 45-65.

SIMON, C. & DANIEL, R. 2009. Achievements and new knowledge unraveled by metagenomic approaches: *Applied Microbiology and Biotechnology*. 85: 265-276.

SIMON, C., HERALCH, J., ROCKSTROH, S. & DANIEL, R. 2009. Rapid identification of genes encoding DNA polymerases by function- based screening of metagenomic libraries derived from Glacial Ice. *Applied and Environmental microbiology.* 75:2964-2968.

SINGH, S. P., SAGAR, K. & KONWAR, B. K. 2013. Strategy in metagenomic DNA isolation and computational studies of humic acid: *Current Research in Microbiology and Biotechnology*. 1: 9-11.

SISTLA, S. & RAO, D. N. 2004. S-Adenosyl-L-methionine-dependent restriction enzymes. *Critical Reviews in Biochemistry and Molecular Biology*. 39; 1: 1–19.

SMITH, H. O. & WILCOX, K. W. 1970. A restriction enzyme from *Hemophilus influenzae.* I. Purification and general properties. *Journal of Molecular Biology*. 51; 2: 379–91.

STRUCTURAL GENOMICS CONSORTIUM, CHINA STRUCTURAL GENOMICS CONSORTIUM, NORTHEAST STRUCTURAL GENOMICS CONSORTIUM, GRASLUND, GUNSALUS, K. C. 2008. Protein production and purification. Nature methods. 5; 2: 135- 46.

STUDIER, F. W., ROSEBERG, A.H., DUNN, J.T., DUDENDORFF, J.W. 1990. Use of T7 RNA polymerase to direct expression of cloned genes. *Methods enzymol*. 185: 60- 89.

STUDIER, F.W. 2005. Protein production by autoinduction in high density shaking cultures. *Protein Expr Purif.* 41;1:207- 34.

SWARTZ, J. R. 1996. *Escherichia coli* recombinant DNA technology. *Escherichia coli* and *Salmonella*: cellular and molecular biology, 2nd ed. Washington, DC. A*merican Society of Microbiology Press*. 1693-1711.

TEKERE, M., LOTTER, A., OLIVIER, J. & VENTER, S. 2015. Bacterial diversity in some of South African thermal springs; Metagenomic Analysis. *Proceedings World Geothermal Congress*. 19-25.

TEKERE, M., LOTTER, A., OLIVIER, J., JONKER, N. & VENTER, S. 2012. Metagenomic analysis of bacterial diversity of Siloam hot water spring, Limpopo, South Africa. *African Journal of Biotechnology.* 78:18005- 18012.

TODD, D.K. 1980. Groundwater Hydrology. 2nd Edition, *John Wiley & Sons*, New York.

TORSTEN, T., GILBERT, J. & MEYER, F. 2012. Metagenomics, a guide from sampling to data analysis. *Microbial informatics and experimentation*. 2:1-3.

TUFFIN, M., ANDERSON, D., HEATH, C. & COWAN, D. A. COWAN. 2009. Metagenomic gene discovery: how far have we moved into novel sequence space: *Biotechnology Journal*. 4: 671-1683.

TYSON, G. W., CHAPMAN, J., HUGENHOLTZ, P., ALLEN, E. E., RAM, R. J., RICHARDSON, P. M., SOLOVYEV, V. V. RUBIN, E. M., ROKHSAR, D. S. & BANFIELD, J. F. 2004. Community structure and metabolism through reconstruction of microbial genomes from the environment: *Nature*. 428: 37-43.

VILLA-KOMAROFF, L., EFSTRATIADIS, A., BROOME, S., LOMEDICO, P., TIZARD, R., NABER, S. P., CHICK, W. L. & GILBERT, W. 1978. A bacterial clone synthesizing proinsulin. *Proceedings of the National Academy of Sciences of the United States of America*. 75; 8: 3727–31.

WILKISON, A., DAY, J. & BOWATER, R. 2001. Bacterial DNA ligases. *Mol Microbiol.* 40; 6: 1241- 8.

WILLIAMS, R. J. 2003. Restriction endonucleases: classification, properties, and applications. *Molecular Biotechnology*. 23; 3: 225 -43.

WILSON, R. H., MORTON, S. K., DEIDERICK, H. A., GERTH, L. M., PAUL, A. H., GERBER, I., PATEL, A., ELLINGTON, D.A., HUNICKE- SMITH, S. P. & PATRICK, M. W. 2013. Engineer DNA ligases with improved activities in vitro. *Protein Engineering, Design & Selection*. 26: 471- 478.

WILSON, S. H., SOBOL, R. W., BEARD, W. A., HORTON, J. K., PRASAD, R. & VAN DE BERG, B. J. 2000. DNA polymerase beta and mammalian base excision repair. *Cold. Spring. Harb. Symp. Quant. Biol.* 65: 143- 155.

WINNACKER, E. L .1987. Isolation, Identification, and Characterization of DNA fragments. From Genes to Clones. VCH. ISBN 0-89573-614-4.

YUAN, R. 1981. Structure and mechanism of multifunctional restriction endonucleases. *Annual Review of Biochemistry*. 50: 285–319.

ZAHOOR, A., LINDNER, S. N. & WENDISCH, F. V. 2012. Metabolic engineering of Corynebacterium glutamicum aimed at alternative carbon sources and new products. *Compu. Struct Biotechnol. J.* 3: e201210004.

ZHANG, R., ZHU, Z., ZHU, H., NGUYEN, T., YAO, F., XIA, K., LIANG, D. & LIU, C. 2005. SNP Cutter: a comprehensive tool for SNP PCR-RFLP assay design. *Nucleic Acids Research*. 33: W489–92.

ZIERENBERG, R. A., ADAMAS, M. W. & ARP, A.J. 2000. Life in extreme environments: Hydrothermal vents. *Proc. Natl. Acad. Sci. USA*. 97; 24: 12961- 12962.

2. CHAPTER TWO: Hot Spring DNA Extraction, Metagenomics Library Construction and Sequencing

Table of Contents

Abstract

The hot spring metagenome of Mphephu, Siloam and Tshipise were explored for the presence of DNA polymerase, DNA ligase and endonuclease II genes using a cultivation-independent approach, metagenomics. Among the various protocols tested, the SDS based CTAB method was found to be the best for metagenome isolation from hot spring soil sediments under investigation. The purity of the isolated metagenomic DNA was somewhat not suitable for gene cloning. To improve on the yield and purity of isolated metagenomic DNA, gel extraction and electroelution were used to clean up the extracted DNA. On average 8 µg, 6 µg, and 5.5 µg of DNA/g were extracted for three locations with 280/260 ratio 1.65 -1.79 after electroelution. A metagenome expression library of approximately 2.16 x10^3 clones was successfully constructed using a CopyControl^{TM} Fosmid Library kit. After sequencing of the metagenome, a total of 5,681,662 reads was produced after raw data processing using CLC Bio genomic workbench 9.1.5. The assembly generated 7338 contigs of which 44 contigs were > 1kb, and 71883 <1kb. The homology to Genes for DNA manipulating enzymes (DNA ligase, DNA polymerase and endonucleases) sequences of the extracted ORFs were searched using BLAST algorithm in NCBI. The list contained 57 distinct genes for DNA polymerase, 29 genes for DNA ligase and more than 100 genes for endonuclease enzymes. The alignment of mined genes revealed that they have 74- 98% identity at the amino acid level with several homologous DNA manipulative enzymes from *Bacillus spp.*

Key words: Hot spring, Metagenomic, Expression library, DNA-manipulative enzymes

2.1. Introduction

Over the past two decades, various extreme environments such as steam vents, hydrothermal vents, hot springs etc. have been explored for their microbial diversity and as a source of industrially essential enzymes (Xie et al. 2011; Sofia et al. 2014). The drawback of such extreme environments is that over 90 % of the microorganisms are not cultivable resulting in culture-dependent techniques being ineffective and cultivation-independent methods indispensable for microbial exploration (Mardanov et al. 2011). In metagenomics, whole genomic DNA can be extracted directly from the environmental sample without a need to cultivate microorganisms. As the composition of different habitats varies in terms of their mineral composition, organic and inorganic compounds, microbial population and biotic factors, standardisation of DNA extraction protocol is considerable for a particular niche or environment (Dias et al. 2014).

There are two types of metagenomic DNA extraction approaches namely, indirect and direct DNA extraction (Gabor et al. 2003; Dias et al. 2014). In the indirect DNA extraction methods, microbial cells are first collected before cell disruption with a lysis solution or reagent, whereas direct DNA extraction do not involve the collection of cells, but the environmental sample is directly lysed to extract metagenomic DNA. Direct DNA extraction methods are time-consuming, labour intensive but they extract high DNA concentrations with high levels of impurities. Despite some drawbacks, direct DNA extraction techniques are commonly employed for extraction of metagenomic DNA from complex environmental samples like hot springs soil sediments, compost soil, etc. due to their excessive DNA yield (Dias et al. 2014). It has been stated in Tebbe and Vahjen (1993) that less than 0.8 μg/ml of humic acid can inhibit the restriction enzymes activity even at high DNA concentrations. This drawback highlights the need for further purification techniques once DNA has been extracted to reduce humic acid concentration for downstream processes.

After isolation and purification of environmental DNA samples, the metagenomic library is constructed and or metagenome sequenced. The construction of the library consists of the cloning of DNA fragments into specific vectors and subsequently inserted into a host cell strains, followed by screening for genes of interest. Activity screening in the function-based metagenomics approach is always accomplished by high throughput screening of library clones on indicator media. It also includes the design of DNA probes or primers which are derived from conserved regions of already characterized genes or protein families. By so doing, only the unique functional classes of proteins can be identified. This strategy has led to the successful mining of genes coding for novel enzymes (Carola and Rolf 2011).

Discoveries of novel natural products and proteins have also been achieved using sequencing of metagenomic clones (Schloss and Handelsman 2003). Current developments in next-generation sequencing (NGS) technologies now render researchers access to the vast databases of DNA sequence information for several numbers of microorganisms. Consequently, the use of metagenomics techniques as tools to identify enzymes of interest has grown in recent years in many areas of biological research (Sebastian et al. 2013). Therefore, the current study was thus aimed at selecting and optimizing DNA extraction protocol in order recover high quality and high molecular weight metagenomic DNA that can be used for down stream processes such as sequencing, library construction and cloning.

2.2. Materials and Methods

2.2.1. **Sampling**

Soil sediments were collected from hot springs in Limpopo province, South Africa. Soil samples were collected at Mphephu, Siloam and Tshipise hot springs in April 2016. The samples were transported to the laboratory (Vaal University of Technology, Vanderbijlpark, South Africa) under sterile conditions, and DNA extractions were performed immediately.

2.2.2. Metagenomic DNA extraction

Five grams of soil samples were mixed with 13.5 ml of DNA extraction buffer containing (0.1 M Tris-HCl [pH 8.0], 0.1 M sodium EDTA [pH 8.0], 0.1 M sodium phosphate [pH 8.0], 1.5 M NaCl, 1% CTAB) and 100 ml of proteinase K (10 mg/ml) in 50 ml Eppendorf tubes by horizontal shaking at 225 rpm for 30 min at 37 °C. After shaking, 1.5 ml of 20% SDS was added to the samples, and then incubated in a 65 °C water bath for 2 h with gentle end-over-end inversions every 30 minutes. The supernatants were collected after centrifugation at 5000 rpm for 10 min at room temperature and transferred into new sterile 50-ml centrifuge tubes. The soil pellets were extracted two more times by adding 4.5 ml of the extraction buffer and 0.5 ml of 20% SDS, vortexing for 10 s, and then incubated at 65 °C for 10 min. Centrifugation was carried out as before. Supernatants from the three cycles of extractions were combined and mixed with an equal volume of chloroform-isoamyl alcohol (24:1, vol/vol). The aqueous phase was recovered by centrifugation and precipitated with 0.6 volume of isopropanol at room temperature for one hour. The pellet of crude DNA was obtained by centrifugation at 14000 rpm for 20 min at room temperature, washed with cold 70% ethanol, and resuspended in sterile deionised water, to give a final volume of 500 µl (Zhou et al. 1996).

2.2.2.1. Determination of metagenomics DNA size

The size of the isolated DNA was analysed by electrophoresis in 0.8 % agarose gel prepared in 1× TBE buffer containing 0.5 µg/mL ethidium bromide (Sambrook and Russell 2001). The TBE Buffer was prepared as follows [108 g Tris base, 55 g boric acid, 7.45 g EDTA and filled up to 1 L with dH2O]. Before electrophoresis, DNA samples were mixed with loading dye and loaded on 0.8 % agarose gel, at 100 V for 60 minutes. A 10kb molecular marker was used to help predict the size of the extracted DNA.After electrophoresis, the gel was visualised and photographed using a digital imaging system UV-transilluminator (SYNGENE G- Box).

2.2.2.2. DNA purification using electroelution

Brown DNA samples (humic acid contaminated) were run on 0.8 % agarose gel for 45 min, at 100 V. Subsequently, DNA bands were cut from the gel using sterile scalpels. About 8- 10 pieces of cut gel slices were placed in the electro eluter as set up according to model 422 electro eluter instruction manual catalogue number 165- 2976, Bio-Rad. After the DNA was eluted, it was pooled and subjected to further purification as follows. One-tenth of 3 M sodium acetate was added to the DNA sample, then gently mixed. Equal volumes of 24: 1 chloroform isoamyl was added, gently mixed then centrifuged for 2 minutes at 14000 rpm. A supernatant was removed and mixed with 0.7 volume isopropanol. DNA was precipitated at 14000 rpm for 20 minutes. The resulting pellet was washed with 70 % ethanol, then centrifuged for 20 minutes at 14000 rpm. DNA was resuspended in 20µl TE buffer.

2.2.2.3. The determination of DNA quantity and quality

DNA samples were quantified using the Qubit dsDNA HS Assay Kit (Life Technologies) and the Qubit 2.0 Fluorimeter (Life Technologies) as per the manufacturer's instruction. One microliter of DNA sample was suspended in Qubit dsDNA HS buffer in clear plastic Qubit Assay Tubes (catalogue number Q32856, Life Technologies) and measured on the fluorimeter. Alternatively, the purity and DNA concentration was determined using a NanoDrop OneC Spectrophotometer (Thermo Scientific, USA) by measuring absorbance at 260 nm and 280 and calculating the (A260/280) ratio.

2.2.3. DNA Sequencing and in silico mining of the genes

Sequencing of pooled DNA was performed using the Illumina MiSeq system at Agricultural Research Council in Pretoria (South Africa). Sequencing raw data was processed using CLC Bio genomic workbench 9.1.5. The sequences were analysed using Bio edit (Hall 1999). The website NCBI algorithm was used to predict the open reading frames (ORFs) and to also compare the sequenced gene to other proteins in the database by using the basic local alignment search tool (BLAST) for protein (Altschul et al. 1997).

2.2.4. Fosmid library construction

The EPI300-T1R Plating strain was supplied as a glycerol stock. Before beginning the Copy Control Fosmid Library Production kit procedure (Epicentre), the EPI300-T1R cells were streaked out on an LB plate without incorporating any antibiotic. The cells were grown at 37 °C overnight and then stored at 4 °C. After assessing the purity and quantity of the metagenomic DNA, the DNA was end-repaired to blunt, 5'-phosphorylated ends. The end-repair step was performed by mixing the sheared metagenomic DNA with sterile water, End-repair buffer, dNTPs, ATP and End-repair enzyme as described on the protocol from the EpiFOS™ Fosmid Library Production Kit (Epicentre). The end- repair mixture was incubated at 25 ºC for 45 minutes then to deactivate the end- repair enzymes, the sample was incubated at 70 °C for 10 minutes. The deactivated sample was then left on the bench for 30 minutes before proceeding to the purification process using the GeneJET DNA purification Kit manual.

The DNA was then purified and concentrated using the GeneJET DNA Purification Kit. The second process after the end-repair step was to ligate the purified blunt-ended DNA to the Cloning-Ready CopyControl pCC2FOS Vector. The ligation took place by mixing 10X Fast-Link Ligation Buffer, 10 mM ATP, CopyControl pCC2FOS Vector (0.5 µg/µl), concentrated purified end-repaired DNA (0.25 µg), Fast-Link DNA Ligase as provided on the EpiFOS™ Fosmid Library Production Kit (Epicentre) manual. The ligation mixture was then incubated at room temperature for 4 hours. Ten millilitres of the ligation mixture reaction was kept on the bench while thawing the MaxPlax Lambda packaging extract for the packaging reaction. The day before the Lambda Packaging reaction, 50 ml of LB broth + 10 mM MgSO4 + 0.2% Maltose was inoculated with a single colony of EPI300-T1R cells, and the flask was shaken overnight at 37 °C at 220 rpm. Then, the ligation reaction mixture was packaged into the thawed MaxPlax Lambda extract as directed by the protocol; then they were used to infect the EPI300-T1R phage resistant strain. A volume of 55 µL infected cells were plated on LB agar supplemented with 12.5 µg/mL chloramphenicol and incubated at 37 °C overnight to select for CopyControl fosmid clones. The number of colony forming units was determined using the following equation:

$$Titre = \frac{(\#\ of\ colonies)(\ dilution\ factor)}{/volume\ phage\ plated(\ \mu l)}$$

The CopyControl fosmid clones were stored by resuspending all colonies from the agar surfaces using approximately 2 mL of LB broth supplemented with 12.5 µg/mL chloramphenicol for each plate. One point six millimetres of the mixture was mixed with 0.4 ml of 100% sterile glycerol to make 20% final concentration of glycerol stock, into cryo-vial tubes. The tube was stored at -80 °C for long-term storage.

