UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA
Denkleiers • Leading Minds • Dikgopolo tša Dihlalefi

Identification and characterisation of nucleic acid manipulating enzymes from metaviromic DNA

By

Mashikoane Pinky Jane Segobola

27435254

Submitted in partial fulfilment of the requirements for the degree

PhD Biotechnology

In the

FACULTY OF NATURAL AND AGRICULTURAL SCIENCES

At the

UNIVERSITY OF PRETORIA

Supervisor:

Prof Don Cowan

Co-supervisors:

Dr Tsepo Tsekoa

Dr Konanani Rashamuse

Dr Evelien Adriaenssens

2019

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA
Denkleiers • Leading Minds • Dikgopolo tša Dihlalefi

DECLARATION

I, Mashikoane Pinky Jane Segobola declare that the thesis, which I hereby submit for the degree *Philosophiae Doctor* (Biotechnology) (Ph.D.) at the University of Pretoria, is my own work and has not been previously submitted by me for a degree at this or any other tertiary institution.

Signature ……..

Date …….

ACKNOWLEDGEMENTS

I am grateful to my supervisor Prof Don Cowan for affording me the opportunity to be his student, for your guidance and constructive criticism, for being patient and inspirational.

I am also very thankful to my co-supervisors, Dr Konanani Rashamuse, Dr Tsepo Tsekoa and Dr Evelien Adriaenssens for their valued guidance, patience and support all through my journey as a student.

I wish to acknowledge the support of these institutions: National Research Foundation (NRF) for financial support, The Department of Science and Technology, The University of Pretoria Genomics Research Institute, The Claude Leon Foundation (to EMA), Cape Nature (Kogelberg Biosphere Reserve) and The Council for Scientific and Industrial Research (CSIR).

To Dr Lusisizwe Kwezi, Dr Ofentse Pooe and Dr Priyen Pilley at the CSIR, I would like to sincerely thank them for all their advices, help and encouragement. I thank the CMEG team at the University of Pretoria for their help with data processing and running the high- throughput computer.

My sincere appreciation to my CSIR laboratory mates Sibongile Mtimka, Gugu Ngwenya, Mulalo Nemutanzhela and Ntombifuthi Shezi for assisting me on some laboratory work.

Much appreciation goes to the rest of the technical staff for their diligent work and technical support.

Much appreciation to Dr. Genis Andres Castillo Villamizar from Georg-August University, Göttingen, Germany. Thank you for assisting with the *E. coli* mutant cells.

I thank the Department of Genetics and CMEG in the University of Pretoria and staff for their administrative assistance.

Lots of people deserve my gratitude for their encouragement, assistance, and friendship. A special mention to Mulalo Nemutanzhela, Albert Mabetha, Kgama Mathiba, Hope

Netshiya, Sipho Rangayi, Londiwe Khumalo, Dineo Moloele, Lerato Maselela and Advita Sigh.

I thank my family, my late Father Maena Cornelius Raphela, my mother Maria Rakgasago Raphela, my brother (Buti Oupa Raphela), my sisters (Nthabiseng, Thato and Thapelo) and my in-laws [GM Segobola (Father-in-law), Mokgadi Segobola (Mother-in-law), Maphure Segobola (Sister-in-law), Dennis Segobola (Brother-in-law)], for their incredible encouragement and understanding throughout my studies.

I also thank all my nephews and nieces for keeping me smiling even in hard times.

I am grateful to my kids, Phenyo and Arabile Segobola for their love, patience, cheering, understanding, support and for giving me reason to work hard every day.

Finally, my husband, Phokela Segobola, has my deepest and warmest gratitude for his patience, love, humour, support, assistance and friendship.

ABSTRACT

The market value of molecular biology enzymes is growing rapidly, due to increasing applications in the Biotechnology industry. Such nucleic acid manipulating enzymes include polymerases, ligases, nucleases, phosphatases, methylases and topoisomerases. In this study, we analysed soils from the Kogelberg Biosphere Reserve that is situated in area of high plant endemism within the Cape Floral Region. These soils are characterised by an acidic pH and are typically oligotrophic and yet support a unique vegetation type termed *fynbos* ('fine bush'). We carried out high throughput nucleic acid sequence analysis of bacterial 16S rRNA gene library and a fosmid library prepared from a soil suspension that was size-selected (0.22 μm) to enrich for viruses. Sequence data were assembled and analysed using the following bioinformatics (CLC genomics workbench, MetaVir, VIROME, MG-RAST, RAST and QIIME). Based on analysis of the 16S rRNA gene marker, there was a high level of bacterial diversity that was dominated by 5 bacterial taxonomic groups; namely: *Actinobacteria*, *Proteobacteria*, *Acidobacteria*, *Planctomycetes* and *Bacteroidetes*. The analysis of viral diversity using sequence data from *PolB, PolB2, terL* and *T7gp17* gene markers revealed many bacteriophages with several members of the order Caudovirales; such as *Siphoviridae*, *Podoviridae* and *Myoviridae*. A combination of sequence- and functional- based screening approaches was used to screen for open reading frames (ORFs) encoding nucleic acid manipulating enzymes. A total of nine (9) ORFs (sequence identify < 60) were identified and belonged to the following enzyme classes: three ligases (*RNALig2*, *RNALig3* and *DNAlig*), three DNA polymerases (PolB, PolA1 and PolA2), and three Nucleases (Restriction endonuclease (RE), Homing endonuclease (HNHc) and Endonuclease 7 (E7)). Various attempts were made to recombinantly express the identified ORFs, including the use of different expression vectors (pET20b(+), pET28a(+), pET30b(+) and pMAL-C5X) and host strains (*E. coli* BL21 DE3, BL21 DE3 *pLysS*, and BL21 AI cells) as well as trying various cultivation and induction conditions. A successful expression strategy was achieved only with *DNAlig* gene fused to a maltose-binding affinity tag using the pMAL-C5X expression vector. The purified recombinant DNALig protein was subsequently purified and assayed for activity. The purified DNALig protein showed an ATP-dependent DNA ligation activity and could actively ligate both restricted blunt-ended and sticky-ended restricted DNA molecule. Through the use of high-throughput next generation nucleic acid sequencing

coupled with sequence- and function- based screening methods, this study was able to highlight the value of analysing the soil metavirome for the discovery of novel nucleic acid manipulating enzymes for biotechnology research and development.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

KEY TERMS AND ABBREVIATIONS USED

AI	- Aluminium
ATP	- Adenosine triphosphate
bp	- base pairs
BLAST	- Basic Local Alignment Search Tool
BSA	- Bovine Serum Albumin
Ca	- Calcium
$CaCl_2$	- Calcium Chloride
CCA	- Canonical Correspondence Analysis
CDS	- coding DNA sequence
Cl_2	- Chlorine
cm	- Centimetre
Da	- Dalton
DGGE	- denaturing gradient gel electrophoresis
°C	- Degrees Celsius
DNA	- Deoxyribonucleic acid
dNTPs	- Deoxyribonucleic-5'-triphosphate
dsDNA	- double-stranded Deoxyribonucleic acid
DTT	- Ditriothritol
EDTA	- Ethylenediaminetetraacetic acid
EtOH	- Ethanol
Fe	- Iron
FISH	- fluorescent in situ hybridisation
GAAS	- Genome relative Abundance and Average Size
g	- gram
× g	- Centrifugal force
hr	- Hour
IPTG	- Isopropyl-β-D-thiogalactopyranoside
K	- Potassium
kb	- kilo base pairs
kDa	- kilo Dalton