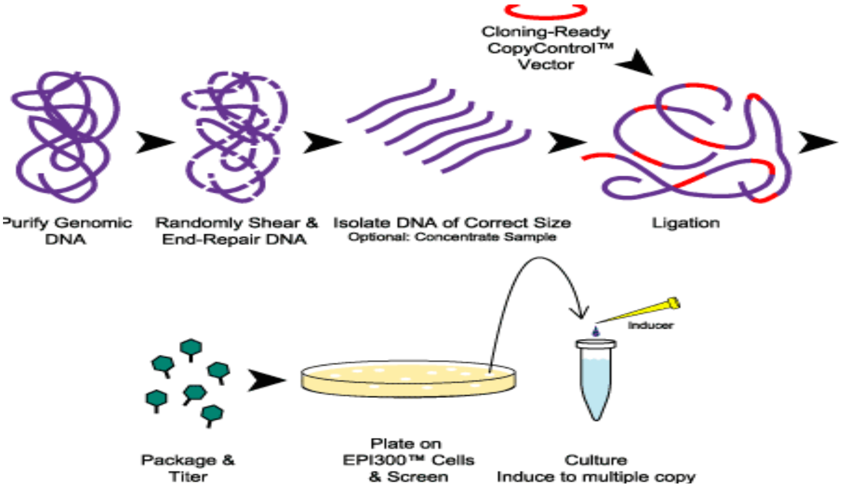

Cloning-Ready
CopyControl™
Vector
Purify Genomic DNA
Randomly Shear & End-Repair DNA
Isolate DNA of Correct Size
Optional: Concentrate Sample
Ligation
Inducer
Package & Titer
Plate on EPI300™ Cells & Screen
Culture
Induce to multiple copy

Figure 2.1. Metagenomic DNA library construction workflow (https://www.cephamls.com/BAC-Fosmid-Library-Construction)

2.2.4.1. Determination of library insert size

After constructing the fosmid library, to validate the success of library preparation, the size of the insert needed to be confirmed. Two clones were randomly picked from the master plate and were each suspended in 50 ml of LB broth supplemented in 12.5 µg/ml chloramphenicol. The culture was incubated at 37 °C for 16 hours and shaken at a speed of 150 rpm. After incubation, 5 ml of the 16 hours culture was transferred to a second sterile Erlenmeyer flask containing 45 ml of the LB broth + 12.5. µg/ml chloramphenicol in order to be induced with 50 µl of the 500X auto-induction solution for high-copy number of plasmids. The mixture was shaken for 5 hours at 37 °C at a speed of 150 rpm. After the induction step, the plasmid (vector + insert DNA) was extracted using the GeneJET Miniprep Plasmid Kit as per manufacturer’s protocol (Epicentre).

Once the plasmid was obtained, the quantification of the plasmid was carried out using NanoDrop™ OneC (Thermo Fisher Scientific, Waltham, MA, USA), and the gel electrophoresis was used to determine the size of the inserted DNA to validate cloning process. Thereafter, a considerable amount of plasmid DNAs were endonuclease restricted with *XbaI* restriction enzymes. The size of the inserted DNA was calculated by the sum of the DNA pieces minus the size of the pCC2FOS vector.

2.3. Results and Discussion

2.3.1. Sampling

Sediments were collected from three hot springs in Limpopo province; Siloam, Tshipise and Mphephu locations during April 2016. The exact location coordinates of sampling areas were taken. The samples were collected from the soil sediments surface (0–10 cm) using sterilised spatulas and sterile 50 mL falcon tubes. Physicochemical and environmental characteristics of the samples such as their pH and temperature were determined. Precautions were taken during sampling and handling of samples to preserve their integrity for microbiological analyses. The samples were labelled and placed in a cooler box for transportation and then stored in the 4 °C until DNA was extracted the following day.

Table 2.1 The three hot springs that were studied and their respective characteristics at the time of sampling.

Spring Name	Sample site coordinates	Temperature	pH
Mphephu	22°54'30.3"S 30°11'03.0"E	44.11° C	8.90
Tshipise	22°53'41.6"S 30°11'39.7"E	59.50° C	8.63
Siloam	22°53'39.4"S 30°11'41.4"E	70.20° C	9.15

Accordingly, the physicochemical and environmental characteristics of the samples such as pH and temperature ranged from 8.63 – 9.15 and 44.11 – 70.20, respectively (**Table 2.1**). Besides, both Mphephu and Tshipise are characterized by sandy soil with pH values that are circumneutral, the temperature of these hot springs differ widely. Although Siloam hot spring has pH value nearly the same as Mphephu, its temperature is the highest making it the hottest spring in the area at the time of sampling. Accordingly, Siloam hot spring is more alkaline than Tshipise and Mphephu and it is characterized by clay and loamy soil. The coordinates of each sampling site are also recorded in **Table 2.1**. The in-depth study of physicochemical characteristics of these hot springs are outlined in Olivier et al. (2011). According to literature and the records of **Table 2.1**, Siloam is by far the hottest spring in South Africa (Olivier et al. 2011).

2.3.2. Metagenomic DNA extraction

As a matter of fact, it is important to optimize the extraction protocol since the chemical and physical characteristics of different soils are not the same (de Castro et al. 2011). Besides, metagenomic DNA meant for a fosmid library construction should be of high quality, literally free of any possible contaminants and of a high concentration, ideally up to 20 µg (de Castro et al. 2011).

Accordingly, metagenomic DNA was extracted from three samples collected from Mphephu, Siloam and Tshipise in triplicates and later pooled together after extraction. It is, however, important to note that the extracted metagenomic DNA was subsequently purified to remove humic acid since direct extraction of DNA can result in co-extraction of humic acids which interfere with cloning efficiency and transformation efficiency (Daniel 2005). Moreover, different extraction methods were utilized to extract metagenomic DNA from the soil though the yields were not as much as that of conventional CTAB method. TheZymoResearch and Nucleospin DNA extraction protocols do yield fairly enough concentrations of DNA, however due to a high humic acid contamination, the yields were compromised during the clean up steps (Daniel 2005). It is also important to mention that some studies revealed that kits can be biased depending on efficiency and the methods of cell lysis and DNA purification which may affect downstream applications. Essentially, kits are not suitable for high molecular weight DNA that can be used in fosmid library construction (Daniel 2005).

Thus, the the SDS based CTAB method was used for the purpose of this study and on average concentrations of 8 µg, 6 µg, and 5.5 µg of DNA/g of soil were recovered for Mphephu, Siloam and Tshipise, respectively. The extraction of DNA from soil sediments produced a brownish coloured DNA with purity ranging between 1.3 and 1.5 using Nanodrop One C. DNA was diluted accordingly and 2µl was ran on 0.8 % agarose gel to determine the integrity and the size of DNA as shown in **Figure 2.2** below.

DNA recovery can be influenced by a number of soil properties such as vegetation cover, water content, chemical properties and a type of soil in question as explained by Zhou et al. (1996). Other factors such as soil particle size and water content were correlated to DNA yields in a study by Burgamann et al. (2001). Other factors that can explain why there were inconsistencies with DNA extractions can be given by the sensitivity of the individual method in assessing the metagenome, whether DNA is freed from the sample or not, and even the diversity of gram positive and negative microorganisms in the environment can affect how DNA it is recovered (Kauffmann et al. 2004). Also, the inconsistencies in yield can be attributed to DNA binding to particles in the soil (Zhou et al. 1996). Clay and humus particles are negatively charged and bind to cations, which lead to reduced DNA yields because of the adsorption free DNA on clay and also on organic matter particles (Sagova- Mareckova *et al.* 2008).

2.3.2.1. Determination of metagenomic DNA size

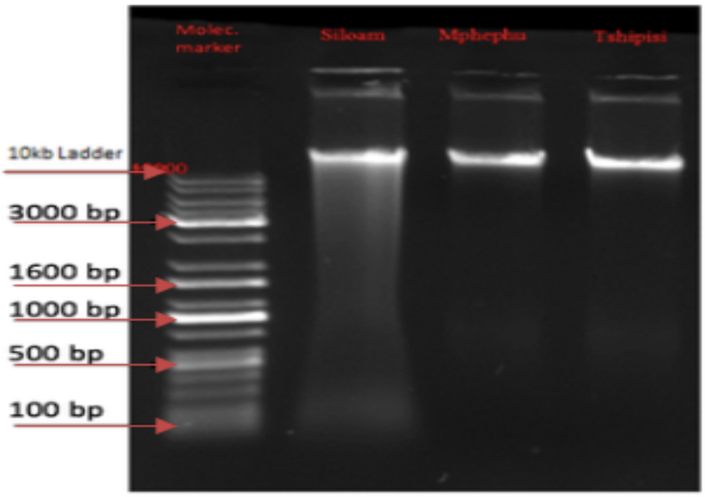


Figure 2.2. 0.8 % agarose gel, 100 voltage for 50 minutes. Lane 1- 4 represent in 1 X TBE buffer; 10kb kappa molecular ladder, DNA extracted from Siloam, Mphephu and Tshipise sediments respectively.

In **Figure 2.2** a 0.8% agarose gel representation of DNA extracted from three locations namely, Siloam, Mphephu and Tshipise using SDS based CTAB extraction method. A 10kb ladder was used from Kappa biosystems (South Africa) to identify the size of the extracted DNA and also to visualize DNA quantity and the integrity. As shown in **Figure 2.2**, high molecular weight DNA was successfully extracted delineated by DNA bands forming above 10kb ladder marker. It can also be deduced that DNA recovered was fairly intact for L3 and L4 (Mphephu and Tshipise) whereas for Siloam a smearing can be seen, which can be explained by **Figure 2.3** (**A**). Siloam sediments are characterized by clay and humus which makes it difficult to extract good quality DNA as compared to Mphephu and Tshipise which were more sandy soil with less humus. The colour of DNA extracted was brownish in colour for all the sampling places however Siloam DNA was the darkest of the three DNA samples.

Since down -stream applications such as PCR amplification, library construction and sequencing require DNA with less impurities to pure DNA, further purification steps were necessary for the three DNA samples. Usually organic matter is the major source of inhibitors that may be coextracted with the microbial DNA present with in the soil. Majorly, humic acids create considerable problem like interference in activity of DNA polymerase used for PCR reactions. As humic acid contains the same charge and size characteristics like DNA, it exhibits absorbance at both 230 and at 260 nm and hence interferes in quantification of DNA. This characteristic can be used to find out the level of contamination of humic acid in an isolated DNA sample (Courtois et al. 2001). To circumvent the challenges of high humic acid coprecipitation along with DNA, gel extraction method coupled with electroelution was used to remove these impurities to acceptable levels as shown in **Figure 2.3**. Before electroelution brownish DNA was observed in the microcentrifuge tube compared to after DNA was purified by electroelution. To have a representation of three locations, DNA samples were pooled so as to capture the metagenome of three hot springs.

2.3.2.2. DNA Purification using Electroelution

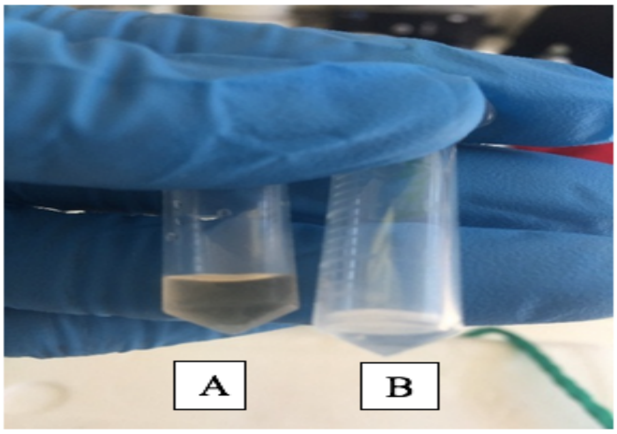


Figure 2.3. DNA sample before (A) and after (B) purification with electroelution.

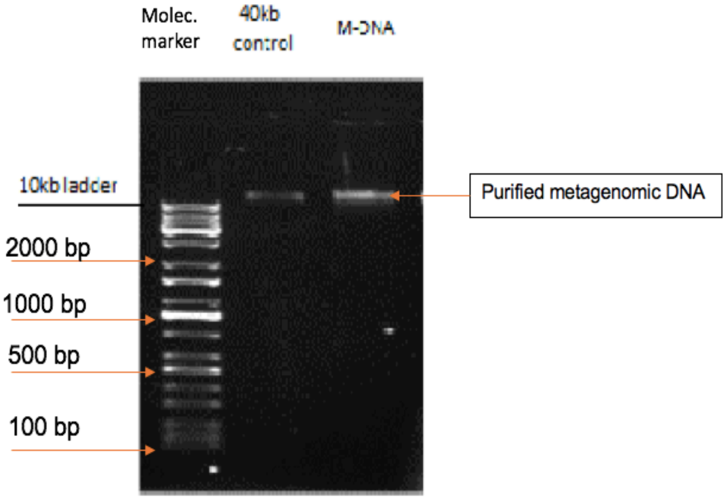


Figure 2.4. A 0.8 % agarose gel, 100 V for 50 minutes. Lane 1- 3 represents; 10kb molecular ladder (kapa biosystems), 40 kb control DNA and pooled metagenomic DNA from 3 location respectively.

In **Figure 2.3.** two microcentrifuge tubes containing DNA samples designated **A** and **B,** are shown to compare between two DNA samples. Centrifuge tube A contains pooled DNA samples from three different environments following extraction using SDS based CTAB method, while centrifuge tube B contains DNA sample after purification with electroelution. Recovered DNA after extraction was brownish in color suggestive of impurities among which is humic acid, carbohydrates etc. that are still attached to the DNA after extraction was completed. The 280/260 ratio for this DNA was very low ranging between 1 and 1.4 for all extracted DNA samples. To improve on the quality of extracted DNA in **Figure 2.3.A** was cleaned up using electroelution method as described in the methodology section and the results were a DNA with reduced brown color and 280/260 ratio between 1.65 -1.79 (**Figure 2.3. B**). Electroelution as performed to purify DNA is a simple and rapid technique for retrieving clean DNA from environmental samples with very different organic and mineral contents. It eliminates the need for labour -intensive steps and the use of expensive kits or dangerous chemicals. By pulling out the DNA selectively with an electrical current, it is possible to recover clean DNA extracts, appropriate for a wide number of downstream applications. The suitability of the samples for PCR is greatly improved, in some cases make possible the use of PCR for otherwise unsuitable samples (Kallmeyer and Smith 2009).

In **Figure 2.4.**, electroelution purified DNA was subjected to 0.8% agarose gel electrophoresis to check the integrity and the size of DNA after purification. A 40 kb control DNA was also loaded near purified DNA to see if extracted DNA has molecular weight necessary for fosmid library construction (about 40kb). As shown in the picture, the purified DNA was above 10kb ladder band and it was the same size as the control DNA (40kb) which indicates that the purified DNA was suitable for fosmid library construction and other downstream applications in term of size and purity as measured on Nanodrop One C spectrophotometer. The band intensity, the absence of smearing or fragmentation of DNA fragment on the gel as well as reduced levels of contaminants in purified DNA supports the argument by Kallmeyer and Smith (2009), that electroelution could be a better alternative for DNA purification.

2.3.3. DNA sequencing, *de novo* assembly and *in silico* gene mining

Sequencing of a pool of DNA was performed using the Illumina MiSeq system at Agricultural Research Council in Pretoria (South Africa). A total of 5,681,662 reads were produced after raw data processing using CLC Bio genomic workbench v 9.1.5. The *de novo* assembly generated 7338 contigs of which 44 contigs were > 1 kb, and 7294 <1 kb. The > 1 kb contigs were mapped to NCBI database to identify bacterial species as well as enzymes available in the library. The homology to Genes for DNA manipulating enzymes (DNA ligase, DNA polymerase and endonucleases) sequences of the extracted ORFs were searched using BLAST algorithm in NCBI. Accordingly, the list contained 57 distinct genes for DNA polymerase, 29 genes for DNA ligase and more than 100 genes for endonuclease enzymes.

Table 2.2. Summary for a BLAST search of putative DNA polymerase I, DNA ligase and endonuclease II

Genes	**Nucleotide length (bp)**	**Amino acid length**	**Mw (kDa)**	**Identity %**	**Amino acid sequence**	**Best hit**
DNA Polymerase I	2690	885	100	80	MGITMQPVLTTSAPSGGNWRYEAKYDGYRGLLKISAAGDVSLISRNAQPLENTFPEITEFAKSMIENLKEHLPITIDGEIVSLTNRFRSRFEYVQKRGLSKKAELIEQAAAKKPCQYLAFDLLVFKGESLTSLPYTERKRVLSDLMKELGLPMAPDPMAHARIQYIPDTSDFHALWNAVKRFDGEGIVAKKKDSRWAENKKTAEWLKLKNYKKAAVFMTGYNMANRYLTIAVYDRGQIKEVGSVSHGLGEQERNAILSIVKQYGTETKPGEYTIDPSICMTVHYLTIHYGTLREVSFVSFEFDMAWEDCTYKRLLLHSRNVHPDLQLTSLDKVIFPKSNKTKADYIGYLNEIGDFLLPFLDNRALTVIRYPHGSGGESFFQKNKPDYAPEFITTIRDDEHEHIICSDYSVLLWLANQLALEFHIPFQTADTTRPTEIVFDLDPPSRSEFPLAVRAANELHRLFEQLGLLSFPKLSGNKGIQIYIPISKNAFTYEETRLFTSFAASYCVSLFPDLFTTERLIKNRGGKLYIDYVQHAPGKTIICPYSTRGNQIGTVAAPLFWDEVHSDLAPSNFTMEAVIKRTKELGCPFESFFRQPQDKQIKAILDHLKEIDRSEN	*Bacillus paralicheniformis*
DNA ligase	1881	622	71	98	MGITMQPVLTTSAPSGGNWRYEAKYDGYRGLLKISAAGDVSLISRNAQPLENTFPEITEFAKSMIENLKEHLPITIDGEIVSLTNRFRSRFEYVQKRGLSKKAELIEQAAAKKPCQYLAFDLLVFKGESLTSLPYTERKRVLSDLMKELGLPMAPDPMAHARIQYIPDTSDFHALWNAVKRFDGEGIVAKKKDSRWAENKKTAEWLKLKNYKKAAVFMTGYNMANRYLTIAVYDRGQIKEVGSVSHGLGEQERNAILSIVKQYGTETKPGEYTIDPSICMTVHYLTIHYGTLREVSFVSFEFDMAWEDCTYKRLLLHSRNVHPDLQLTSLDKVIFPKSNKTKADYIGYLNEIGDFLLPFLDNRALTVIRYPHGSGGESFFQKNKPDYAPEFITTIRDDEHEHIICSDYSVLLWLANQLALEFHIPFQTADTTRPTEIVFDLDPPSRSEFPLAVRAANELHRLFEQLGLLSFPKLSGNKGIQIYIPISKNAFTYEETRLFTSFAASYCVSLFPDLFTTERLIKNRGGKLYIDYVQHAPGKTIICPYSTRGNQIGTVAAPLFWDEVHSDLAPSNFTMEAVIKRTKELGCPFESFFRQPQDKQIKAILDHLKEIDRSEN	*Bacillus Paralicheniformis*
Endonuclease II	636	207	23	82	MFCLETTIGEMFPDAECELVHDNPFELVIAVALSAQCTDALVNKVTKTLFKKYKKPEDYLAVPLEELQQDIKSIGLYRNKAKNIQKLCKMLLEEYGGEVPKDRDELVKLPGVGRKTANVVVSVAFGVPAIAVDTHVERVSKRLGICRWKDSVTEVEKTLMKKVPESEWSVTHHLIFFGRYHCKAQRPKCEECPLFLCAEAS	*Bifidobacterium adolescintis*

The DNA Polymerase I ORF was 2670 bp in length and encoded a polypeptide of 885 amino acids with 100 kDa predicted molecular mass. Sequence analysis of DNA ligase revealed that it has 74- 80% identity at the amino acid level with several homologous DNA ligases from the following bacteria; *Bacillus paralicheniformis* (80%), *Bacillus glycinifermentans* (77%), *Bacillus coagulans* (76%), and *Bacillus caldoxylosilyticus* (74%).

The DNA Ligase ORF was 1881 bp in length and encoded a polypeptide of 622 amino acids with 71 kDa predicted molecular mass. Sequence analysis of DNA ligase revealed that it has 79 - 98% identity at the amino acid level with several homologous DNA ligases; *Bacillus paralicheniformis* (98%), *Bacillus haynesii* (96%), *Bacillus swezeyi* (82%), and *Bacillus sonorensis* (79%).

The Endonuclease II ORF was 636 bp in length and encoded a polypeptide of 207 amino acids with 24 kDa predicted molecular mass. Sequence analysis of endonuclease II revealed that it has 82% identity at the amino acid level with two homologous endonuclease II; *Bafidobacterium adolescintis* (82%) and *Clorobaculum limnaeum* (82%).

2.3.4. Fosmid library construction

A fosmid library was constructed using a CopyControl pCC2FOS^{TM} vector which resulted in a library size of approximately 2.16 x10^{3} clones. The size of the metagenomic library is consistent with what other studies were able to construct, demonstrating the capability of cloning genomic DNA into fosmids as an alternative to capturing DNA of unculturable microorganisms present in the environment. In a study by Rondon et al. (2000) a fosmid library was constructed from soil sediments and it contained 3 624 fosmid clones with insert sizes ranging from 23 kb. According to Pang et al. (2008), fosmids are useful for constructing stable metagenomic libraries from complex environmental samples, however, during fosmid library construction a large amount of DNA can be lost due to handling and processing (Rondon et al. 2000).

2.3.4.1. Determination of library insert size

Agarose gel electrophoresis was conducted for *XbaI* digested fragments of Fosmid DNA. Average insert size of the fosmid was analysed from the size of the digested fragments with reference to the marker and was found to be 25kb. Gel picture showing the digested fragments is given below in **Figure 2.6.**

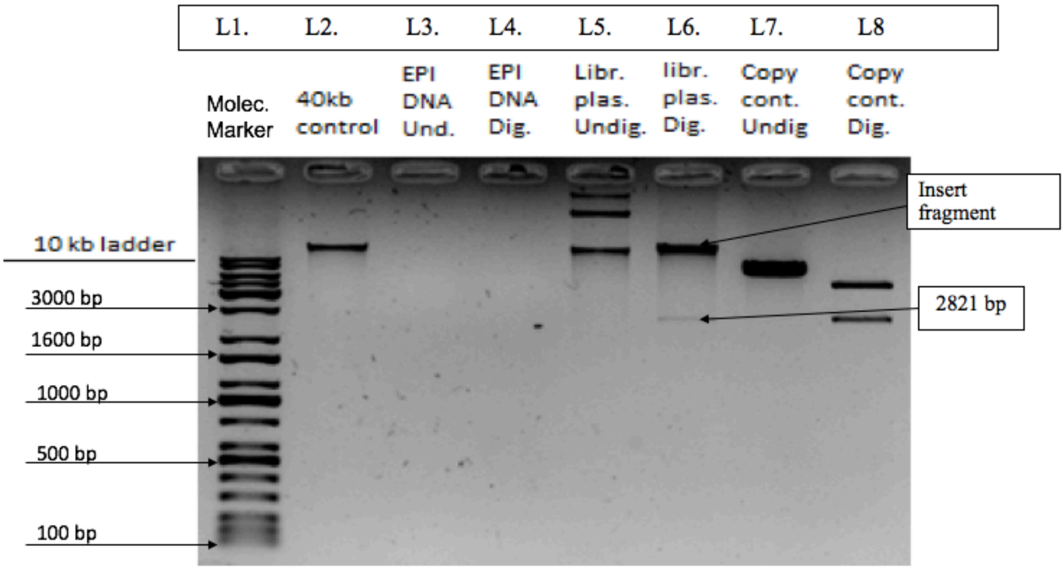


Figure 2.5. 0.8 % agarose gel, 100 voltage for 50 minutes. Lane 1is a 10 kb molecular marker, Lane 2 is a 40 kb Copy control DNA, Lane 3 is an undigested EPI cell Plasmid DNA, Lane 4 is *Xba*I digested EPI cell Plasmid, Lane 5 is undigested hot spring library plasmid, Lane 6 is XbaI digested hot spring library plasmid, Lane 7 is an undigested Copy control vector and Lane 8 is *Xba*I digested copy control vector respectively).

To confirm if the clones contained both the vector and the insert DNA, clones were randomly selected and have their plasmid DNA extracted and subsequently digested with *Xba*I enzyme as presented in **Figure 2.5** lane 6. The choice of the restriction enzymes was carefully selected ensuring that the enzyme cuts the vector backbone at two position; *Xba*I cuts the 8.1 kb vector at positions 413 and 3234. When digested, the plasmid showed two bands on the gel, with one aligning just below 3kb band of the 10kb ladder (KAPA biosystems), which corresponds with the difference between position 3234 and 413 bp of the fosmid vector. The other band was compared with the ladder and the 40kb control DNA (lane 2) and its size was estimated to range between 25kb and 40kb on the agarose gel.

To validate digestion, a negative control (L7) which comprises copy control vector was also digested with the same enzyme and resolved on the gel (L8) to see bands separation. Two bands were observed on the gel, the smallest band corresponded with the predetermined 2821 bp, while the biggest band was about 5 kb in size.

2.4. Conclusion

The aim of this chapter was to bio-prospect the metagenome obtained from the hot spring soil sediments and also to discover possible new genes that might code for DNA manipulating enzymes. Discovery of DNA manipulating enzymes is essential as they play a major role in the day to day molecular biology laboratory. The soil sediments are richer in microbial diversity than any other ecosystem and the majority of microorganisms from soil and hot springs are not well categorized due to the inability to culture them on the medium. The hot spring ecological unit is an attractive reservoir for the discovery of novel enzymes. Bio-prospecting a soil metagenome is the great platform to overcome the inability to culture microorganisms and search for novel enzymes. Due to the importance of finding thermostable enzymes, the hot spring was used in this study to search for DNA manipulating enzymes.

In this study, total metagenomic DNA from soil was successfully extracted but due to the low quality of extracted DNA following the use of SDS- based CTAB extraction method, it was necessary to modify the method. Gel extraction purification coupled with electroelution, resulted in good quality DNA for metagenomic library construction using the fosmid based system. The library was also validated to show that it contained the vector and the insert DNA inside. For the purpose of this project, the library was only constructed and stored for future function- based screening of the clones for traits of interest. Since the library is constructed from hot spring sediments, it will be suitable for screening for a wide variety of thermostable biorefinery enzymes and many enzymes of industrial importance. Sequence based screening of the metagenome was adopted to mine for DNA manipulating genes (DNA polymerase, DNA ligase and endonuclease II) from the metagenomic sequence data. Sequence analysis of primary structures showed similarity of ORFs discovered in this study with proteins from the NCBI-Blast-P database. The percentage identity shows that these enzymes are of known microorganisms, although they slightly differ with the organisms in NCBI. The complete ORFs were identified and sent to GeneScript for synthesis to make them ready for cloning and protein expression. In can be concluded that the soil sediments from hot springs can be successfully extracted and purified to recover a good quality DNA to be used for downstream applications. The same DNA can also be used to construct the fosmid library and also in high- throughput sequencing.

2.5. References

ALTSCHUL, S. F., MADDEN, T.L., SCHAFFER, A.A., ZHANG, J., ZHANG, Z., MILLER, W. & LIPMAN, D.J. 1997. Gapped BLAST and PSI- BLAST: a new generation of protein database search programs. *Nucleic Acids Res*. 25; 17: 3389- 402.

BURGMANN, H., PESARO, M., WIDMER, F. & ZEYER, J. 2001. A strategy for optimizing for optimizing quality and quantity of DNA extracted from soil. *J. Microbiol. Methods*. 52: 389-393.

COLORA, S. & ROLF, D. 2011. Metagenomic analysis: Past and future trends. *Applied and Environmental Microbiology*. 77; 4:1153- 1161.

COURTOIS, S., FROSTEGÅRD, A., GÖRANSSON, P., DEPRET, G., JEANNIN, P. & SIMONET, P. 2001. Quantification of bacterial subgroups in soil: comparison of DNA extracted directly from soil or from cells previously released by density gradient centrifugation. *Environmental Microbiology*. 3; 7:431–439.

DANIEL, R. 2005. The metagenomics of the soil. *Nature Reviews Microbiology*. 3:470- 478.

De CASTRO, A.P., QUIRINO, B.F., ALLEN, H., WILLIAMSON, L.L., HANDELSMAN, J. & KRUGER, R.H. 2001. Construction and validation of two metagenomic DNA libraries from Cerrado soil with high clay content. *Biology Letters*. 33: 2169- 2175.

DIAS, R.S., SILVA, L.C.F., ELLER, M.R., OLIVERIA, V.M., De PAULA, S.O.S. & SILVA, C.C. 2014. Metagenomics: Library Construction and Screening. *Methods*. 45-65. *Environ Microbiol*. 6; 9: 879-86.

GABOR, E.M., de VRIES, J.E. & JANSSEN, D. B. 2003. Efficient recovery of environmental DNA for expression cloning by indirect extraction methods. *FEMS Microbiology Ecology* 44: 153- 163.

HALL, T.A. 1999. BioEdit: a user friendly biological sequence alignment editor and analysis. *Nucleic Acids Symposium Series*. 41: 95-98.

KALLMEYER, J. & SMITH, C. 2009. An improved electroelution method for separation of DNA from humic substances in marine sediment DNA extracts. *FEMS Microbiol Ecol*. 69: 125- 131.

KAUFFMANN, I.M., SCHMITT, J. & SCHMID, R.D. 2004. DNA isolation for cloning in different hosts. *Applied Microbiol Biotechnol*. 64; 5: 665- 70.

MARDANOV, A.V., GUMEROV, V.M., BELETSKY, A.V., PEREVALOVA, A. A., KARPOV, G.A., BONCH-OSMOLOVSKAYA, E.A., RAVIN, N.A. & SKRYABIN, K.G. 2011. Uncultured archaea dominate in the thermal groundwater of Uzon Caldera, Kamchatka. *Extremophiles* 15:365–372.

OLIVIER, J., VENTER, J.S. & JONKER, C.Z. 2011. Thermal and chemical characteristics of hot water spring in northern part of Limpopo province, South Africa. *Department of Environmental Sciences. Unisa*. 37; 4:427- 436.

PANG, M., ABDULLAH, N., LEE, C. W. & NG, C. C. 2008. Isolation of high molecular weight DNA from forest topsoil for metagenomic analysis: *Asia Pacific Journal of Molecular Biology and Biotechnology*. 16:35-41.

RONDON, M. R., AUGUST, P. R., BETTERMANN, A. D., BRADY, S. F., GROSSMAN, T. H., LILES, M. R., LOIACONO, K. A., LYNCH, B. A., MACNEIL, I. A. & MINOR, C. 2000. Cloning the soil metagenome: a strategy for accessing the genetic and functional diversity of uncultured microorganisms: *Applied and Environmental Microbiology*. 66:2541-2547.

SAGOVA- MARECKOVA, M., CERMAK, L., NOVOTNA, J., PLHACKOVA, K., FORSTOVA, J. & KOPECKY, J. 2008. Innovative Methods for Soil DNA Purification Tested in Soils with Widely Differing Characteristics. *Applied and Environmental Microbiology*. 74; 9: 2902-2907.

SAMBROOK, J. & RUSSELL, R.W. 2001. Molecular cloning: A laboratory manual, 3rd ed. *Cold spring harbor laboratory press, cold spring harbor, N.Y.* 2; 8:658-662.

SCHLOSS, D. P. & HANDELSMAN, J. 2003. Biotechnological prospects from metagenomics. *Current opinions in Biotechnology*.14:303-310.

SEBASTIAN, R., KIM, J.Y., KIM, T.H. & LEE, K.T. 2013. Metagenomics: a promising approach to assess enzymes biocatalyst for biofuel production. *Asian J. Biotechnol.* 5: 33–50.

SOFIA, U.M., TORIL, E.G., ALEJANDRA, G.M., BAZAN, A.A. & DONATI, E.R. 2014. Archaeal and Bacterial diversity in five different hydrothermal ponds in Copaline region in Argentina. *Syst App Microbiol*. 37: 429- 441.

TEBBE, C. & VAHJEN, W. 1993. Interference of humic acids and DNA extracted directly from soil in detection and transformation of recombinant DNA from bacteria and a yeast. Applied and environmental microbiology. 59. 2657-65.

XIE, W., WANG, F., GUO, L., CHEN, Z., SIEVERT, S.M., MENG, J. 2011. Comparative metagenome of microbial communities inhabiting deep- sea hydrothermal vents chimneys with contrasting chemistries. *ISME J*. 5, 414- 426.

ZHOU, J., BRUNS, M.A. & TIEDJE, J.M. 1996. DNA recovery from soils of diverse composition. *Appl Environ Microbiol*. 62; 2:316-22.

https://www.cephamls.com/BAC-Fosmid-Library-Construction

Chapter 3: Expression, purification and functional analysis of DNA manipulating enzymes

Table of Contents

Chapter 3

Abstract

DNA-manipulating enzymes are proteins that catalyse the manipulation of DNA in molecular biotechnology or genetic engineering. They are usually grouped into a four broad classes depending on their reaction. They are one of the commonly used enzymes in molecular biology laboratories. In this study we synthesized, cloned, expressed, and purified DNA-manipulating enzymes; DNA polymerase, DNA ligase and endonuclease II. DNA fragments for the genes were acquired from hot a spring metagenomic data, and were cloned into pET- 30a(+) which resulted in pET-30a(+)-*hs-dp-vut*, pET-30a(+)-*hs-lig-vut* and pET-30a(+)-*hs-en-vut* constructs. Analysis of nucleotide sequence revealed that recombinant proteins; DNA polymerase, DNA ligase, endonuclease II have molecular mass of 99901.4, 71017.6 and 23608 Da respectively. The protein expression was carried out under the control of the T7*lac* promoter in *Escherichia coli* BL21 (DE3) then induced by 0.5 mM isopropylthio-β-D-galactoside (IPTG) at 16 ºC for 16 hours. The expressed proteins were found almost entirely in the insoluble form in cell lysate. The inclusion bodies were solubilized with 8M urea and the recombinant protein was purified by Ni-NTA column using 500mM imidazole while the refolding was performed in 14 kDa cut off dialysis membrane into 1X PBS, 0. 5 M L- arginine, pH 7.4. Analysis of the SDS/PAGE gel for DNA polymerase, DNA ligase and endonuclease II has shown that the purity of the proteins were 85- 95 %.