KEGG	- Kyoto Encyclopaedia of Genes and Genomes
KO	- KEGG Orthology
λ	- lambda
l	- litre
LB	- Luria-Bertani
LEW	- Lysis-Equilibration
M	- molar
M5nr	- M5 non-redundant database
MDA	- Multiple displacement amplification
Mg	- Magnesium
$MgCl_2$	- magnesium chloride
$MgSO_4$	- Magnesium sulphate
MG-RAST	-Metagenomics-Rapid Annotation using Subsystem Technology
mL	- millilitre
min	- minute
mM	- mill molar
µg	- Microgram
µL	- Micro litre
MnCl	- manganese chloride
Mn	- Manganese
Na	- Sodium
NaCl	- sodium chloride
NCBI	- National Center for Biotechnology Information
NGS	- Next-generation sequencing
nt	- nucleotide
ORFs	- Open reading frames
OTU	- operational taxonomic unit
Pac Bio	- Pacific Biosciences
PEG	- Poly ethylene glycol
PCA	- Principal Component Analysis

PCR	- Polymerase chain reaction
pmol	- pico mole
QC	- Quality control
QIIME	- Quantitative Insights Into Microbial Ecology
rpm	- Revolutions per minute
rRNA	- Ribosomal ribonucleic acid
SDS-PAGE	- Sodium Dodecyl Sulphate Polyacrylamide Gel Electrophoresis
SO4	- Sulphate
ssDNA	- single-stranded Deoxyribonucleic acid
TA	- Tris acetic acid
TAE	- Tris-acetate-EDTA
TEM	- Transmission Electron Microscopy
T-RFLP	- Terminal Restriction Fragment Length Polymorphism
U	- Unit
UPGMA	- Unweighted-Pair Group Method Using Average linkages
VIROME	- Viral Informatics Resource for Metagenome Exploration
VLPs	- Virus like Particles

CHAPTER 1: LITERATURE REVIEW

1.1 The unique South African *fynbos* biome

The Cape Floral Region is positioned within the south-western part of South Africa, covering approximately 6% of the land area in the country (Goldblatt, 1997; Ramond *et al.*, 2015). The Cape Floral Region is the smallest of the six globally recognised Floral Kingdoms (Myers *et al.*, 2000) and is characterised by its unique plant biodiversity (Stafford *et al.*, 2005). Most of the plant species found in the Cape Floral Region are endemic to the south-western part of South Africa, and many plant species are extremely rare and in danger of extinction (Cowling *et al.*, 2003). The most prevalent vegetation type in this region is *fynbos* (Wintle *et al.*, 2011). *Fynbos* is short, scrubby vegetation which is characterised by the presence of restios (*Restionaceae*), proteas (*Proteaceae*) and heaths (*Ericaceae*). It is a particularly species rich vegetation type, having over 8700 plant species, with a large number of endemic species and only found in this region. This vegetation is also highly threatened and has therefore been recognised as a global biodiversity "hotspot" (Myers 1990).

The Kogelberg Biosphere Reserve (coordinates 34°04' to 34°24'S; 18°48' to 19°12'E) is approximately 40 km from the city of Cape Town and is part of the Cape Floral Region (Mucina and Rutherford, 2006; Ramond *et al.*, 2015) (Figure 1.1). The Kogelberg Biosphere Reserve is made up of sandy and acidic soil that are characterised by low nutrient levels (phosphorus, potassium and nitrogen) (Keeley, 2013). Several studies reported the existence of microbial communities associated with this region (Stafford *et al.*, 2005; Slabbert *et al.*, 2010; Ramond *et al.*, 2015; Miyambo *et al.*, 2016; Moroenyane *et al.*, 2016; Postma *et al.*, 2016a).

The research of Stafford *et al.* (2005) explored how the soil microbes are associated with the plant rhizosphere of specific *fynbos* species and these soils are colonised by bacteria belonging to *Proteobacteria, Firmicutes, Actinobacteria* and *Acidobacteria*. A more recent study by Slabbert *et al.* (2010) showed that the high beta diversity of above-ground plant species influence the bacterial diversity in *fynbos* soils. However, these studies focused on rhizosphere, mycorrhizal or fungal diversity (Allsopp and Stock, 1995; Caravaca *et al.*, 2002; Spriggs *et al.*, 2003). The viral diversity of *fynbos* soil has, to date been unexplored.

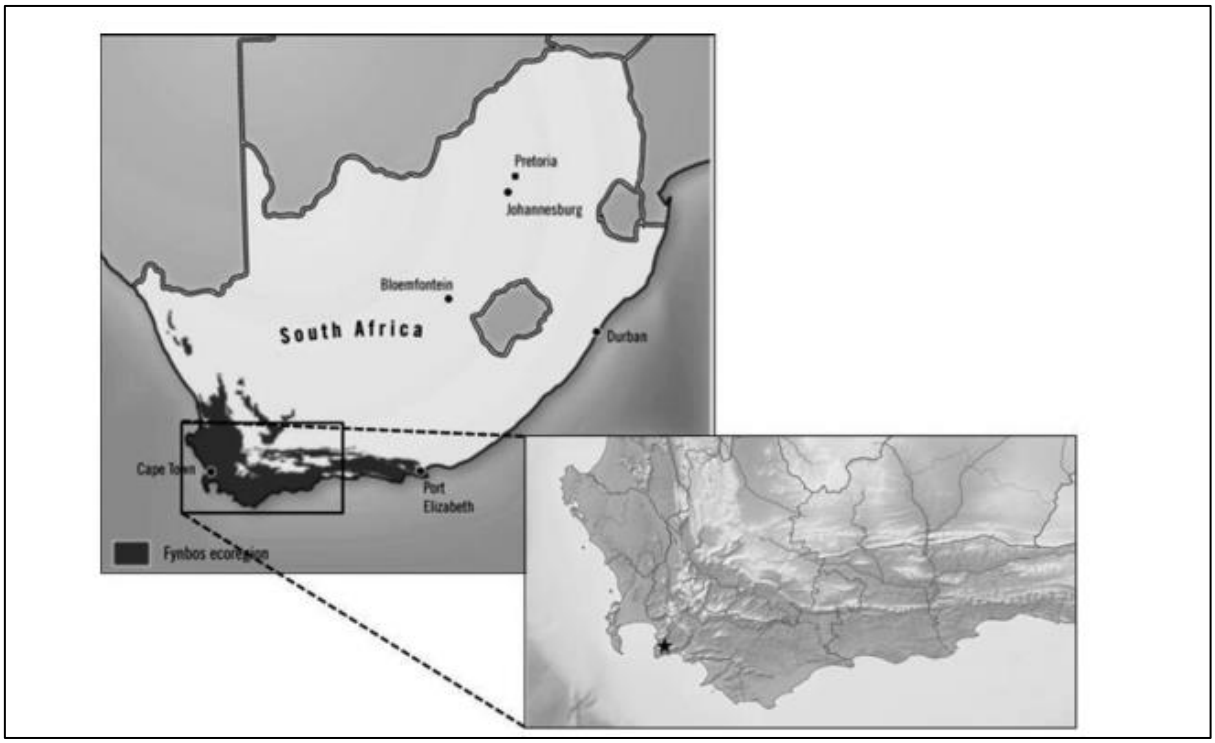


Figure 1.1: **Map of South Africa with the *fynbos* biome further enlarged in grey. The Kogelberg Biosphere Reserve (coordinates: S34°16.489/E019°02.405), is denoted by a black star and indicates the location of sample collection. (Sources: http://www.grida.no/graphicslib/detail/*Fynbos*-ecoregion-in- south-africa_1320;_Ramond *et al.*, 2015)**

The unique plant biodiversity and the possibility of unique microorganisms associated with the *fynbos* plants in this region, led us to explore the viral community in soils from the Kogelberg Biosphere Reserve. The proposed study aims to use metaviromics coupled with high-throughput sequencing approaches to explore the viral diversity in these soils, as well as to use bioprospecting techniques to identify and isolate novel nucleic acid manipulating enzymes that may provide novel products for biotechnology. This study will also aid our understanding of the identity and function of viruses in soil microbiology, since viruses may transfer genes from host to host where they can alter microbe growth and protein expression, and cause mortality of soil microbes (Penadés *et al.*, 2015).

1.2 Overview of viruses

Viruses are regarded to be the most diverse organisms in the world and are considered the most common inhabitants in almost every ecosystem on earth (Paul, 1988) (Koonin *et al.*,

2006). Viruses vary in structure and consist of genetic material (DNA or RNA) surrounded by a protein coat or capsid which protects the genetic information. Furthermore, when the viruses are outside the cell, their nucleic acid is encapsulated by lipid or glycoprotein around the protein coat (Fuhrman and Campbell, 1998). Viruses completely depend on their hosts to reproduce as they do not contain a ribosome and infect their host by transduction and lysogenic conversion (Williamson *et al.*, 2005). Viruses can infect a wide range of living things, from animals and plants to bacteria and archaea (Breitbart and Rohwer, 2005). Bacteriophages or viruses that can infect bacteria have been shown to influence bacterial diversity and species distribution, since they function as obligate intracellular parasites and undergo replication within cells, thereby utilising the host's bacterial replication mechanisms (Louten, 2016) and (Fuhrman, 1999).

Bacteriophages occur abundantly in the biosphere (Ackermann, 2006). Once isolated, bacteriophages are amendable to study using techniques of molecular biology and biotechnology, and are receiving considerable interest for applied biotechnology; such as the use of bacteriophages as antibacterial agents, phage display systems, vehicles for vaccine delivery as well as for diagnostic purposes (Haq *et al.*, 2012).

1.2.1 Classification of bacteriophages

The International Committee on the Taxonomy of Viruses (ICTV), specifically, the Bacterial and Archaeal Viruses Subcommittee (BAVS) has classified a broad number of bacteriophages (ICTV 2017). Seven orders, subdivided into 29 families and 82 unassigned families, were recognised by the ICTV in 2017 (Table 1.1). The *Caudovirales* represents bacterial and archaeal types with head-tail morphologies and is subdivided into 3 families (*Myoviridae, Podoviridae* and *Siphoviridae*). In addition, the other orders include *Bunyavirales* (9 families), *Herpesvirales* (3 families), *Ligamenvirales* (2 families), *Mononegavirales* (9 families), *Nidovirales* (4 families), *Tymovirales* (4 families) and *Picornavirales* (6 families). Furthermore, 85 families have not been assigned to any order. Table 1.1 represents an overview of the 7 orders.

Table 1.1: The latest official release of the ICTV viral taxonomy (2017). There are 8 orders subdivided into 40 families and additional 85 families which have not been assigned to any order (ICTV 2017).

Order	Families	Number of genus and species
Order: *Bunyavirales*	(9 Families)	
	Family: *Feraviridae*	(1 Genus and 1 Species)
	Family: *Fimoviridae*	(1 Genus and 9 Species)
	Family: *Hantaviridae*	(1 Genus and 41 Species)
	Family: *Jonviridae*	(1 Genus and 1 Species)
	Family: *Nairoviridae*	(1 Genus and 12 Species)
	Family: *Peribunyaviridae*	(2 Genera and 52 Species)
	Family: *Phasmaviridae*	(1 Genus and 6 Species)
	Family: *Phenuiviridae*	(4 Genera and 24 Species)
	Family: *Tospoviridae*	(1 Genus and 4 Species)
Order: *Caudovirales*	(3 Families)	
	Family: *Myoviridae*	(6 Subfamilies and 39 Genera not in a Subfamily)
	Family: *Podoviridae*	(3 Subfamilies and 20 Genera not in a Subfamily)
	Family: *Siphoviridae*	(6 Subfamilies and 94 Genera not in a Subfamily)
Order: *Herpesvirales*	(3 Families)	
	Family: *Alloherpesviridae*	(4 Genera and 12 species)
	Family: *Herpesviridae*	(3 Subfamilies and 17 Genera)

	Family: *Malacoherpesviridae*	(2 Genera and 1 Species)
Order: *Ligamenvirales*	(2 Families)	
	Family: *Lipothrixviridae*	(3 Genera and 8 species)
	Family: *Rudiviridae*	(1 Genus and 3 Species)
Order: *Mononegavirales*	(9 Families)	
	Family: *Bornaviridae*	(1 Genus and 8 Species)
	Family: *Filoviridae*	(3 Genera and 7 Species)
	Family: *Mymonaviridae*	(1 Genus and 1 Species)
	Family: *Nyamiviridae*	(3 Genera and 5 Species)
	Family: *Paramyxoviridae*	(7 Genera and 50 Species)
	Family: *Pneumoviridae*	(2 Genera and 5 Species)
	Family: *Rhabdoviridae*	(18 Genera and 131 Species)
	Family: *Sunviridae*	(1 Genus and 1 Species)
	Family: *Unassigned*	(5 Genera and 5 Species)
Order: *Nidovirales*	(4 Families)	
	Family: *Arteriviridae*	(5 Genera and 17 Species)
	Family: *Coronaviridae*	(2 Subfamilies and 4 Genera)
	Family: *Mesoniviridae*	(1 Genus 6 Species)
	Family: *Roniviridae*	(1 Genus and 1 Species)
Order: *Picornavirales*	(6 Families)	
	Family: *Dicistroviridae*	(3 Genera and 15 Species)
	Family: *Iflaviridae*	(1 Genus and 15 Species)
	Family: *Marnaviridae*	(1 Genus and 1 Species)
	Family: *Picornaviridae*	(35 Genera)
	Family: *Secoviridae*	(1 Subfamily and 5 Genera not in a Subfamily)

	Family: Unassigned	(2 Genera and 4 species)
Order: *Tymovirales*	(4 Families)	
	Family: *Alphaflexiviridae*	(7 Genera and 50 Species)
	Family: *Betaflexiviridae*	(2 Subfamilies and 12 Genera)
	Family: *Gammaflexiviridae*	(1 Genus and 1 Species)
	Family: *Tymoviridae*	(4 Genera and 39 Species)
Virus families not assigned to an order	(85 Families)	

Bacteriophage genomes are diverse in terms of their composition and structure (Hulo *et al.*, 2011). These genomes can be linear or circular and have single-or double-stranded DNA and ribonucleic acid (RNA) (Hulo *et al.*, 2011). Furthermore, these genomes may encode as few as four to hundreds of genes and can differ significantly in structure (tailed, polyhedral, filamentous or pleomorphic) (Bertani, 1953; Hulo *et al.*, 2011).

A number of physical and biological criteria have been considered in the classification of viruses (Baltimore, 1971). Viruses may be classified in terms of their genetic composition using the Baltimore system of virus classification (Baltimore, 1971) with RNA or DNA as their genetic information and occurring in double-stranded (ds) or single-stranded (ss) form. Viruses are divided into 7 classes according to the Baltimore system (Table 1.2). Class I consists of dsDNA viruses, class II of ssDNA viruses, class III of dsRNA viruses, class IV of positive (+)-sense ssRNA viruses, class V of negative (-)-sense ssRNA viruses, class VI of RNA reverse transcribing viruses and class VII of DNA reverse transcribing viruses (Baltimore, 1971).