Key words: DNA-manipulating enzymes, metagenomic data, cloning, expression, purification

3.1. Introduction

DNA-manipulating enzymes are proteins that are used to modify DNA molecule in protein engineering or molecular biotechnology. They can be grouped into four broad classes depending on the type of the reaction they catalyze. They are commonly arranged as nucleases, ligases, polymerases and modifying enzymes.

DNA polymerase is an enzyme responsible for the replication making or copying DNA molecules from deoxyribonucleotides (dNTPs), the building blocks of DNA (Cotterill and Kearsey 2009). They usually function in pairs to synthesise two identical DNA strands from a single DNA molecule (Garcia- Diaz and Bebenek 2007). It is responsible for polymerization during polymerase chain reactions (PCR) and it has revolutionised molecular biology with their ability to amplify small amounts of DNA invitro (Kaguni 2018). Traditionally, the parental thermostable DNA polymerases were isolated from a heat stable bacterium called *Thermus aquatics*. But recently, DNA polymerases have been produced by recombinant DNA technology to improve their specificity and production (Drouin 2007).

DNA ligases are present in almost all living organisms, and it is required for survival functions and maintaining the integrity of the DNA backbone structure (Al- Manasra and Al- Razen 2012). They are housekeeping enzymes that are essential for survival roles and cellular processes linked to breaks filling of 5′-phosphate and 3′-hydroxyl termini at single-strand breaks in double-stranded DNA, or at two fragments containing either complementary single strand or blunt ends, which are essential roles in DNA replication, recombination, and repair (Rossi et al. 1997; Seo et al. 2007; Pascal 2008). This process allows the joining of similar and foreign DNA sequences (Pascal 2008). DNA ligases are one of the crucial discoveries in molecular biology and biotechnology due to the role it plays in the molecular cloning of important genes (Al- Manasra and Al- Razen 2012). DNA ligases can also be used in projects involving gene synthesis (Bang and Church 2008). They are essential in most next-generation sequencing (NGS) platforms, either during sample preparation or for adapter ligation (e.g. Ion Torrent sequencing), or for the sequencing reaction itself (e.g. SOLiD sequencing). Importantly, it is the ligation of cohesive or blunt-ended dsDNA fragments that are most commonly needed in molecular biology protocols (Lohman et al. 2011).

Type II restriction endonucleases are indispensable tools in creating recombinant DNA molecules (Russel 2001). Among the 232 different specificities, nearly half of the restriction-modification (R-M) systems have been cloned and expressed (Pingould and Jeltsch 2001). They cut the phosphodiester linkages of a double helix DNA. To achieve blunt ends, it cleaves the centre of a strand, or it can cut DNA at a staged position to yield the overhangs called sticky ends (Ninfa et al. 2010). Restriction enzymes are used in molecular biology laboratory to aid in the insertion of genes into plasmid vectors in the process of gene cloning and for protein production projects (Zhang et al. 2005). This enables flexible insertion of gene fragments into the vector of choice.

Since the isolation of these enzymes from their native host is often costly and results in very small yields upon purifications, overexpression of widely used restriction endonucleases in recombinant *E. coli* cells is advantageous for the high yield recovery of these enzymes (Som et al. 1987; Hsieh et al. 2000; Gholizabeh et al. 2010). The parallel use affirnity tags with recombinant DNA techniques, allows the facile modification of proteins of interest leading to efficient identification, production, and isolation from host system (Structural Genomics Consortium et al. 2008).

The most traditionally used expression systems are based on pET vectors which facilitate expression of a target gene under the control of the lac operator and T7 RNA polymerase promoter (Structural Genomics Consortium et al. 2008). When the conditions are repressive, T7 RNA polymerase is not produced, and transcription of the target gene is very insignificant. After induction, most of the cellular protein synthesis machinery is devoted to the production of the target protein (Studier et al. 1990). When using T7 systems, protein expression can be induced either by manipulating the carbon sources during E. coli growth or with the chemical inducer isopropyl-β-d-thiogalactoside (IPTG) (Studier 2005).

Purification of recombinant proteins is mostly accomplished using a purification tag which can be located either at the N- or C-terminus of a protein of interest. Recombinant proteins produced in *E. coli* systems can also be purified using conventional chromatographic methods based ion exchange, on size exclusion, and hydrophobic interaction that separate proteins according to size, charge and hydrophobicity, respectively (Rosano and Ceccarelli 2014).

As a chromatographic procedure, IMAC has the benefits of having robust, mild elution conditions, specific binding and the ability to regulate selectivity by using chromatography buffers with low imidazole concentrations. There is a wide array of common resins with slightly different binding capacities and binding strengths, but all tolerate harsh cleaning procedures (Structural Genomics Consortium et al. 2008). Most purification steps can be integrated by high-performance liquid chromatography; the most commonly used devices are the ÄKTA systems from GE Healthcare (Structural Genomics Consortium et al. 2008). After protein purification, samples are visualized by SDS-PAGE. If the protein is stained with a dye such as Coomassie brilliant blue, the intensity of the bands will usually be proportional to the amount of protein. This allows the purity of the sample to be estimated and whether the purified protein is of the expected size (Bradford 1976).

This chapter is therefore dedicated to cloning, expression and purification of recombinant putative DNA polymerase*,* DNA ligase and endonuclease II proteins transformed into *E.coli* BL21 (DE3).

3.2. Materials and Methods

3.2.1. Gene synthesis and cloning

The putative DNA manipulating enzymes genes (a 1881 base pairs DNA ligase gene, 2670 base pairs DNA polymerase I and 636 base pairs endonuclease II) sourced from the in silico mining of metagenome sequence data were then sent to be synthesized and cloned into pET-30a (+) by the services of GenScript (Piscataway, USA). The constructs were designed *in silico* using SnapGene and the recombinant plasmids were designated pET-30a(+)-*hs-dp-vut*, pET-30a(+)-*hs-lig-vut* and pET-30a(+)-*hs-en-vut*; for DNA polymerase I, DNA ligase and endonuclease II respectively. The genes were designed such that they all having 6 His-tag on the C- terminus. The genes were individually cloned onto *Xba*I/*Hind*III linearized pET-30a (+) (Novagen, Germany).

3.2.2. Transformation.

Upon arrival the constructs were individually transformed into *E. coli* BL21 (DE3) competent cells. The protocol by Sambrook et al. (1989) with few modifications was used for transformation of the constructs. The appropriate volumes of competent cells, as well as an extra volumes for test plasmid positive control were removed from the freezer to determine transformation efficiency. Immediately, tubes containing competent cells were placed on ice. The cells were allowed to thaw on ice for 2–5 min. For each transformation, 1.5 ml centrifuge tube was prechilled on ice to control temperature variations. Twenty microliters of thawed competent cells were transferred into each pre-chilled centrifuge tube. One microliters (1 – 10 ng) purified recombinant plasmid was added to BL 21 expression host competent cells. The mixture was stirred gently and returned to ice to incubate for 5 minutes. The tubes were then heated for exactly 30 seconds in a 42°C water bath without shaking. Subsequently, the tubes were placed on ice for 2 min. Eighty microliters of room temperature super optimal broth with catabolite repression (SOC medium) was added to each tube, while the tubes were placed on ice during handling. The tubes were incubated at 37°C with shaking at 250 rpm for 60 min before plating. Twenty microliters of transformation mixture was spread on LB medium supplemented with kanamycin (50mg/µl). The plates were then incubated overnight at 37°C.

The following day, colonies were randomly picked from the plates to confirm transformation. Randomly picked colonies were grown overnight in 50 ml LB broth containing kanamycin at the final concentration of (50mg/µl). The culture was then span down at maximum speed, while the supernatant was discarded and the pellet retained for plasmid extraction using the GeneJET Miniprep Plasmid Kit, according to the user manual. Once the plasmid was obtained, the quantification of the plasmid was carried out using Nanodrop One C. Thereafter, 2µl of 100ng/µl plasmid DNA was digested with *Xba*I/*Sma*I restriction enzymes. This digestion process was carried out as per Fast Digest Thermo Scientific user protocol. Five microliters of the digestion product was ran on gel electrophoresis to determine the presence of the transformed construct to validate transformation process.

3.2.3. Expression and Purification of putative DNA polymerase, DNA ligase and Endonuclease II.

The insoluble fractions (inclusion bodies) of cell protein extract were analyzed on SDS-PAGE to determine the expression levels of DNA ligase protein. To achieve this, glycerol stock of *E.coli* BL21 (DE3) transformed with target construct (pET-30a(+)-*hs-dp-vut*, pET-30a(+)-*hs-dl-vut* and pET-30a(+)-*hs-en-vut*) was thawed and inoculated in 4 ml LB medium containing 50 µl/ml kanamycin and incubated overnight at 37 °C with shaking at 200rpm. Four milliliters of overnight cultured cells were inoculated into 500ml LB broth containing 50 µg/ml kanamycin and were incubated at 37 °C while shaking at 200 rpm.

When OD $_{600}$ value of the culture had reached 1.2, isopropylthio-β-galactoside (IPTG) (Promega, USA) at the final concentration of 0.5 mM was added to induce protein expression for 16 hrs. at 16 °C with shaking at 200rpm. The cells were harvested at 5000 rpm, 4 °C for 30 minutes, and were then resuspended in lysis buffer (50 mM Tris- HCl, 150 mM NaCl, pH 8.0). The cells were cracked open by sonication for 30 minutes, using Bransonic Ultrasonic Bath (model 1800) followed by centrifugation at 5000 rpm at 4 °C for 30 minutes. Because membrane proteins were targeted, the supernatant was discarded and the inclusion bodies were collected for further processing. The collected inclusion bodies were then resuspended in denature buffer (50 mM Tris- HCl, 8 M Urea, 150mM NaCl, pH 8.0) with sonication for 30 minutes, followed by centrifugation at 5000 rpm, 4 °C for 30 minutes. Subsequently, the supernatant was loaded onto Ni- NTA HisTrap FF column chromatography using AKTA start protein purification system (GE Healthcare Life Sciences) for target protein binding, followed by washing with buffer (50mM Tris- HCl, 8 M Urea, 150 mM NaCl, 1% Triton X-114, pH 8.0). The target protein was then eluted with a stepwise gradient of imidazole (20mM imidazole, 50mM imidazole; 8M urea and 500mM imidazole; 8M urea). According to conductivity curve and UV absorption peaks on the AKA start purification system, different elution fractions were collected.

The eluted fractions were pooled and refolded by dialysis into 1X PBS, 0. 5 M L- arginine, pH 7.4. The refolding was performed in 14 kDa cut off dialysis membrane for 2 hours and then the buffer was replaced with a fresh one for additional 16 hours. Once refolding was completed and His tag removed by thrombin, the sample were centrifuged at 13 000rpm for 30 minutes and filtered through a 0. 22µm filter and the target DNA ligase was in aliquots of 0.3 ml/tube. The proteins were separated by SDS-PAGE (4% acrylamide-stacking and 12% acrylamide-separating) and stained in Coomassie staining solution (10% acetic acid, 50% methanol, 0.25% Coomassie Brilliant Blue R 250) for an hour and destained overnight with destaining solution (10% acetic acid, 50% methanol). A sodium dodecyl sulphate-polyacrylamide gel electrophoresis (SDS-PAGE) was performed according to Laemmli (1970) to determine the subunit molecular mass and purity of the protein.

3.3. Results and Discussion

3.3.1. Gene synthesis and Cloning

Gene synthesis is a technique or an approach in synthetic biology that is employed to make synthetic gene fragments in the laboratory. The method is based on the process known as solid-phase DNA fragment synthesis, and it differs from polymerase chain reaction (PCR) and molecular cloning in that it does not have to begin with pre-existing DNA sequences. This technique makes it possible to produce a double-stranded DNA molecule with no apparent limits with regards to nucleotide sequence or size of the fragment. It has successfully been used to generate functional bacterial or yeast chromosomes containing approximately one million base pairs. Gene synthesis involves a combination of both molecular biology and organic chemistry techniques to produce a gene, and entire genes may be synthesised "de novo," without the need for template DNA (Kosuri and Church 2014).

It is an essential tool in many areas of recombinant DNA technology including molecular engineering, gene therapy, heterologous gene expression, and vaccine development. It is also a powerful engineering tool for creating and designing new DNA sequences and protein functions. The approach can be more economical than traditional cloning and mutagenesis procedures (Kosuri and Church 2014). In this study gene synthesises and cloning was performed at GeneScript.

3.3.1.1. DNA polymerase

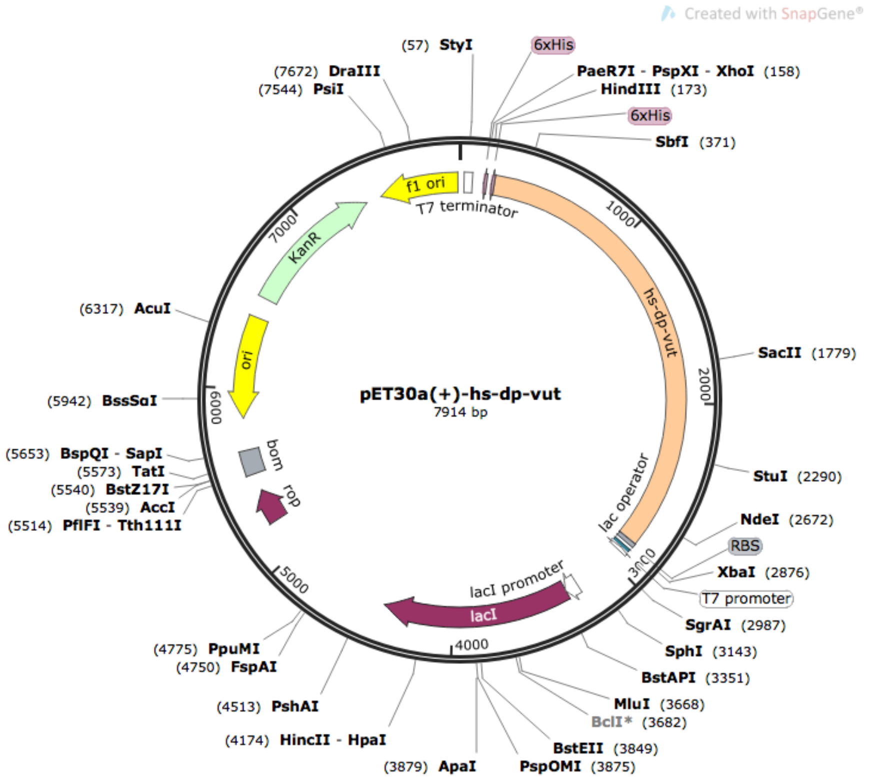


Figure 3.1. A map of a 7914 bp pET 30a (+)-*hs-dp-vut* construct after cloning polymerase gene into pET 30a (+).

To facilitate cloning of a 2670 bp open reading frame (ORF) DNA polymerase gene into pET-30a (+), the insert fragment was designed in such a way that it possesses *Nde*I and *Hind*III restriction sites at the N-terminus and C- terminus respectively. The 6-histidine tag was incorporated at the C- terminus of the gene just before the stop codon TAATGA (Tandem termination codon). The DNA sequence of the fragment to be cloned and expressed was as below.

5’-*Nde*I- ATG- DNA polymerase gene – 6 Histag- TAATGA -*Hind*III-‘3.

The amino acid sequence was retrieved following BLASTx algorithm from NCBI using DNA polymerase open reading frame. The size and molecular weight of the protein were 885 amino acids and 99901.2 Da respectively. The resulting DNA fragment was then chemically synthesized and cloned to pET-30a (+) by services of GenScript (Piscataway, USA) and a 7914 bp construct (**figure 3.1**) was then built. See appendix B3 and B4 for DNA polymerase nucleotide and ammino acid sequence.

3.3.1.2. DNA ligase

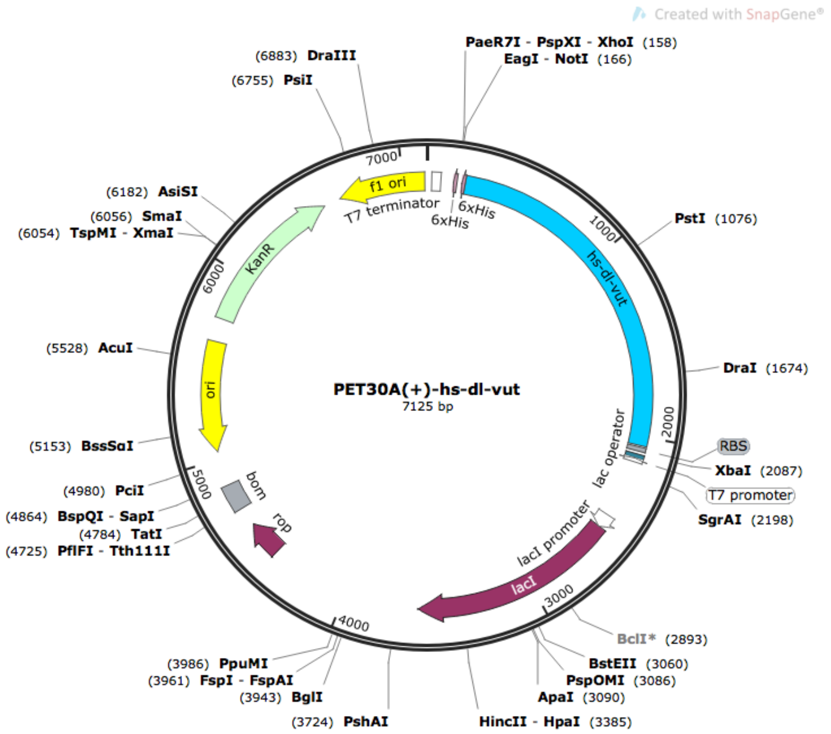


Figure 3.2. A map of a 7125 bp pET 30a(+)-*hs-dl-vut* construct after cloning DNA ligase gene into pET 30a (+).

A DNA ligase gene (1881bp) to be cloned was designed such that it possesses *Nde*I and *Hind*III restriction sites at the N-terminus and C- terminus, respectively. To help facilitate affinity purification of the recombinant DNA ligase enzyme, the 6- Histidine tag was also incorporated at the C- terminus before the TAATGA (Tandem termination codon), the sequence is as outlined below.

5'-*Nde*I- ATG- DNA ligase gene – 6 Histag- TAATGA -*Hind*III-'3.

The cloning strategy adopted here, ensures that the predicted 622 amino acids and 71017.6 Da DNA ligase enzymes is produced under the control of a T7 promoter as regulated by an IPTG inducible operator sequence. The complete nucleotide and amino acid sequences (see appendix B5 and B6) were retrieved following BLASTx search algorithm from NCBI using open reading frame for DNA ligase gene. The resulting DNA fragment was then sent to be synthesized and cloned into pET-30a (+) at GenScript (Piscataway, USA). The complete construct is shown below in **figure 3.2**.

3.3.1.3. Endonuclease II

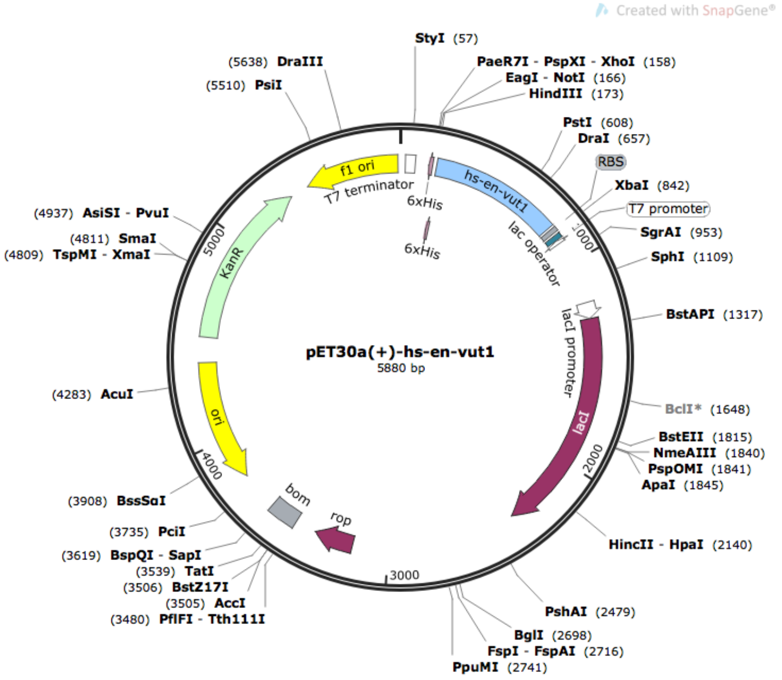


Figure 3.3. A map of a 5880 bp pET 30a(+)-*hs-en-vut* construct after cloning DNA endonuclease II gene into pET 30a (+) using SnapGene software for in silico cloning.

To clone and express a 636 bp DNA ligase gene into pET-30a (+), the insert fragment was designed such that it possesses 6- Histidine tag at the C- terminus to help facilitate affinity purification of the recombinant endonuclease II enzyme, the 6- Histidine tag was also incorporated at the C- terminus before the TAATGA (Tandem termination codon). The orientation of DNA fragment is as follows. See appendix B7 and B8 for DNA and protein sequence.

5'-*Nde*I- ATG- endonuclease II – 6 Histag- TAATGA -*Hind*III-'3.

The presence of hexahistidine tag allows recombinant protein to be purified using a nickel-chelating resin. As stated by Esposito and Chatterjee (2006), many strategies have been developed over the years that promote solubility of the targeted recombinant proteins and these strategies employ the addition of fusion tags to aid with purification. Some expression vector systems allow the expression of the protein of interest such that purification and solubility is improved. Sometimes 6- histidine tag is combined with solubility tags for both affinity and solubility function (Esposito & Chatterjee 2006). The cloning strategy adopted here, ensures that the predicted 207 amino acids and 23608.4 Da endonuclease II enzyme is produced under the control of a T7 promoter as regulated by an IPTG inducible operator sequence. The amino acid sequence in was also retrieved using BLASTx search algorithm from NCBI using translated nucleotide query. The size and molecular weight of the protein were predicted in silico as described above. The resulting DNA fragment was then chemically synthesized and cloned into pET-30a (+) by services of GenScript (Piscataway, USA). The built construct is shown **Figure 3. 3**.

3.3.2. Transformation and quality control

After transformation of the target constructs into *E.coli* BL21 (DE3), the clones were grown overnight and randomly selected and have their plasmid extracted and verified by restriction digestion of *Xba*I and *Sma*I. The digestion product for pET-30a(+)-*hs-dp-vut*, pET-30a(+)-*hs-dl-vut* and pET-30a(+)-*hs-en-vut* were analyzed on 1.2 % agarose gel in **Figures 3.4**, **3.5**, **3.6** respectively. The two restriction enzymes that were selected to verify the success of transformation were chosen due to the fact that they are single cutters and they cut a 5422 bp pET 30a (+) backbone at positions 384 bp and 4353 bp respectively. So the presence of 4353-384 = 3969 bp in the digestion product, would suggest that the recombinant plasmid was successfully transformed into *E.coli*. In **Figure 3.4** the pET-30a (+)-*hs-dp-vut* recombinant plasmid (Lane 1) was double digested with *Xba*I/*Sma*I and the digestion product of three bands was observed. The first band corresponded with 4 kb marker, the second band aligned with about 2.5kb marker while the third band was positioned at about 1.2 kb. The presence of about 4kb band together with the sum total of all the bands in Lane 2 suggests that the extracted plasmid was actually the recombinant pET-30a (+)-*hs-dp-vut* plasmid.

In **Figure 3.5**, the pET-30a (+)-*hs-dl-vut* recombinant plasmid in lane 1 was also double digested with XbaI/SmaI and the digestion product was two bands. The 4kb band was observed which corresponded with a calculated 3969 bp product. The second band aligned with about 2.5kb marker and the total sum of the two bands is well over the molecular size of the 5422 bp pET 30a (+) and about the size of the construct of about 7.3 kb . The presence of about 4kb band also suggests that the extracted plasmid was actually the pET-30a (+)-*hs-dl-vut* construct.

In **Figure 3.6**, the pET-30a (+)-*hs-en-vut* construct (lane 1) was double digested with *Xba*I/*Sma*I and the digestion product of two bands was also observed. The first band corresponded with 4 kb marker, while the second band aligned with about 2 kb marker. The presence of about 4kb band together with the sum total of the bands in Lane 2 suggests that the extracted plasmid was actually the recombinant pET-30a (+)-*hs-en-vut* plasmid.

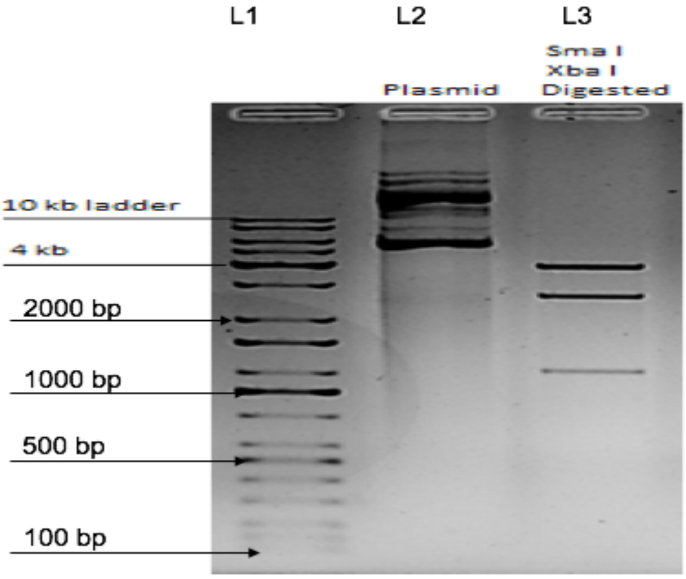


Figure 3.4. M: 10kb molecular marker, Lane 1: Plasmid DNA and Lane 2: *Xba*I/*Sma*I digested pET-30a (+)-*hs-dp-vut* construct. Electrophoresis was run on 1.2 % agarose at 100V for 60 minutes.

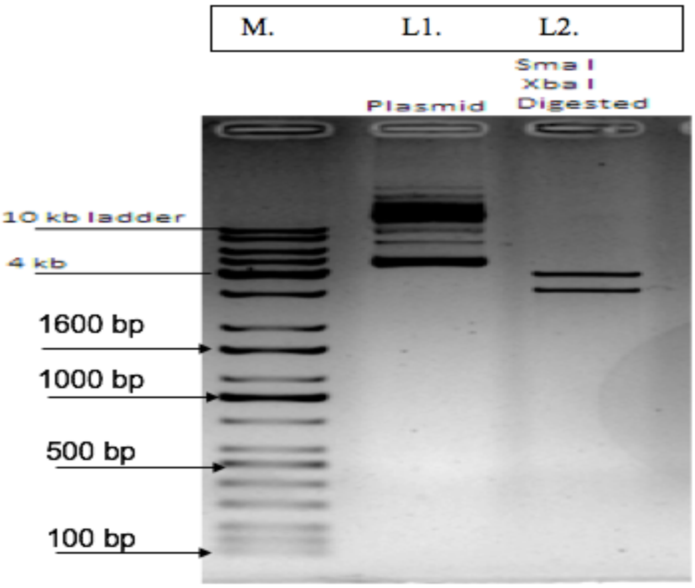


Figure 3.5. M: 10kb molecular marker, Lane 1: Plasmid DNA and Lane 2: *Xba*I/*Sma*I digested pET-30a (+)-*hs-dl-vut* construct. Electrophoresis was run on 1.2 % agarose at 100V for 60 minutes.

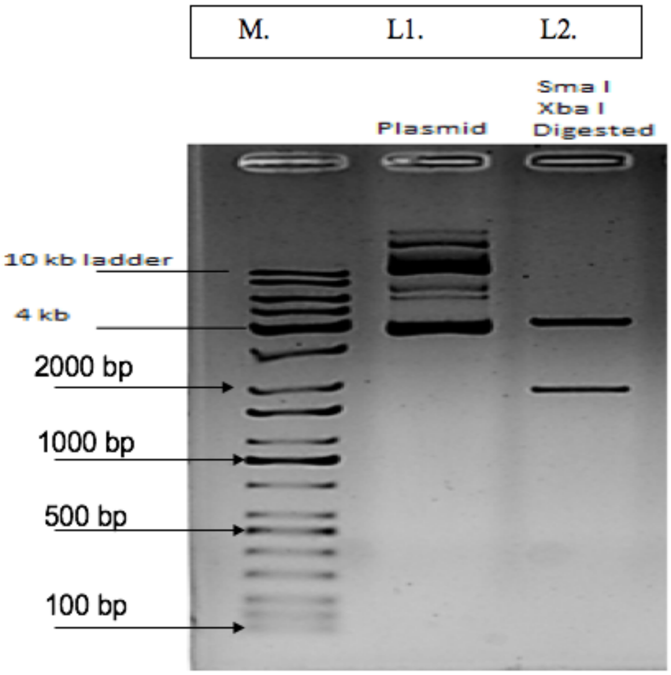


Figure 3.6. M: 10kb molecular marker, Lane 1: Plasmid DNA and Lane 2: *Xba*I/*Sma*I digested pET-30a (+)-*hs-dp-vut* construct. Electrophoresis was run on 1.2 % agarose at 100V for 60 minutes.

3.4.3. Expression and purification of DNA polymerase, DNA ligase and endonuclease II.

In this study, we cloned three full- length sequence of DNA polymerase, DNA ligase and endonuclease II protein, from hot spring metagenome and expressed it as a His-tagged fusion protein in *E. coli* cells. However, we found that the protein were expressed in an insoluble form in inclusion bodies. Therefore, we had to purify the recombinant proteins in a denatured form and then renatured them again. According to Carrio and Villaverde (2005), many factors contribute to inclusion bodies accumulation in the cell, the use of high temperature during protein expression, high inducer concentration and expression under robust promoter systems often results in expression of the desired protein at a high translational rate. These factors exhaust bacterial protein quality control system and the partially folded and misfolded protein molecules aggregate to form inclusion bodies (Carrio and Villaverde 2005).

Although the protein expressed in the form of inclusion bodies is mostly considered undesirable, their formation can be advantageous, as their isolation from cell homogenate is a convenient and effective way of purifying the protein of interest. The inclusion bodies formation in the cell offers several advantages. It allows the expression of a very high level of protein, in some cases more than 30% of the cellular proteins, it facilitates easy extraction of aggregated proteins from the cells because of the differences in their size and density as compared with cellular contaminants (Palmer and Wingfield 2004). Also, when the protein is expressed as inclusion bodies, it has lower degradation as compared to proteins expressed soluble. It renders the protein resistant to proteolytic attack by cellular proteases, and lesser contaminants which help in decreasing the number of purification steps during protein purification. Thus, because of the advantages mentioned above, recombinant proteins expressed as inclusion bodies have been widely utilized for the commercial production of proteins (Singh and Panda 2005).

3.3.3.1. Expression and Purification of putative DNA polymerase.

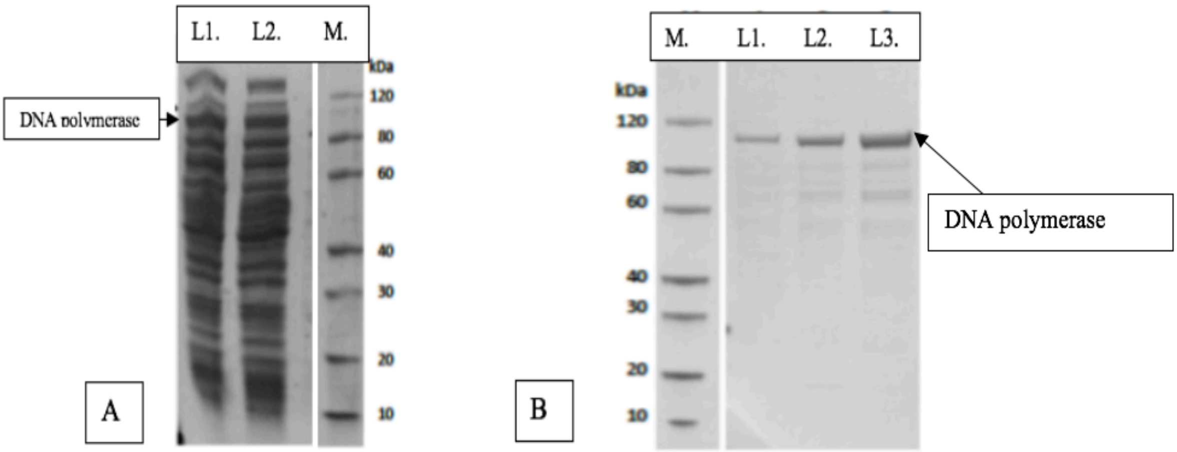


Figure 3.7. The SDS-PAGE analysis showing different protein fractions during DNA polymerase purification with Ni column. (**A**) M: Protein marker, Lane 1: supernatant after centrifugation, Lane 2: flow through. (**B**) M: Protein marker, Lane 1: A wash with 20 mM imidazole, Lane 2: Elution with 50 mM imidazole, 8M urea and Lane 3: Elution with 500 mM imidazole; 8M urea.

3.3.3.2. Expression and Purification of putative DNA Ligase.

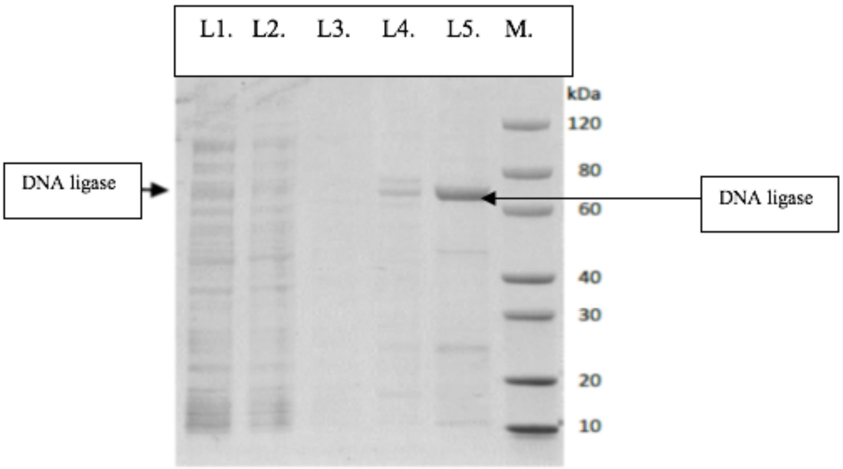


Figure 3.8. The SDS-PAGE analysis showing different protein fractions during DNA ligase purification with Ni column. M: Protein marker, Lane 1: supernatant after centrifugation, Lane 2: flow through, Lane 3: A wash with 20 mM imidazole buffer, Lane 4: Elution with 50 mM imidazole, 8M urea and Lane 5: Elution with 500 mM imidazole, 8M urea.