Table 1.2: Classification of viruses in terms of their genetic contents using the Baltimore system of virus classification

Class	Description of genome	Example of virus
I	Double-stranded DNA viruses	Lamda, T4, Herpesvirus, Poxvirus
II	Single-stranded (+) sense DNA viruses	ØX174, Parvoviruses
III	Double-stranded RNA viruses	Ø6, Reoviruses
IV	Single-stranded (+) sense RNA viruses	MS2, Picornaviruses
V	Single-stranded (-) sense RNA viruses	Orthomyxoviruses, Rhabdoviruses
VI	Single-stranded (+) sense RNA viruses	Retroviruses
VII	Double-stranded DNA viruses	Hepadnaviruses

1.2.2 Bacteriophage replication

A bacteriophage life cycle involves either a lytic, lysogenic or combined lytic-lysogenic cycle (Figure 1.2). In the lytic cycle, the phage infects a bacterium and thereafter transforms the bacterium to make more phages. The bacterial cell is lysed and destroyed immediately after the replication and release of the virion (Bertani, 1953) and the progeny virions (viral particles) can find a new host to infect. In virulent bacteriophages like T4, lysis inhibition can occur when there is a high concentration of extracellular phage (Bertani, 1953) and the phage progeny does not lyse out of the cell immediately (Snyder *et al.*, 2015).

In the lysogenic cycle, the viral genome integrates with the host's DNA and replicates along with the host's DNA without lysing the cell. Alternatively, the viral genome may be maintained as a plasmid in the cell, but located outside of the host chromosomal genome (a process called pseudolysogeny) (Payeta and Suttle, 2013). Bacteriophages that utilise the lysogenic cycle are called temperate phages. The prophage is effectively the bacteriophage genetic material which is integrated into the bacterium's genome and is capable of producing phages upon targeted activation (Shao *et al.*, 2016). The lysogenic cycle enables the phage genome to continue to persist and replicate in all of the bacterial daughter cells (Bertani, 1953). At the end of the cycle, the virus remains dormant until the host's condition deteriorates and the prophage becomes active (Bertani, 1953). The phage lambda, which

infects *Escherichia coli*, is an example of a bacteriophage that enters both the lysogenic and the lytic cycles (Snyder *et al.*, 2015).

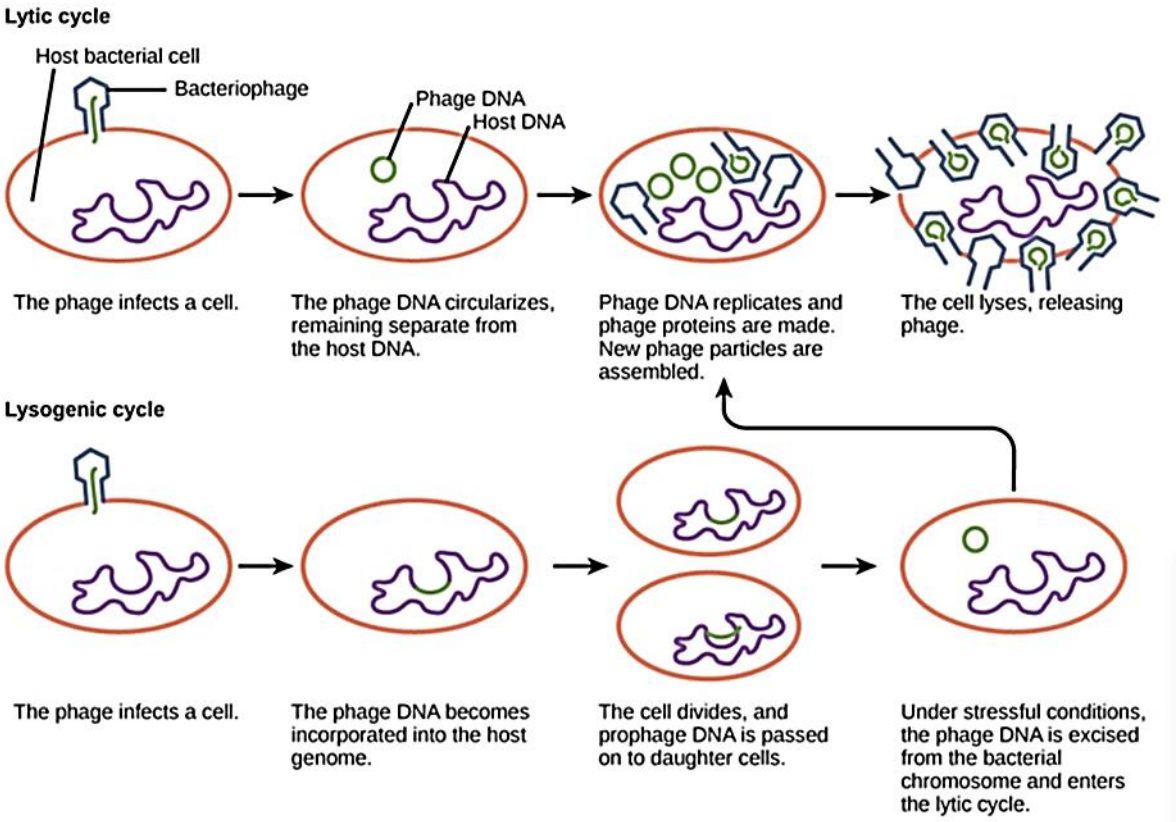


Figure 1.2: The life cycle of a Bacteriophage. The lytic cycle involves the replication of the phage and the lysis of the host cell. The phage's DNA is integrated into the host's genome and passed on to subsequent generations (lysogenic cycle). Because of environmental stresses, the prophage may be induced and thus enters into the lytic cycle (Source: courses.lumenlearning.com) (https://courses.lumenlearning.com/boundless-biology/chapter/virus-infections-and-hosts/)

1.2.3 Phage infection

Phage infection consists of several steps (Figure 1.3). The first step comprises attachment, where bacteriophages attach to specific receptors located on the surface of the host cell (Figure 1.3). The second step involves penetration, where the bacteriophage injects its nucleic acid into the bacterial cell (Figure 1.3). The third step is the transcription of the phage's DNA, where host cell enzymes are used to make additional DNA that is transcribed to messenger RNA (mRNA) (Figure 1.3). Further steps involve a replication mechanism

(Figure 1.3). This process entails the synthesis of proteins and nucleic acids where bacterial ribosomes start translating viral mRNA into protein (Figure 1.3). Thereafter, the virions assemble where the construction of new virus particles takes place (Figure 1.3). In the final step, the virions are released where bacteriophages could be set free through cell lysis, extrusion or budding (Weinbauer, 2004).

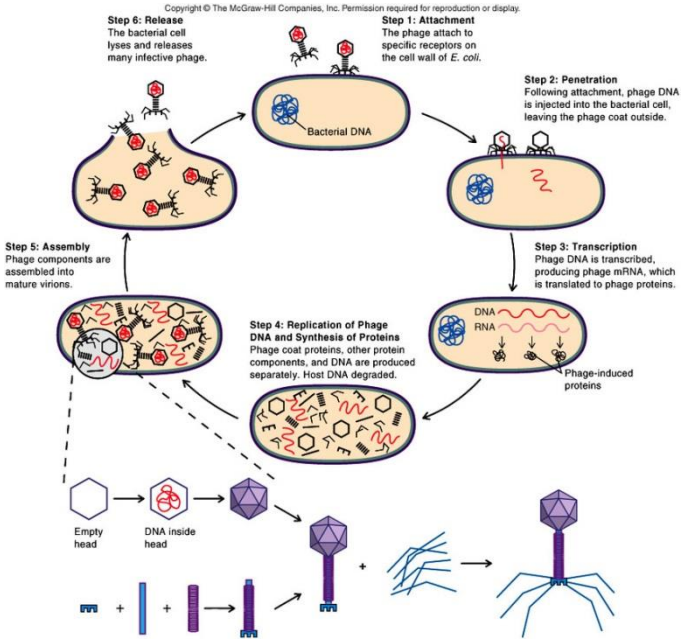


Figure 1.3: Steps involved in phage infection (Source: The McGraw-Hill companies. Inc.) (https://emilybio11.weebly.com /microbiology.html).

1.3 Molecular approaches to bacteriophage ecology

The first approach that led to the discovery of an abundance of bacteriophages in the environment involved the use of electron microscopy to view water samples from natural aquatic environments (Bergh *et al.*, 1989). Later studies also reveal the abundance of bacteriophages in seawater using epifluorescence microscopy (Marie *et al.*, 1999), and tailed phages dominated the marine virus populations (Chibani-Chennoufi *et al.*, 2004). Williamson *et al.*, (2005) studied Delaware six soil types using Epifluorescence microscopy (EFM) and transmission electron microscopy (TEM) and reported higher abundance of bacteriophages in the moist and organic matter rich forested soils than the dry and organic matter poor agricultural soils.

Molecular biology approaches to assessing bacteriophage diversity are restriction length polymorphism (RFLP), denaturing gradient gel electrophoresis (DGGE), suppression subtractive hybridisation (SSH), representational difference analysis (RDA) (Mokili *et al.*, 2012) and random polymerase chain reaction amplification of polymorphic DNA (RAPD-PCR) (Srinivasiah *et al*., 2013). The limitations associated with these methods include poor sensitivity at a low viral burden or when there is little DNA sequence variation between the viral samples (Delwart, 2007). The lack of common genetic markers for viruses such as the 16S rRNA or 18S rRNA (Doolittle, 1988) that are essential for protein synthesis and universal to all microbes, limits the overall insight into bacteriophage diversity (Breitbart, Miyake, *et al.*, 2004; Dorigo *et al.*, 2004; Wilhelm and Matteson, 2008). Nonetheless, there are conserved signature genes for specific viral groups which can be used as gene markers to assess environmental viral diversity (Adriaenssens and Cowan, 2014). Combining such strategies with high-throughput sequencing can provide information about the abundance of viruses and can provide insight into the functional roles of viruses in the soil ecosystem (Zablocki *et al.*, 2016). High-throughput sequencing methods applied to samples containing many viral types is termed viral metagenomics or metaviromics (Santos *et al.*, 2010).

The term viral metagenome (also called metavirome) refers to the total viral gene pool in a given environment. Metaviromics are the techniques used to study viral genetic material sourced directly from the environment (Edwards and Rohwer, 2005). Introducing metaviromics has revolutionised the field of environmental viral ecology by enabling the viral communities to be fully explored using a variety of environmental samples without reliance on culture-dependent techniques (Rosario and Breitbart, 2011).

Metaviromics involves random shotgun sequencing of fragments from a pool of viral genomes in any given community (Breitbart *et al.*, 2002). Metaviromics has enabled a much clearer picture of bacteriophage diversity and has revealed previously unknown viral diversity (Edwards and Rohwer, 2005; Willner *et al.*, 2009; Rosario and Breitbart, 2011; Mokili *et al.*, 2012). The study of Breitbart *et al.*, 2002 revealed many viral sequnces with similarities to known viruses, particularly all major families of tailed bacteriophages, but over 80% of viral sequences were unique and did not match the sequences of public genomic databases at the time. Similar bacteriophage signature sequences were found in freshwater, sediments and terrestrial samples (Miyake, *et al.*, 2004). Several subsequent metaviromic

studies of environmental samples have revealed previously unknown diversity of viral genomes (Hendrix *et al.*, 1999; Breitbart *et al.*, 2002; Breitbart, Felts, *et al.*, 2004; Mann, 2005; Rohwer and Barott, 2013; Watkins *et al.*, 2016; Koonin and Dolja, 2018)

1.4 The metavirome concept

1.4.1 A brief history of metagenomics

Historically, culture-dependent techniques were used in the analysis of microorganisms. However, estimates showed that these techniques allowed for less than 1% of the diverse microbes from environmental samples to be grown, isolated, and studied (Rappé and Giovannoni, 2003). To circumvent the challenges associated with culture-dependent techniques, an alternative technique, known as metagenomics, was developed (Schloss and Handelsman, 2003; Riesenfeld *et al.*, 2004). Using this technique, culture-independent methods could be employed directly to environmental samples from microbial communities (Schloss and Handelsman, 2003; Riesenfeld *et al.*, 2004; Thomas *et al.*, 2012).

Previously, ribosomal RNA found within all bacteria and archaea was used for analysing natural microbial populations independently from culture (Pace *et al.*, 1986). This method was also used to determine bacterial phylogenies (Woese *et al.*, 1990). The 16S rRNA analysis was first implemented in 1991 by Schmidt *et al.* (1991). The authors obtained DNA samples from a marine picoplankton community and analysed them for the presence of bacteria (Schmidt *et al.*, 1991). This led to numerous investigations on bacterial diversity in environmental samples (Béjà *et al.*, 2002; Hugenholtz *et al.*, 2002; Venter *et al.*, 2004; Gasol and Moran, 2015; Bobrova *et al.*, 2016; Lee *et al.*, 2016).

Advances in sequencing technology have improved genomic studies beyond 16S rRNA analysis to include complete sequencing of all DNA isolated from an environmental sample (Ju and Zhang, 2015). A pioneer of this approach was Venter *et al.* (2004) who studied the metagenome of the Sargasso Sea and demonstrated the utility of this approach in its ability to be effectively applied to the discovery of genes and microbial species and that could also aid our understanding of the marine environment (Venter *et al.*, 2004).

Along with the success of sequence-based metagenomics studies, activity-based or "functional" analysis of metagenomic libraries has rapidly developed (Schoenfeld *et al.*, 2010). This development has allowed for the expression and characterisation of genes within

environmental DNA libraries (Lam *et al.*, 2015); and has led to metagenomic studies being broadly categorised as either sequence-based or activity-based (functional) (Handelsman, 2004).

Screening of metaviromics libraries have been used effectively for the identification and characterisation of novel genes. Previous study by Schmitz, (2010) identified 26 actively expressed lysins through viral metagenomics. Their study represented one of the first functional screens of lysins from a viral metagenomic sample. Another study described functional metaviromics study as viral metapopulations which are isolated from natural thermal environments (Moser *et al.*, 2012), by identifying and characterizing replication operons and developing the gene products as thermostable enzymes used for nucleic acid amplification and sequencing (Moser *et al.*, 2012). Schoenfeld *et al.*, (2010)'s review discussed the use of metaviromics as a tool for new enzymes discovery by focusing on how to improve sampling and recombinant DNA cloning methods, functional and genomics-based screens, and expression systems, in order to accelerate discovery of new enzymes and other viral proteins for use in biotechnology (Schoenfeld *et al.*, 2010). These previous functional metaviromics studies, therefore, highlight an opportunity to use functional metaviromics to mine relevant enzymes to use for biotechnology or molecular biology purposes (Schmitz *et al.*, 2010)

1.4.2 A brief history of metaviromics

The first metaviromics of an environmental sample was carried out by Breitbart *et al.* (2002). A combination of differential filtration and density-dependent gradient centrifugation was employed to isolate the metavirome DNA (Breitbart *et al.*, 2002). Partial short-gun sequencing was used to analyse uncultured viral communities from marine water columns collected near the shores of the oceans (Breitbart *et al.*, 2002). Thereafter, a metaviromic DNA library was created by utilizing a linker-amplified shotgun libraries (LASLs) approach (Breitbart *et al.*, 2002). The DNA was mechanically sheared for the creation of the metaviromic library, in order to capture viral nucleic acids irrespective of the presence of specific nucleic acid sequences and restriction enzyme sites (Breitbart *et al.*, 2002). The results highlighted the possibility to sequence the entire genome of an uncultivated viral community from marine samples (Breitbart *et al.*, 2002).