3.3.3.3. Expression and Purification of putative endonuclease II

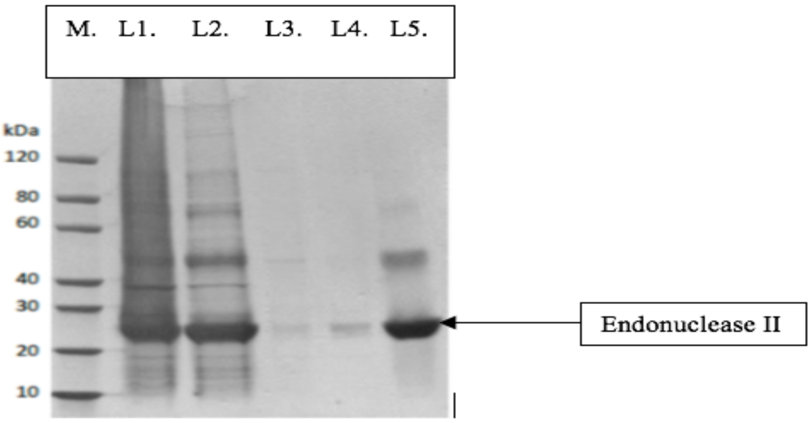


Figure 3.9. The SDS-PAGE analysis showing different protein fractions during endonuclease II purification with Ni column. M: Protein marker, Lane 1: supernatant after centrifugation, Lane 2: flow through, Lane 3: A wash with 20 mM imidazole buffer, Lane 4: Elution with 50 mM imidazole; 8M urea and Lane 5: Elution with 500 mM imidazole; 8M urea.

After the solubilization of the inclusion bodies with denature buffer, the supernatant that was recovered subsequent to centrifugation, was loaded onto SDS-PAGE to visualize the total protein profile (lane 1 of **Figure 3.7** A, **Figure 3.8** and **Figure 3.9**). In **Figure 3.7** lane 1, the recombinant protein DNA polymerase was identified by overexpression of one of the protein bands which migrated closer to the predicted molecular weight of 99901.2 Da as predicted in silico. The molecular weight of the bands was predicted against a 120 kDa protein marker (New England, Biolabs). In **Figure 3.8** lane 1, the putative DNA ligase protein was identified as an overexpressed protein band located in the area of predicted 71017.6 Da as estimated using 120 kDa protein marker. Finally, in **Figure 3.9** lane 1, the putative endonuclease II protein was observed as an overexpressed protein band situated in the area of 25 kDa which corresponds with the in silico predicted molecular weight of 23608.4 Da. The three proteins were therefore successfully expressed under the control of T7 at the IPTG concentration of 0.5 mM. IPTG acts upon the lac promoter on the T7 RNA polymerase gene present in the genomic DNA of BL21 (DE3). This T7 RNA polymerase produced acts on T7 promoter of pET30a (+) and it initiates the synthesis of our recombinant proteins. The next step was to purify the extracted proteins using AKTA start protein purification system.

The supernatant was also loaded onto AKTA start purification system, where His Trap FF 5ml column was used to bind recombinant protein to the Ni- NTA matrix which selectively binds the target, while everything else flows through. For quality purposes, the proteins that could not bind to the column (flow through), were collected and viewed on SDS- PAGE. The protein profile of the flow through is shown in lane 2 of **Figure 3.7** (**A**), **Figure 3.8** and **Figure 3.9**. In the flow through, a protein profile similar to that of the supernatant was observed however it was at a lesser concentration than that of the supernatant. The column was then washed with the wash buffer containing 20 mM imidazole and the flow through fractions were collected and also visualized on SDS-PAGE as shown in **Figure 3.7** lane 1, **Figure 3.8** lane 3 and **Figure 3.9** lane 3. The protein profiles of the fractions after the washing step revealed residual proteins that did not bind to the column during binding step, also suggesting that the recombinant proteins of interest were still bound to the column. Therefore, washing the column with the binding buffer containing 20 mM imidazole resulted in the elution of many contaminants but not of the tagged proteins.

The bound recombinant proteins were eluted with two buffer concentration of imidazole; 50 mM and 500mM respectively. The elution fractions were collected and loaded onto SDS PAGE gel as presented by lane 2 and lane 3 in **Figure 3.7 (B)**, lane 4 and lane 5 in **Figure 3.8** and **Figure 3.9**. A significant elution of recombinant proteins was observed with 50 mM imidazole however using 500 mM imidazole resulted in the release of a significant additional amount of recombinant protein, thus suggesting that a rather large concentration of imidazole is necessary to achieve a complete elution of expressed recombinant Histag recombinant protein. The eluted recombinant proteins still contained a few contaminating proteins as observed by the presence of other bands as depicted on SDS-PAGE gels **Figure 3.7**, **3.8** and **3.9**. By analysis of the SDS-PAGE gel, the purity of DNA polymerase, DNA ligase and endonuclease III was estimated to be 85, 95, and 80% respectively. Many researchers have studied the expression and optimization of DNA manipulating enzymes in bacteria, however culture based techniques were used to recover the genes and also, some of data is not readily available since they are either patented or not released by the companies producing them.

In a study by Desai and Pfaffle (1995), pUC18 plasmid was used for the cloning and expression of DNA polymerase. The protein expression was under 0.5 mM IPTG induction for 16-20 h. High levels of enzyme production were reported although their system expressed recombinant DNA polymerase even before the addition of the inducer (Desai and Pfaffle 1995). On the other hand, Moazen et al. (2012), also used pET expression system for the optimum expression of *Taq* DNA polymerase after 2 h of IPTG the induction. Seo et al. 2007, cloned and expressed the gene coding *Staphylothermus marinus* DNA ligase using pET-22b (+) in *Escherichia coli* BL21-CodonPlus (DE3)-RIL. They also showed that their DNA ligase could catalyze blunt-end intermolecular joining of DNA sequences in the presence of tricine-NaOH buffer and Mn(2+), using either ATP or ADP however none of them reported using metagenomic approach to mine for the genes .

3.4. Conclusion

In this study, the genes encoding DNA polymerase, DNA ligase and endonuclease II derived from a hot spring metagenomic sequences were successfully synthesized and cloned in to pET-30a (+) and expressed using a T7-based promoter system in *E. coli* BL21 (DE3). *In silico* analysis of the sequences and SDS-PAGE gel revealed that molecular weight for DNA polymerase, DNA ligase, and endonuclease II were 99 901.4, 71 017.6 and 23 608 Da, respectively. The purification protocol resulted in proteins with 80- 95 % purity by SDS PAGE analysis.

3.5. References

AL- MANASRA, A. & AL- RAZEN, F. 2012. Cloning and Expression of a new bacteriophage (SHP*h*) DNA ligase isolated from sewage. *Journal of Genetic Engineering and Biotechnology*, 10:177- 184.

BRADFORD, M.M. 1976. A rapid sensitive method for the quantification of microgram quantities of protein utilizing the principle of protein- dye binding. *Anal Biochem*. 72:248-254.

CARRIO, M.M. & VILLAVERDE, A. 2005. Localization of chaperons Dnak and GroEL in bacterial inclusion bodies. *J. Bacteriol*. 187; 10: 3599-3601.

CHEN, R. 2012. Bacterial expression systems for recombinant protein production: *E. coli* and beyond. *Biotechnology Advances*. 30: 1102-1107.

COTTERILL, S. & KEARSEY, S. 2008. DNAReplication: A database of information and resources for the eukaryotic DNA replication community. Nucleic acids research. 37.

DEMAIN, A. L. AND VAISHNAV. P. 2009. Production of recombinant proteins by microbes and higher organisms. *Biotechnology Advances*. 27: 297-306.

DESAI, U. J. & PFAFFLE, P. K. 1995. Single-step purification of thermostable DNA polymerase expressed in *Escherichia coli*. *Biotechniques*.19:780–4.

DROUIN, R., DRIDT, W. AND SAMASSEKOU, O. 2007. DNA polymerases for PCR applications. Industrial Enzymes: Structure, Function and Applications. 379- 401.

ESPOSITO, D. AND CHATTERJEE, D. K. 2006. Enhancement of soluble protein expression through the use of fusion tags. *Current Opinion in Biotechnology*. 17: 353-358.

GARCIA- DIAZ, M. & BEBENEK, K. 2007. Multiple function of DNA polymerase. *Critical Reviews in Plant Sciences*. 26; 2:105-122.

GHOLIZADEH, A., FAIZI, M.H. & BAGHBAN, K. B. 2010. Induced expression of *Eco*RI endonuclease as an active maltose-binding fusion protein in *Escherichia coli. Microbiology*. 79: 167- 72.

HSIEH, P., XIAO, J., O'LOANE, D. & XU, S. 2000. Cloning, Expression, and Purification of a Thermostable Nonhomodimeric Restriction Enzyme, *Bsl*I. *Journal of Bacteriology*. 182;4 :949-955

KAGUNI, M. J. 2018. The Macromolecular Machines that Duplicates the *Escherichia Coli* Chromosome as targets for Drug discovery. *Antibiotics*. 7;1: 23

KOSURI, S. & CHURCH, G.M. 2014. Large- scale de- novo DNA synthesis technologies and application. *Nature Methods*.11: 499-507.

LOHMAN, G. J., TABOR, S., & NICHOLS, N. M. 2011. DNA ligases. *Curr. Protoc. Biol.* 3; 3: 14.

MOAZEN, F.,RASTEGARI, A., HOSEINI, S.M., PANJEHPOUR, M., MIROLIAEI, M., & SADEGHI, H. M. 2012. Optimization of Taq DNA polymerase enzyme expression in Escherichia coli. *Advanced biomedical research.* 1-82.

NINFA, J. A., BALOU, D. P. & BENORE, M. 2010. Fundamental Laboratory Approaches for Biochemistry and Biotechnology. Hoboken, N.J: *John Wiley & Sons*. 341.

PALMER, I & WINGFIELD, P.T. 2004. Preparation of inducible (Inclusion- Body) proteins from *Escherichia coli*. *Curr Protoc Protein sci*. 6: Unit 6.3.

PASCAL, J. M. 2008. DNA and RNA ligases: Structural variations and shared mechanisms. *Curr. Opin. Struct. Biol.* 18; 1: 96-105.

PINGOUD, A. & JELTSCH, A. 2001. Structure and function of type II restriction endonucleases. *Nucleic Acids Research*. 29; 18: 3705–27.

ROSANO, G. L. AND CECCARELLI, E. A. 2014. Recombinant protein expression in microbial systems. *Frontiers in Microbiology*. 5: 1-2.

ROSSI, R., MONTECUCCO, A., CIARROCCHI, G. & BIAMONTI, G. 1997. Functional Characterization of the T4 DNA Ligase: A new insight into the mechanism of action. *Nucleic Acids Res.* 1; 11: 2106- 13.

RUSSELL, D.W. & SAMBROOK, J. 2001. Molecular cloning: a laboratory manual. Cold Spring Harbor, N.Y: Cold Spring Harbor Laboratory. ISBN 0-87969-576-5.

SEO, M.S., YUN, M.S., JEONG, J.K., CHOI, J., LEE, S. M., KIM, J.H., LEE, J. H. & SINGH, S.M. & PANDA, K.A. 2005. Solubilization and refolding of bacterial inclusion body proteins. *J Biosc Bioeng*. 99; 4:303- 310.

SOM, S., BHAGWAT, A.S. & FRIEDMAN, S. 1987. Nucleotide Sequence and Expression of the Gene Encoding the *Eco*RII Modification Enzyme, *Nucleic Acids Res.*12: 313–332.

STRUCTURAL GENOMICS CONSORTIUM, CHINA STRUCTURAL GENOMICS CONSORTIUM, NORTHEAST STRUCTURAL GENOMICS CONSORTIUM, GRASLUND, GUNSALUS, K. C. 2008. Protein production and purification. Nature methods. 5; 2: 135- 46.

STUDIER, F.W. 2005. Protein production by autoinduction in high density shaking cultures. *Protein Expr Purif.* 41;1:207- 34.

ZHANG, R., ZHU, Z., ZHU, H., NGUYEN, T., YAO, F., XIA, K., LIANG, D. & LIU, C. 2005. SNP Cutter: a comprehensive tool for SNP PCR-RFLP assay design. *Nucleic Acids Research*. 33 (Web Server issue): W489–92.

GENERAL CONCLUSION

The aim of this research project was to mine genes encoding for DNA manipulating enzymes such as DNA polymerase, DNA ligase and endonuclease II from a hot spring using metagenomic techniques. Discoveries of DNA-manipulating enzymes is very important for the role they play in molecular biology, life sciences R&D laboratories, molecular diagnostics tools for diseases diagnosis, development of novel point-of-care, as well as in epigenetics research. The high temperatures of hot springs as well as potentially essential and untapped microbial communities have attracted bioprospecting of these environments for thermostable enzymes. Hot spring soil sediments are richer in microbial diversity and the majority of microorganisms from extreme environments are not well categorized due to the inability to culture them in cultivation media. The hot spring ecological unit is an attractive source for the discovery of novel enzymes. Bio- prospecting a hot spring is the great platform to overcome the inability to culture microorganisms and mine for novel enzymes and isoforms of existing enzymes. Due to the importance of finding thermostable enzymes, the soil sediments associated with hot spring were used in this study to search for DNA-manipulating enzymes. The construction of metagenomic expression library serve as a potential source for screening wide variety of thermostable biomolecules in the future.

In this study, total metagenomic DNA from hot spring was successfully extracted and due to the quality of extracted DNA, a metagenomic library was successfully constructed using the fosmid based metagenomic library construction system. Metagenomic DNA was also successfully sequenced, *de novo* assembled, *in silico* analysed and genes encoding for DNA manipulating enzymes identified. Heterologous gene expression was used to express DNA manipulating enzymes using a pET30a (+) expression vector and *E. coli* strain. The histidine tagged recombinants were overexpressed under the control of the T7 expression promoter and 0.5 mM IPTG. All of the cloned genes (DNA polymerase, DNA ligase and endonuclease) were successfully expressed. The application of affinity purification supported purification of extracted proteins. In conclusion, the outcome of this work proves that hot spring metagenome can be accessed without the need to first cultivate the microorganisms and that is a valuable source for DNA-manipulating enzymes and other enzymes.

RECOMMENDATIONS

The expressed and purified DNA-manipulating enzymes can be functionally analyzed and studied further for their potential use as alternatives to commercially available ones. Moreover, metagenomic library usually harbors thousands of genes, further research can be undertaken to explore the constructed hot spring metagenomic fosmid library for thermostable enzymes of wide industrial and pharmaceautical applications. Due to lack or difficulty in screening for other important enzymes captured in the metagenomic library, the focus should also be directed towards finding ways of phenotypically screening for such enzymes.

APPENDIX

Appendix A: Reagents used in the study

Appendix A1: List of Buffers and solutions used in the study

Buffers and solutions	Composition

10X TBE
(pH 8.3)
	108 g Tris base;
	55 g Boric acid;
	7.45 g EDTA
	In 1000 ml of dH_2O

10X TGS
(pH 8.3)
	30.0 g of Tris base;
	144.0 g of glycine;
	10.0 g of SDS
	In 1000 ml of dH_2O.

6X DNA Loading dye	30% (v/v) glycerol 0.25% (w/v) bromophenol blue 0.25% (w/v) xylene cyanol FF

Binding Buffer
(pH 7.4)
	20 mM sodium phosphate;
	0.5 M NaCl;
	20-40 mM imidazole
	In 1000 ml of dH_2O

Coomassie staining solution	10% (V/V) Acetic acid;
	50% (V/V) Methanol;
	0.1% (W/V) coomassie blue R250;
	40% dH_2O

CTAB
(pH 8.0)
	100 mM Tris-HCl;
	1.4 M NaCl;
	20 mM EDTA;
	2% CTAB
	In 1000 ml of dH_2O

De-staining solution	10% (V/V) Acetic acid;
	50% (V/V) Methanol;
	40% dH_2O

1 M EDTA **(pH 8.0)**	186.1 g of disodium EDTA•2H_2O; In 400 mL of dH_2O Adjust the pH to 8.0 with NaOH

Elution Buffer
(pH 7.4)
	20 mM sodium phosphate;
	0.5 M NaCl;
	500 mM imidazole
	In 1000 ml of dH_2O

End-Repair 10X Buffer
(pH 7.5)
	330 mM Tris-acetate [pH 7.5];
	660 mM potassium acetate;
	100 mM magnesium acetate;
	5 mM DTT
	In 100 ul

Laemmli sample buffer
(pH 6.8)
	65.8 mM Tris-HCl;
	26.3% (w/v) glycerol;
	2.1% SDS;
	0.01% bromophenol blue

Phage Dilution Buffer
(pH 8.3)
	10 mM Tris-HCl [pH 8.3];
	100 mM NaCl;
	10 mM $MgCl_2$
	In 1000 ml of dH_2O

TE **(pH 8.0)**	10 mM Tris-HCl (pH 8.0); 1 mM EDTA (pH 8.0
1 M Tris-HCl **(pH 8.0)**	121.1 g of Tris base in 800 ml of H_2O; Adjust pH with 32% HCl

Appendix A2: List of antibiotics and inducer used in the study

Reagents	Preparation
Chloramphenicol (34 mg/ml): antibiotic	0.680g of chloramphenicol; 20 ml of 100% ethanol Filter-sterilised and aliquoted in 2 ml sterile microcentrifuge tubes. 12.5 µg/ml is the final concentration used.
Kanamycin (100 mg/ml): antibiotic	2g of kanamycin; 20 ml of distilled water Filter-sterilised and aliquoted in 2 ml sterile microcentrifuge tubes. 50 µg/ml is the final concentration used
Isopropyl-β-D-thiogalactopyranoside (IPTG) 1M: inducer	4.77g of IPTG 20 ml of distilled water Filter-sterilised and aliquoted in 2 ml sterile microcentrifuge tubes. 0.1 M is the final concentration used

Appendix A3: List of media used in the study

Media	Formulation per Liter	Preparation

LB agar (pH 7.5)	10 g Peptone;
	5 g Yeast Extract;
	10 g Sodium Chloride;
	12 g Bacteriological Agar
		All the components are mixed and autoclaved at 121 °C for 15 minutes. Medium is cooled at 50-60 °C to be poured in sterile petri dishes.