Viral enzymes were important for the early development of molecular biology and biotechnology and have become essential tools (Schoenfeld *et al.*, 2011). However, the viral enzymes found, to date, represent only a fraction of the potential diversity that actually exists (Schoenfeld *et al.*, 2011). The main technical challenges to the discovery of novel viral enzymes are related to the cultivation of new viral host systems (Schoenfeld *et al.*, 2011). Traditional approaches demand the suitability of both the host and the virus for cultivation. Often the host cannot successfully form lawns and viruses cannot successfully form plaques, which leads to the failure of isolating the new virus (Schoenfeld *et al.*, 2011). Once the viruses are isolated, major experimental work is needed to define factors such as multiplicity of infection, burst size and infection kinetics. These are vital parameters to facilitate the detection of viral enzymes. Even with the optimisation of the induction at particular time points, subsequently after infection, there is often a degree of difficulty in differentiating between the viral enzymes and those from the host cell (Schoenfeld *et al.*, 2010).

The introduction of next-generation sequencing (NGS) technologies has substantially improved the unit cost and throughput of nucleic acid sequencing (Angly *et al.*, 2006). This has enabled the development of metaviromics approaches to study viral community structure and for the discovery of novel genes and viruses (Rosario and Breitbart, 2011). To date, there are only a handful of metaviromics studies. Therefore, the proposed study, which is based on using metaviromic techniques to explore the viral community of the Kogelberg Biosphere Reserve soil sample, will be of great importance to advance this field.

1.4.3 Metavirome analysis

Metavirome analysis is composed of the following steps: (1) sample preparation, (2) metavirome library construction (3) functional and/or sequence-based screening (4), high-throughput sequencing, and (5) bioinformatics analysis (Figure 1.4) (Mokili *et al.*, 2012).

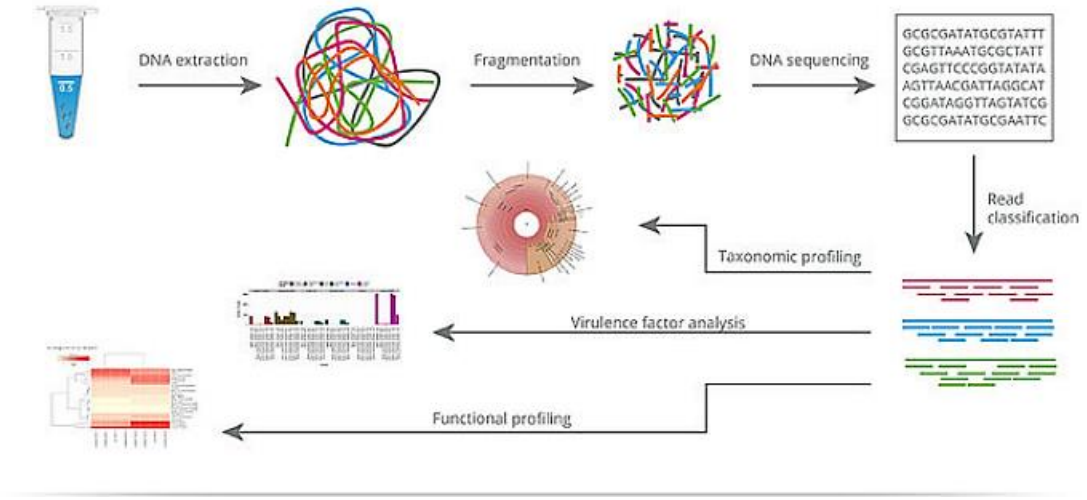


Figure 1.4: General steps in a metaviromics strategy to investigate viral communities in environmental samples (Source: Biodesign Collection by Greg Sieber) (https://i.pinimg.com/736x/6c/1a/1b/ 6c1a1baf2906c6c098c4390e47999944.jpg).

1.4.4 Sample preparation

1.4.4.1 Isolation of viruses

The first step is to successfully isolate total viral community genomes from the environment by carefully separating the viral nucleic acid from that of eukaryotic and prokaryotic origin (Casas and Rohwer, 2007). This step often results in the significant loss of viral nucleic acid, because free viruses form a small fraction of the environmental sample (Thurber *et al.*, 2009) and physical enhancement techniques that are based on size selection only enrich for free viruses and not intracellular viruses (viroids) or viral genomes integrated into host cell genomes (viral prophage) (Hall *et al.*, 2014).

Environmental samples are filtered to remove larger microbial cells and to concentrate the sample by various methods; such as: ultracentrifugation, tangential flow filtration (TFF) and depth ultrafiltration (UF) (Paul *et al.*, 1991; Casas and Rohwer, 2007; Thurber *et al.*, 2009). Thereafter, density centrifugation in sucrose gradients or caesium chloride (CsCl) can be used for the separation of intact viral particles (Edwards and Rohwer, 2005). Other techniques such as polyethylene glycol (PEG) precipitation methods have been implemented as alternatives to prevent the disintegration of CsCl sensitive viruses, and the loss of very small or large viruses during filtration (Casas and Rohwer, 2007).

The quality of a viral sample produced using various purification methods can be assessed by using various techniques; such as: microscopy techniques (Thurber *et al.*, 2009), pulsed-field gel electrophoresis (Steward, 2001), epifluorescence microscopy (Breitbart *et al.*, 2002), quantitative flow cytometry (Brussaard, 2004), and quantitative PCR of 16S and 18S rDNA to detect contamination with bacterial genomes.

1.4.4.2 Metavirome DNA extraction

There is a wide range of nucleic acid extraction methods and the diversity of metavirome nucleic acid is influenced by biases inherent in these methods (Head *et al.*, 1998; Frostegård *et al.*, 1999; Martin-Laurent *et al.*, 2001; LaMontagne *et al.*, 2002; Carrigg *et al.*, 2007). Thus, this could potentially lead to the underestimation of the total microbial or viral diversity in the studied environment irrespective of the method employed in calculating the species or operational taxonomic unit (OTU) diversities (Delmont *et al.*, 2011; Bag *et al.*, 2016).

Environmental samples typically yield relatively small amounts of viral DNA (Edwards and Rohwer, 2005). The DNA is usually quantified by ultraviolet (UV) absorption (e.g. Nanodrop), intercalating dyes (e.g. Qubit; Invitrogen, SYBR Green), 5′ hydrolysis probes (e.g. Taqman®) coupled with real-time quantitative PCR (qPCR; Kapa Biosystems) or droplet digital emulsion PCRs (ddPCR; Bio-Rad). Additionally, the concentration limitations can be overcome by the extraction of DNA from large amounts of environmental sample or the use of library preparation methods from low input DNA such as Illumina's NextEra device (Tan and Yiap, 2009). For example, if an aqueous environment is under study, the filtering of several hundred litres of water can be sufficient to obtain enough DNA for cloning and sequencing (Rohwer and Edwards, 2002).

Depending on the amount of viral DNA produced, whole-genome amplification (WGA) techniques such as multiple displacement amplification (MDA) are commonly used prior to preparing the metaviromic library studies (Edwards and Rohwer, 2005; Brum and Sullivan, 2015). The limitation of the MDA step is the bias of preferentially amplifying small circular ssDNA templates (Yilmaz *et al.*, 2010; Kim and Bae, 2011; Marine *et al.*, 2014). Sequence-independent single primer amplification (SISPA) is another commonly used random amplification approach in combination with NGS (Reyes and Kim, 1991). However, Rosseel

et al. (2013), combined random amplification (SISPA) with NGS and reported biased amplification of certain virus groups (Rosseel *et al.*, 2013). Alternative methods such as linker amplification (LA) or tagmentation (TAG) (Duhaime *et al.*, 2012) suffer bias as these methods are highly selective against ssDNA templates (Kim *et al.*, 2011).

1.4.5 Metavirome library construction

Library construction remains the main metaviromics technique. It begins with isolating environmental DNA, cloning the DNA fragments into vectors, packaging the vectors into lambda phage followed by the transduction into *E. coli* (Figure 1.5). Vectors appropriate for cloning DNA fragments larger than 10 kb namely, cosmids (25–35 kb), fosmids (25–40 kb), or bacterial artificial chromosomes (BACs) (100–200 kb) can be utilised (Rondon and Al, 2000; Kim *et al.*, 2004). Nevertheless, library construction is time consuming and costly. Fosmid libraries exhibit bias (such as GC bias), which are represented in the library and can affect conclusions derived from analysis of the clone libraries and the ability to identify novel gene. This could be due to fewer hydrogen bonds in AT-rich sequences resulting in the fragmentation and subsequent loss of AT-rich sequences during the cloning process. Transcriptional activity of the cloned DNA and toxicity from expressed genes can possibly be another reason for bias in libraries (Lam *et al.*, 2015). Therefore, to overcome the library construction challenges and biases, direct sequencing that does not require the creation of a metaviromics library can be used (Fantle *et al.*, 2003). However, the limiting factor of the direct sequencing are the generation of short read lengths and the constant confidence in base calling in sequence reads. Other limiting factors of direct sequencing include the high (e.g. more than gigabase per run) amount of data generated by the NGS systems in the form of short reads. The high amount of short reads pose challenges to software developers and more efficient computer algorithms (Ansorge, 2009).

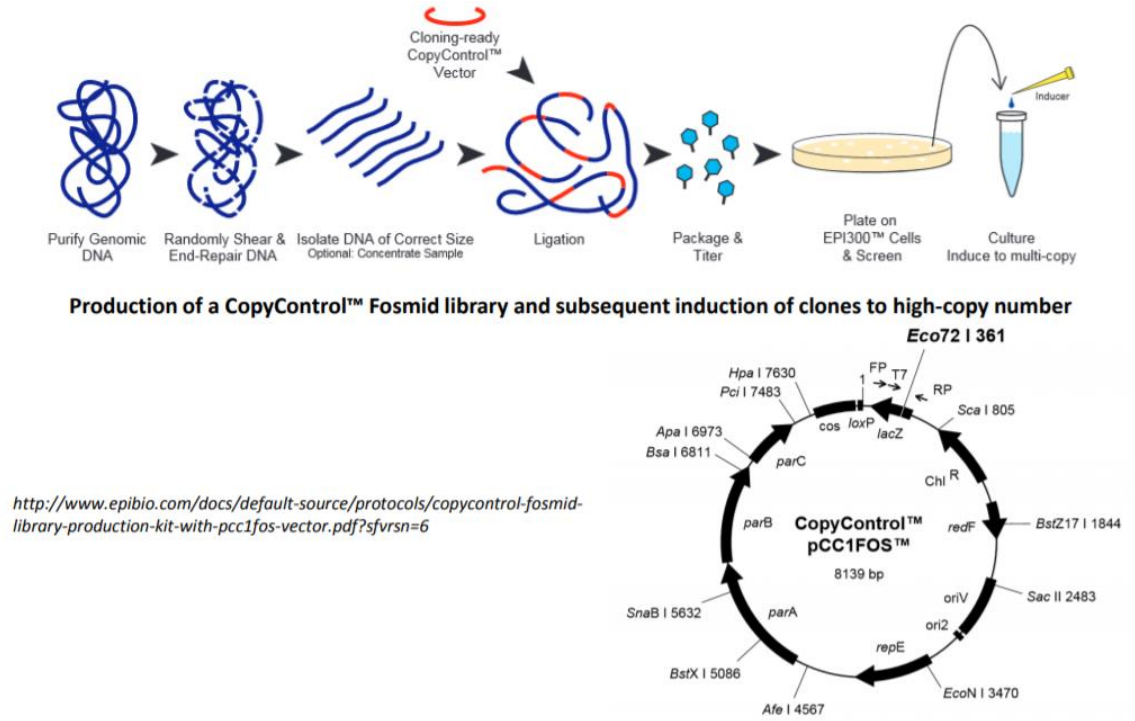


Figure 1.5: Fosmid library construction using pCC1FOS (Epicentre Biotechnologies).

1.4.6 Metavirome high-throughput sequencing

Sequence-based metaviromics relies on Next-generation sequencing (NGS) technologies for the analysis of the genetic makeup of viral communities (Fantle *et al.*, 2003). NGS is a broad term involving a number of modern sequencing technologies that allow for cost-effective DNA sequencing when compared to the traditional Sanger sequencing approach (Buermans and den Dunnen, 2014). The advancement of high-throughput sequencing technologies over the last decade has dramatically increased the sequencing speed and decreased the cost per base (Lam *et al.*, 2012). The most common forms of NGS sequencing used for metaviromics are Illumina, Ion Torrent and Ion Proton sequencing. The limitation to these techniques, in comparison to Sanger sequencing, is due to the generation of shorter sequence fragments (Liu *et al.*, 2012). The Ion Torrent Personal Genome Machine (PGM) System (only single reads) and 454 pyrosequencing produces 400 base pair (bp) reads, whereas the Illumina MiSeq produces approximately 400bp paired reads, while SOLiD produces 25-75bp reads (Rodrigue *et al.*, 2010).

Long-read sequencing produces, on average, fragment lengths of 10,000bp up to a maximum of 100,000bp. These read lengths can now be obtained through the use of more recent

sequencing technology advancements such as Oxford Nanopore MinIO and Pacific Biosciences (Giordano *et al.*, 2017). The assembled sequences achieved from these sequencing platforms result in higher contig continuity and more complete long fragments of the genome (Giordano *et al.*, 2017). These sequencing platforms are thus capable of extending paths into problematic or repetitive regions of genomes and play an important role in overcoming the high fragmentation of NGS-based assemblies (Giordano *et al.*, 2017).

A wide range of environments have been investigated using metaviromics. These environments include sea and aquatic environments (Angly *et al.*, 2006; Culley *et al.*, 2006; Dinsdale *et al.*, 2008; Hurwitz *et al.*, 2013), extreme thermal environments (e.g. hot springs and marine thermal vents) (Schoenfeld *et al.*, 2008; Gudbergsdóttir *et al.*, 2015) and soil (Fierer *et al.*, 2007; Williamson *et al.*, 2007; Fancello *et al.*, 2013). Previous study of prairie, desert, and rainforest soils used metavirome approaches to assess diversity of the viral communities in these soil biomes (Fierer *et al.*, 2007). Their comparative study revealed that soil viruses are taxonomically diverse and distinct when compared to viral communities from other environments that have used metaviromic approaches (Fierer *et al.*, 2007). The abundance and diversity of viruses in six Delaware soils showed that the soil virus communities were dominated by bacteriophages (Williamson *et al.*, 2005).

1.4.7 Sequence-based screening of metaviromes

Identification of novel genes with desired function can be achieved by cloning, and screening by PCR nucleic acid hybridisation-based methods or more recently, sequence-based screening (Culligan *et al.*, 2014). Gene motifs that characterise genes of desirable function (Altschul *et al.*, 1990) can be used to design consensus primers or probes to isolate novel genes from metaviromic DNA (Wooley *et al.*, 2010). If only partial gene fragments are recover from the metaviromic library, genome walking can be performed (Culligan *et al.*, 2014), but this approach relies on the specificity and coverage of the primer or probe used (Culligan *et al.*, 2014; Rashid and Stingl, 2015). Nevertheless, various novel enzymes have successfully been discovered using these classic techniques (Knietsch *et al.*, 2003; Simon and Daniel, 2011).