LB broth (pH 7.5)	10 g Peptone;
	5 g Yeast Extract;
	10 g Sodium Chloride
		All the components are mixed and autoclaved at 121 °C for 15 minutes.

SOC agar (pH 7.0)	10 mM magnesium chloride, 10 Mm magnesium sulfate, 2.5 mM potassium chloride, 10 mMsodium chloride, 2% tryptone, 0.5% yeast extract, 20 mM glucose (to be added after autoclaving)	All components are mixed and autoclaved at 121 °C for 15 minutes. The medium is cooled then 20 mM glucose is added after autoclaving.

Appendix A4: List of microorganisms used in the study

Strains	Features	Supplier
EPI300™-T1R Phage T1-resistant *Escherichia coli* Plating strain	*[F– mcrA Δ(mrr-hsdRMS-mcrBC) (StrR) φ80dlacZΔM15 ΔlacX74 recA1 endA1 araD139 Δ(ara, leu)7697 galU galK λ– rpsL nupG trfA tonA dhfr]*	Epicentre
***Escherichia coli* BL21 (DE3) competent cells**	F^{-} *ompT gal dcm lon* $hsdS_{B}(r_{B}^{-}m_{B}^{-})$ *λ(DE3 [lacI lacUV5-T7p07 ind1 sam7 nin5])* $[malB^{+}]_{K\text{-}12}(\lambda^{S})$ *pLysS[T7p20* $ori_{p15A}](Cm^{R})$	Lucigen

Appendix A5: List of all vectors used in the study

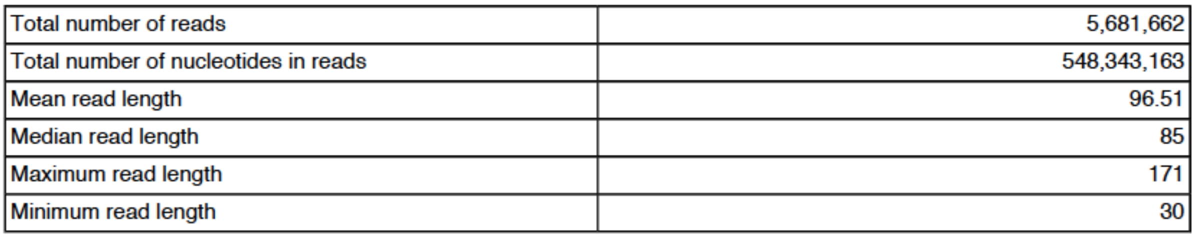
Vectors	Features	Selective marker	Supplier
pCC2FOS	Copy controlled vector, linearized and dephosphorylated at Eco72I restriction site. Requires EPI300™-T1R E. coli strain for high copy number induction, used for construction of fosmid library.	Chloramphenicol	Epicentre
pET30a(+)	Expression vector of N and C-terminally His-tagged protein	Kanamycin	Merck

Appendix A6: Preparation of 12% separating gels and 4% stacking gels for SDS-PAGE

	Solution components	Volume (ml)	
		12% Resolving gel	4% stacking gel
	Distilled water	3.3	3.4
	30% Bis-acrylamide mix	4.0	0.83
	1.5M Tris-HCl (pH 8.8)	2.5	0.63
	10% SDS	0.1	0.05
	10% Ammonium persulfate (APS)	0.1	0.05
	TEMED	0.005	0.005

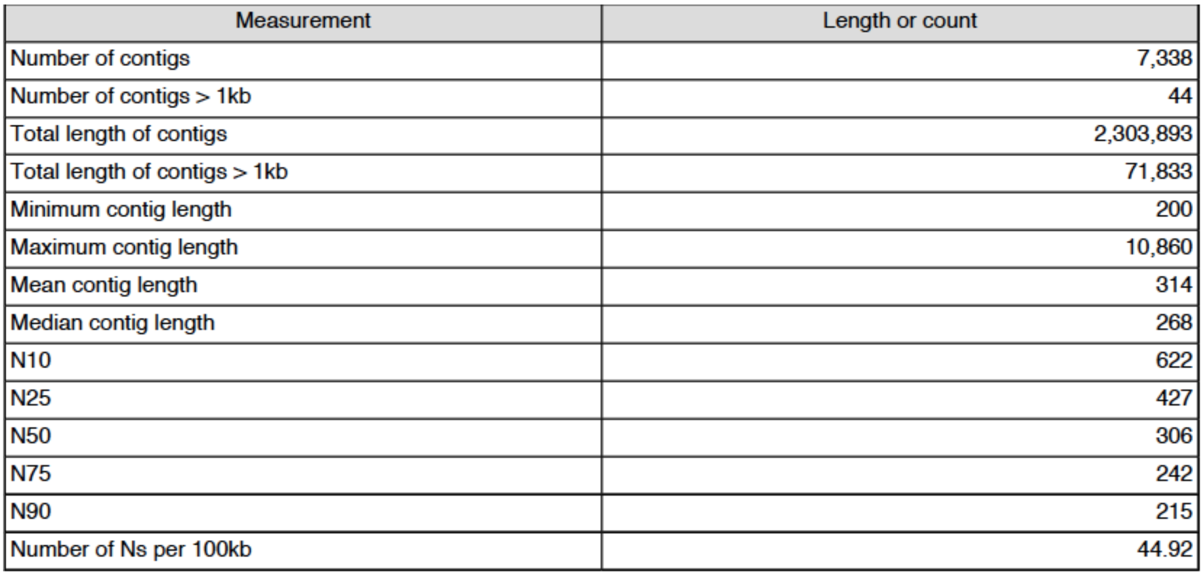
Appendix B: Summary of statistical analysis of the *De novo* assembly and alignments of DNA sequences

Appendix B1: Basic statistics on input reads for metagenomic DNA from hot spring



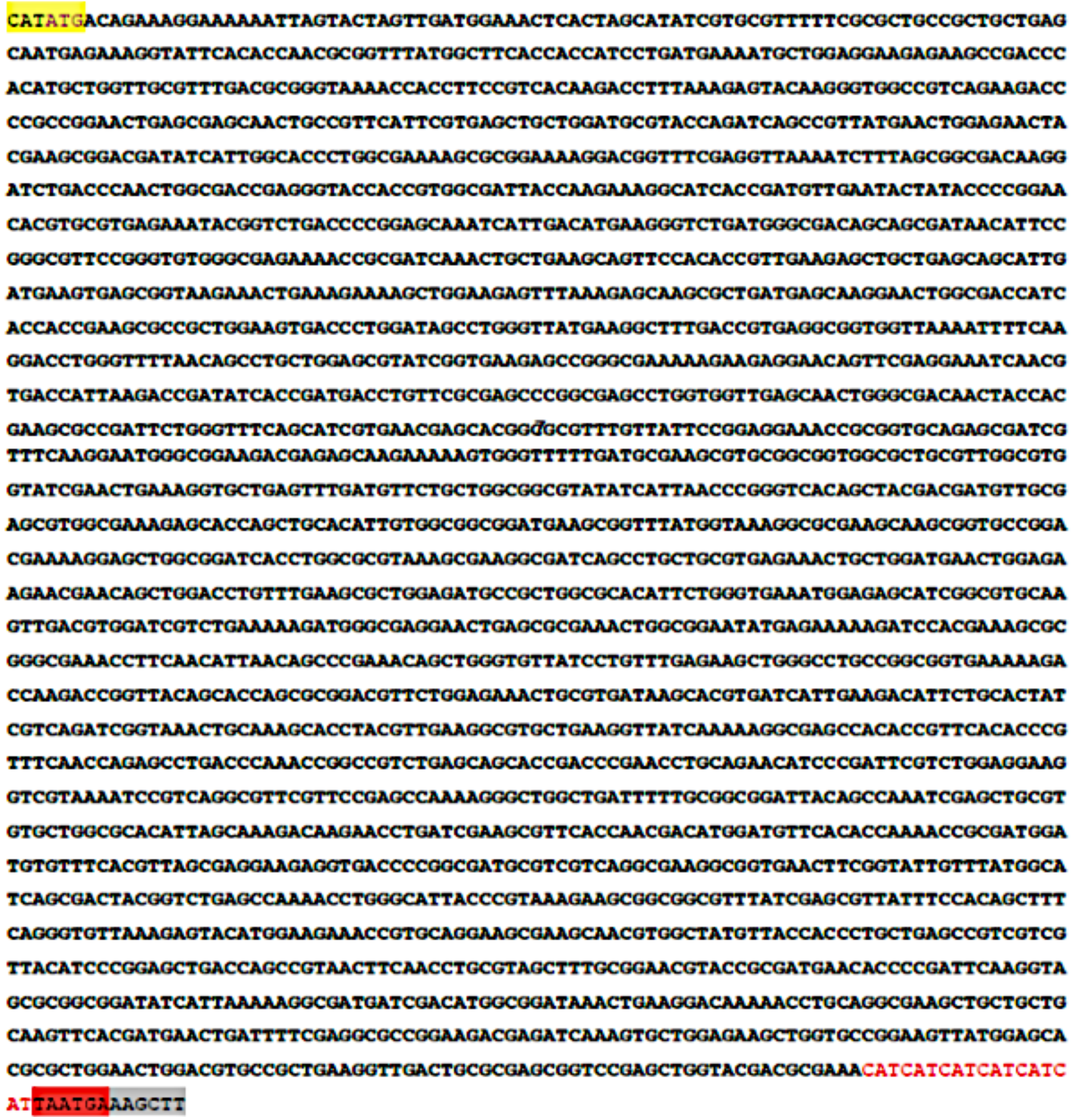
Total number of reads	5,681,662
Total number of nucleotides in reads	548,343,163
Mean read length	96.51
Median read length	85
Maximum read length	171
Minimum read length	30

Appendix B2: Basic statistics on contigs



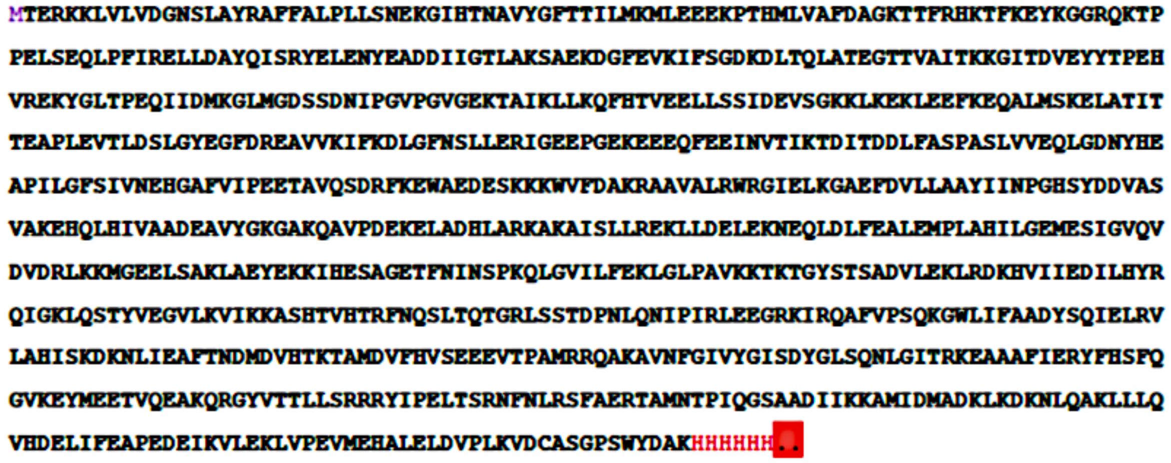
Measurement	Length or count
Number of contigs	7,338
Number of contigs > 1kb	44
Total length of contigs	2,303,893
Total length of contigs > 1kb	71,833
Minimum contig length	200
Maximum contig length	10,860
Mean contig length	314
Median contig length	268
N10	622
N25	427
N50	306
N75	242
N90	215
Number of Ns per 100kb	44.92

Appendix B3: A 2670 bp Putative DNA polymerase gene fragment with *Nde*I and *Hind*III restriction sites and 6- His tag the C- terminal to be used for cloning.



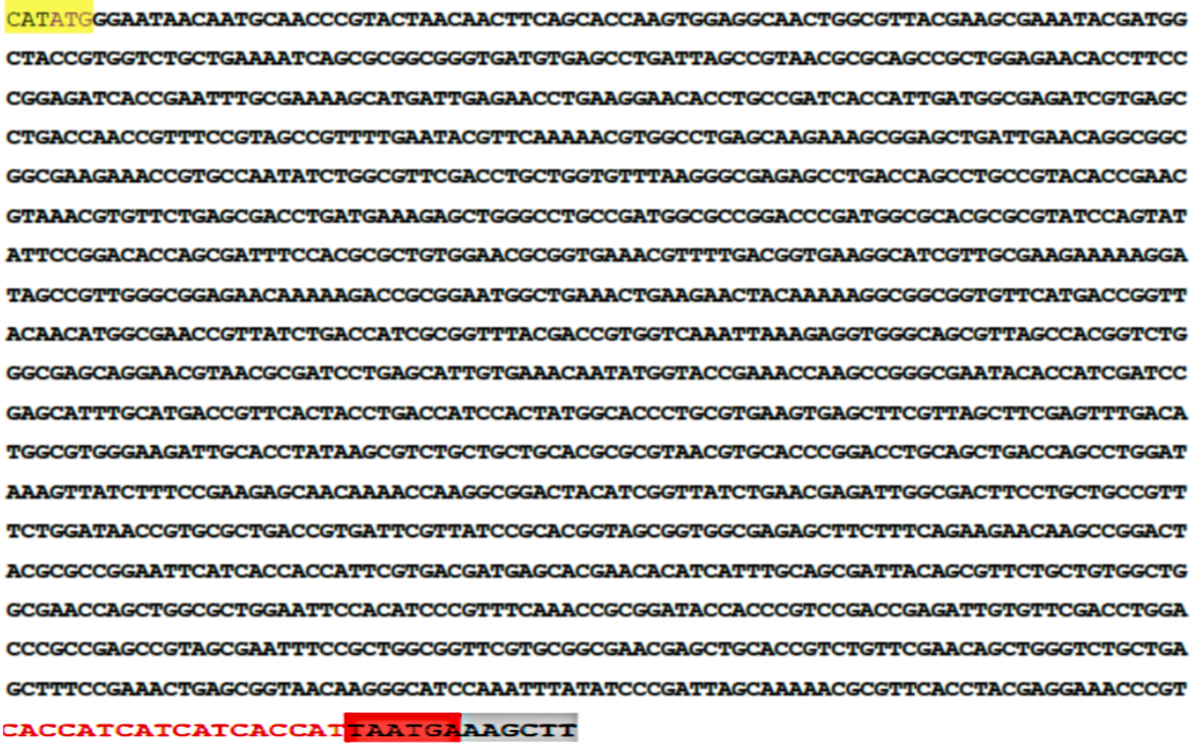
```
CATATGACAGAAAGGAAAAAATTAGTACTAGTTGATGGAAACTCACTAGCATATCGTGCGTTTTTCGCGCTGCCGCTGCTGAG
CAATGAGAAAGGTATTCACACCAACGCGGTTTATGGCTTCACCACCATCCTGATGAAAATGCTGGAGGAAGAGAAGCCGACCC
ACATGCTGGTTGCGTTTGACGCGGGTAAAACCACCTTCCGTCACAAGACCTTTAAAGAGTACAAGGGTGGCCGTCAGAAGACC
CCGCCGGAACTGAGCGAGCAACTGCCGTTCATTCGTGAGCTGCTGGATGCGTACCAGATCAGCCGTTATGAACTGGAGAACTA
CGAAGCGGACGATATCATTGGCACCCTGGCGAAAAGCGCGGAAAAGGACGGTTTCGAGGTTAAAATCTTTAGCGGCGACAAGG
ATCTGACCCAACTGGCGACCGAGGGTACCACCGTGGCGATTACCAAGAAAGGCATCACCGATGTTGAATACTATACCCCGGAA
CACGTGCGTGAGAAATACGGTCTGACCCCGGAGCAAATCATTGACATGAAGGGTCTGATGGGCGACAGCAGCGATAACATTCC
GGGCGTTCCGGGTGTGGGCGAGAAAACCGCGATCAAACTGCTGAAGCAGTTCCACACCGTTGAAGAGCTGCTGAGCAGCATTG
ATGAAGTGAGCGGTAAGAAACTGAAAGAAAAGCTGGAAGAGTTTAAAGAGCAAGCGCTGATGAGCAAGGAACTGGCGACCATC
ACCACCGAAGCGCCGCTGGAAGTGACCCTGGATAGCCTGGGTTATGAAGGCTTTGACCGTGAGGCGGTGGTTAAAATTTTCAA
GGACCTGGGTTTTAACAGCCTGCTGGAGCGTATCGGTGAAGAGCCGGGCGAAAAAGAAGAGGAACAGTTCGAGGAAATCAACG
TGACCATTAAGACCGATATCACCGATGACCTGTTCGCGAGCCCGGCGAGCCTGGTGGTTGAGCAACTGGGCGACAACTACCAC
GAAGCGCCGATTCTGGGTTTCAGCATCGTGAACGAGCACGGGGCGTTTGTTATTCCGGAGGAAACCGCGGTGCAGAGCGATCG
TTTCAAGGAATGGGCGGAAGACGAGAGCAAGAAAAAGTGGGTTTTTGATGCGAAGCGTGCGGCGGTGGCGCTGCGTTGGCGTG
GTATCGAACTGAAAGGTGCTGAGTTTGATGTTCTGCTGGCGGCGTATATCATTAACCCGGGTCACAGCTACGACGATGTTGCG
AGCGTGGCGAAAGAGCACCAGCTGCACATTGTGGCGGCGGATGAAGCGGTTTATGGTAAAGGCGCGAAGCAAGCGGTGCCGGA
CGAAAAGGAGCTGGCGGATCACCTGGCGCGTAAAGCGAAGGCGATCAGCCTGCTGCGTGAGAAACTGCTGGATGAACTGGAGA
AGAACGAACAGCTGGACCTGTTTGAAGCGCTGGAGATGCCGCTGGCGCACATTCTGGGTGAAATGGAGAGCATCGGCGTGCAA
GTTGACGTGGATCGTCTGAAAAAGATGGGCGAGGAACTGAGCGCGAAACTGGCGGAATATGAGAAAAAGATCCACGAAAGCGC
GGGCGAAACCTTCAACATTAACAGCCCGAAACAGCTGGGTGTTATCCTGTTTGAGAAGCTGGGCCTGCCGGCGGTGAAAAAGA
CCAAGACCGGTTACAGCACCAGCGCGGACGTTCTGGAGAAACTGCGTGATAAGCACGTGATCATTGAAGACATTCTGCACTAT
CGTCAGATCGGTAAACTGCAAAGCACCTACGTTGAAGGCGTGCTGAAGGTTATCAAAAAGGCGAGCCACACCGTTCACACCCG
TTTCAACCAGAGCCTGACCCAAACCGGCCGTCTGAGCAGCACCGACCCGAACCTGCAGAACATCCCGATTCGTCTGGAGGAAG
GTCGTAAAATCCGTCAGGCGTTCGTTCCGAGCCAAAAGGGCTGGCTGATTTTTGCGGCGGATTACAGCCAAATCGAGCTGCGT
GTGCTGGCGCACATTAGCAAAGACAAGAACCTGATCGAAGCGTTCACCAACGACATGGATGTTCACACCAAAACCGCGATGGA
TGTGTTTCACGTTAGCGAGGAAGAGGTGACCCCGGCGATGCGTCGTCAGGCGAAGGCGGTGAACTTCGGTATTGTTTATGGCA
TCAGCGACTACGGTCTGAGCCAAAACCTGGGCATTACCCGTAAAGAAGCGGCGGCGTTTATCGAGCGTTATTTCCACAGCTTT
CAGGGTGTTAAAGAGTACATGGAAGAAACCGTGCAGGAAGCGAAGCAACGTGGCTATGTTACCACCCTGCTGAGCCGTCGTCG
TTACATCCCGGAGCTGACCAGCCGTAACTTCAACCTGCGTAGCTTTGCGGAACGTACCGCGATGAACACCCCGATTCAAGGTA
GCGCGGCGGATATCATTAAAAAGGCGATGATCGACATGGCGGATAAACTGAAGGACAAAAACCTGCAGGCGAAGCTGCTGCTG
CAAGTTCACGATGAACTGATTTTCGAGGCGCCGGAAGACGAGATCAAAGTGCTGGAGAAGCTGGTGCCGGAAGTTATGGAGCA
CGCGCTGGAACTGGACGTGCCGCTGAAGGTTGACTGCGCGAGCGGTCCGAGCTGGTACGACGCGAAACATCATCATCATCATC
ATTAATGAAAGCTT
```

Appendix B4: The *in silico* representation of 885 amino acids DNA polymerase protein sequence with a 6- His tag for protein purification.



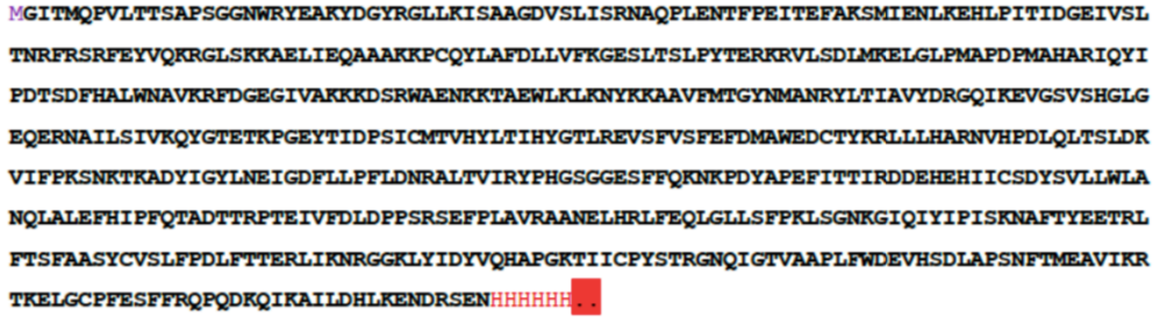
```
MTERKKLVLVDGNSLAYRAFFALPLLSNEKGIHTNAVYGFTTILMKMLEEEKPTHMLVAFDAGKTTFRHKTFKEYKGGRQKTP
PELSEQLPFIRELLDAYQISRYELENYEADDIIGTLAKSAEKDGFEVKIFSGDKDLTQLATEGTTVAITKKGITDVEYYTPEH
VREKYGLTPEQIIDMKGLMGDSSDNIPGVPGVGEKTAIKLLKQFHTVEELLSSIDEVSGKKLKEKLEEFKEQALMSKELATIT
TEAPLEVTLDSLGYEGFDREAVVKIFKDLGFNSLLERIGEEPGEKEEEQFEEINVTIKTDITDDLFASPASLVVEQLGDNYHE
APILGFSIVNEHGAFVIPEETAVQSDRFKEWAEDESKKKWVFDAKRAAVALRWRGIELKGAEFDVLLAAYIINPGHSYDDVAS
VAKEHQLHIVAADEAVYGKGAKQAVPDEKELADHLARKAKAISLLREKLLDELEKNEQLDLFEALEMPLAHILGEMESIGVQV
DVDRLKKMGEELSAKLAEYEKKIHESAGETFNINSPKQLGVILFEKLGLPAVKKTKTGYSTSADVLEKLRDKHVIIEDILHYR
QIGKLQSTYVEGVLKVIKKASHTVHTRFNQSLTQTGRLSSTDPNLQNIPIRLEEGRKIRQAFVPSQKGWLIFAADYSQIELRV
LAHISKDKNLIEAFTNDMDVHTKTAMDVFHVSEEEVTPAMRRQAKAVNFGIVYGISDYGLSQNLGITRKEAAAFIERYFHSFQ
GVKEYMEETVQEAKQRGYVTTLLSRRRYIPELTSRNFNLRSFAERTAMNTPIQGSAADIIKKAMIDMADKLKDKNLQAKLLLQ
VHDELIFEAPEDEIKVLEKLVPEVMEHALELDVPLKVDCASGPSWYDAKHHHHHH..
```

Appendix B5: A 1881 bp Putative DNA ligase gene fragment with *Nde*I on the N-terminus and *Hind*III restriction site on the C- terminus. The 6- Histag (red) were also incorporated to help facilitate purification.



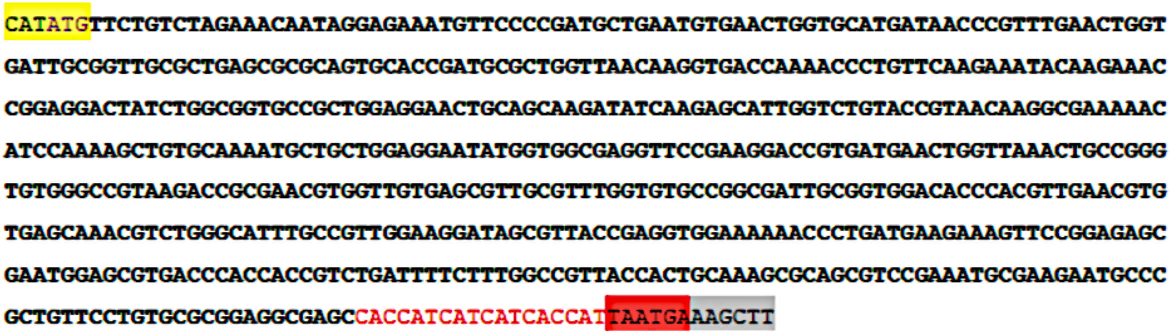
```
CATATGGGAATAACAATGCAACCCGTACTAACAACTTCAGCACCAAGTGGAGGCAACTGGCGTTACGAAGCGAAATACGATGG
CTACCGTGGTCTGCTGAAAATCAGCGCGGCGGGTGATGTGAGCCTGATTAGCCGTAACGCGCAGCCGCTGGAGAACACCTTCC
CGGAGATCACCGAATTTGCGAAAAGCATGATTGAGAACCTGAAGGAACACCTGCCGATCACCATTGATGGCGAGATCGTGAGC
CTGACCAACCGTTTCCGTAGCCGTTTTGAATACGTTCAAAAACGTGGCCTGAGCAAGAAAGCGGAGCTGATTGAACAGGCGGC
GGCGAAGAAACCGTGCCAATATCTGGCGTTCGACCTGCTGGTGTTTAAGGGCGAGAGCCTGACCAGCCTGCCGTACACCGAAC
GTAAACGTGTTCTGAGCGACCTGATGAAAGAGCTGGGCCTGCCGATGGCGCCGGACCCGATGGCGCACGCGCGTATCCAGTAT
ATTCCGGACACCAGCGATTTCCACGCGCTGTGGAACGCGGTGAAACGTTTTGACGGTGAAGGCATCGTTGCGAAGAAAAAGGA
TAGCCGTTGGGCGGAGAACAAAAAGACCGCGGAATGGCTGAAACTGAAGAACTACAAAAAGGCGGCGGTGTTCATGACCGGTT
ACAACATGGCGAACCGTTATCTGACCATCGCGGTTTACGACCGTGGTCAAATTAAAGAGGTGGGCAGCGTTAGCCACGGTCTG
GGCGAGCAGGAACGTAACGCGATCCTGAGCATTGTGAAACAATATGGTACCGAAACCAAGCCGGGCGAATACACCATCGATCC
GAGCATTTGCATGACCGTTCACTACCTGACCATCCACTATGGCACCCTGCGTGAAGTGAGCTTCGTTAGCTTCGAGTTTGACA
TGGCGTGGGAAGATTGCACCTATAAGCGTCTGCTGCTGCACGCGCGTAACGTGCACCCGGACCTGCAGCTGACCAGCCTGGAT
AAAGTTATCTTTCCGAAGAGCAACAAAACCAAGGCGGACTACATCGGTTATCTGAACGAGATTGGCGACTTCCTGCTGCCGTT
TCTGGATAACCGTGCGCTGACCGTGATTCGTTATCCGCACGGTAGCGGTGGCGAGAGCTTCTTTCAGAAGAACAAGCCGGACT
ACGCGCCGGAATTCATCACCACCATTCGTGACGATGAGCACGAACACATCATTTGCAGCGATTACAGCGTTCTGCTGTGGCTG
GCGAACCAGCTGGCGCTGGAATTCCACATCCCGTTTCAAACCGCGGATACCACCCGTCCGACCGAGATTGTGTTCGACCTGGA
CCCGCCGAGCCGTAGCGAATTTCCGCTGGCGGTTCGTGCGGCGAACGAGCTGCACCGTCTGTTCGAACAGCTGGGTCTGCTGA
GCTTTCCGAAACTGAGCGGTAACAAGGGCATCCAAATTTATATCCCGATTAGCAAAAACGCGTTCACCTACGAGGAAACCCGT
CACCATCATCATCACCATTAATGAAAGCTT
```

Appendix B6: A representation of 622 amino acids DNA ligase protein sequence with a 6- His tag at the C-terminus.



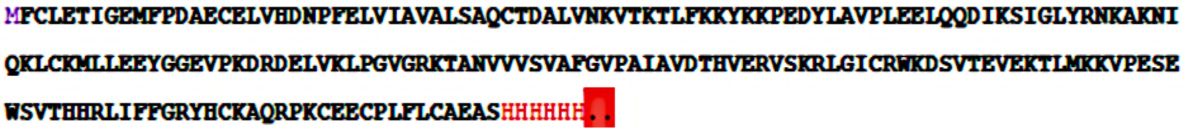
```
MGITMQPVLTTSAPSGGNWRYEAKYDGYRGLLKISAAGDVSLISRNAQPLENTFPEITEFAKSMIENLKEHLPITIDGEIVSL
TNRFRSRFEYVQKRGLSKKAELIEQAAAKKPCQYLAFDLLVFKGESLTSLPYTERKRVLSDLMKELGLPMAPDPMAHARIQYI
PDTSDFHALWNAVKRFDGEGIVAKKKDSRWAENKKTAEWLKLKNYKKAAVFMTGYNMANRYLTIAVYDRGQIKEVGSVSHGLG
EQERNAILSIVKQYGTETKPGEYTIDPSICMTVHYLTIHYGTLREVSFVSFEFDMAWEDCTYKRLLLHARNVHPDLQLTSLDK
VIFPKSNKTKADYIGYLNEIGDFLLPFLDNRALTVIRYPHGSGGESFFQKNKPDYAPEFITTIRDDEHEHIICSDYSVLLWLA
NQLALEFHIPFQTADTTRPTEIVFDLDPPSRSEFPLAVRAANELHRLFEQLGLLSFPKLSGNKGIQIYIPISKNAFTYEETRL
FTSFAASYCVSLFPDLFTTERLIKNRGGKLYIDYVQHAPGKTIICPYSTRGNQIGTVAAPLFWDEVHSDLAPSNFTMEAVIKR
TKELGCPFESFFRQPQDKQIKAILDHLKENDRSENHHHHHH..
```

Appendix B7: A nucleotides representation 636 bp Putative Endonuclease II gene fragment with 6- His tag at the C- terminal as synthesised at GenScript.



```
CATATGTTCTGTCTAGAAACAATAGGAGAAATGTTCCCCGATGCTGAATGTGAACTGGTGCATGATAACCCGTTTGAACTGGT
GATTGCGGTTGCGCTGAGCGCGCAGTGCACCGATGCGCTGGTTAACAAGGTGACCAAAACCCTGTTCAAGAAATACAAGAAAC
CGGAGGACTATCTGGCGGTGCCGCTGGAGGAACTGCAGCAAGATATCAAGAGCATTGGTCTGTACCGTAACAAGGCGAAAAAC
ATCCAAAAGCTGTGCAAAATGCTGCTGGAGGAATATGGTGGCGAGGTTCCGAAGGACCGTGATGAACTGGTTAAACTGCCGGG
TGTGGGCCGTAAGACCGCGAACGTGGTTGTGAGCGTTGCGTTTGGTGTGCCGGCGATTGCGGTGGACACCCACGTTGAACGTG
TGAGCAAACGTCTGGGCATTTGCCGTTGGAAGGATAGCGTTACCGAGGTGGAAAAAACCCTGATGAAGAAAGTTCCGGAGAGC
GAATGGAGCGTGACCCACCACCGTCTGATTTTCTTTGGCCGTTACCACTGCAAAGCGCAGCGTCCGAAATGCGAAGAATGCCC
GCTGTTCCTGTGCGCGGAGGCGAGCCACCATCATCATCACCATTAATGAAAGCTT
```

Appendix B8: A representation of 207 amino acids Endonuclease II protein sequence with a 6- His tag for protein purification.



```
MFCLETIGEMFPDAECELVHDNPFELVIAVALSAQCTDALVNKVTKTLFKKYKKPEDYLAVPLEELQQDIKSIGLYRNKAKNI
QKLCKMLLEEYGGEVPKDRDELVKLPGVGRKTANVVVSVAFGVPAIAVDTHVERVSKRLGICRWKDSVTEVEKTLMKKVPESE
WSVTHHRLIFFGRYHCKAQRPKCEECPLFLCAEASHHHHHH..
```