Since the advent of NGS technologies, sequence-based screening of metagenomes is now commonly used to screen for novel genes (Dinsdale *et al.*, 2008; Rajendhran and

Gunasekaran, 2008; Kodzius and Gojobori, 2015; Madhavan and Sindhu, 2017). Sequence-based metagenomics is not without limitations, despite its advantages and usefulness. A significant challenge is the size and the complexity of the sequence data, posing difficulties in the bioinformatics analysis of the data (Rashid and Stingl, 2015). Another challenge lies in the assembling of genomes from many genotypes with significant sequence similarity, which is typically a problem of having short sequence reads and this request the assemblage of contigs (Rashid and Stingl, 2015). An additional limitation is the occurrence of undesirable host DNA or in the metaviromic sequencing data. Lastly, the cost of generating metagenomes in the case of diverse communities, such as soil environments is relatively high in comparison to the traditional cloning, screening and sequencing strategy (Rashid and Stingl, 2015). However, high-throughput screening in combination with both function-based and sequence-based methods can detect novel microbial enzymes at a higher frequency when compared to conventional screening and this often justifies the extra costs involved (Uchiyama and Miyazaki, 2009).

1.4.8 Function-based screening of metaviromes

The metaviromics techniques have allowed researchers to gain access to the genomic diversity of virus communities. However, given the DNA sequence, there are still significant limitations and barriers to confidently identify gene function. These include the fact that several genes share similar sequences or motifs and genes may be dormant as pseudogenes and not expressed to produce a functional protein. In order to avoid this limitation of DNA sequencing approaches to metaviromics, functional metaviromics clone and express viral DNA in surrogate hosts and screen for function and biological activity (Cheng *et al.*, 2017). Over the past years, many novel enzymes have been recovered from metavirome libraries; including DNA polymerases (Pols), resolvases (e.g. T4 endonuclease VII and T7 endonuclease I) and T4 RNA ligase (Schmitz *et al.*, 2010; Schoenfeld *et al.*, 2010). However, developing more activity-based screening methods and scaling up the throughput of the available approaches are key strategies to identify novel enzymes in metavirome samples (Schoenfeld *et al.*, 2010).

Functional screening enables the identification of new enzymes from various environments in libraries without reliance on sequence homology with previously isolated genes (Uchiyama and Miyazaki, 2009) that is needed for primer or probe based screening methods.

Function-based screening for genes with desirable traits is a straightforward way to isolate genes by the direct detection of phenotypes, complementation of heterologous genes and the induction of gene expression (Felczykowska *et al.*, 2015). However, this approach is more laborious than sequence-based screening procedures (Ferrer *et al.*, 2005; Fernández-Álvaro *et al.*, 2010).

Agar plate screening is used in the function-based screening method for the identification of colonies which express enzymes with desired properties (Leemhuis *et al.*, 2009). In agar plate screening, the colonies are plated on medium containing the enzyme substrate of interest, and positive clones expressing enzymes acting on the substrate are identified by a zone of clearance or fluorescence (Leemhuis *et al.*, 2009). The limitation of this screening method is the inherent biases of cloning and expression in a suitable host and the difficulties in the high-throughput screening (Leemhuis *et al.*, 2009). A microtiter plate-based approach can also be used to increase the throughput of expression screening of libraries to identify desirable enzyme activities (Leemhuis *et al.*, 2009). Either clones or pools of clones can be screening for expressing enzyme of interest using a medium containing a substrate detectable by spectrophotometry or fluorometry (Leemhuis *et al.*, 2009; Uchiyama and Miyazaki, 2009; Culligan *et al.*, 2014).

Complementation screening is an efficient strategy for identifying targeted genes (Fantle *et al.*, 2003). Complementation screening employs a host with a specific functional deficiency that is restored through cloning of extraneous DNA. The complementation of cold-sensitive *E. coli* mutants for example, has the ability to allow for difficult and rapid screening of metagenome libraries for novel enzymes (Simon *et al.*, 2009). The study by Simon *et al.* (2009) used *Escherichia coli polA* mutant strain to screen the metagenome libraries for the presence of DNA polymerase-encoding genes through complementation. The recovered DNA polymerases had novel sequences compared to existing known polymerases, thereby demonstrating the value of this approach (Simon *et al.*, 2009).

Metaviromics has revealed an enormous diversity of novel virus genomes harbouring new genes which will potentially provide novel enzymes for molecular biology (Blondal *et al.*, 2003; Schmitz *et al.*, 2010; Schoenfeld *et al.*, 2011; Moser *et al.*, 2012). Functional characterisation of metavirome sequences is an important step towards the understanding of viral ecology and for the identification of new viral enzymes (Schoenfeld *et al.*, 2011). To

date, metaviromics studies have remained focused on sequence-based screening. In this regard, functional screening of metaviromes could offer a large source of untapped novel recombinant molecules. One study by Schmitz *et al.* (2010) devised a novel functional screening for the cloning of lytic enzymes from metaviromes and provided a general model for lysin identification through metaviromics. These studies highlight the potential of metaviromics for cloning of enzymes of biotechnology, research or academic value.

1.4.9 Bioinformatics analysis

Bioinformatics analysis of metavirome NGS sequence data involves 3 main steps: (1) read trimming or removal based on quality scores (such as Phred score), (2) *de novo* assembly of reads into contiguous sequences known as contigs; and (3) taxonomic and functional assignments of reads or contigs (Ewing *et al.*, 2005; Wooley *et al.*, 2010). The metavirome identification depends on similarity searches against a database using the Basic Local Alignment Search Tool (BLAST) (Altschul *et al.*, 1990). Comprehensive databases such as GenBank or servers containing only viral sequences such as ProViDE, VIROME, MEGAN, VirusHunter, MetaVir, VirSorter may be used (Rose *et al.*, 2016).

ProViDE assigns viruses at different taxonomic levels by using virus-specific alignment parameters and thresholds matched to a protein database from BLAST (Ghosh *et al.*, 2011). MEGAN is a generally applicable metagenomics classifier, which uses BLAST results to infer the lowest common ancestor (LCA) for a given sequence and provides functional analyses through a graphical interface (Huson *et al.*, 2007; Huson *et al.*, 2011; Rose *et al.*, 2016). VirusHunter is a Perl script-based tool which uses BLAST prior to assembly for viral identification (Zhao *et al.*, 2013). VirSorter is used to find virus contigs in microbial datasets and thereby identify prophages and viruses through comparison to custom datasets (Roux *et al.*, 2016).

VIROME is an automated web-based application which is designed to explore metagenome sequence data from viral assemblages which occur within a number of different environmental contexts (Wommack *et al.*, 2012). VIROME uses several databases such as ribosomal databases and UniRef100 and MetaGenomeOnline for ORF classification. This server performs a quality check, an assessment of bacterial or eukaryotic DNA

contamination and provides taxonomic and functional assignments of reads or contigs (Wommack *et al.*, 2012).

MetaVir is an automated web application designed to identify viral sequences by predicting coding sequences and analysing them with BLASTp searches against the Viral RefSeq protein database (Roux, Tournayre, *et al.*, 2014). The limitations of this pipeline are the restricted number of reads that can be analysed at one time and the tool is time consuming (Roux, Tournayre, *et al.*, 2014). It performs taxonomic and functional assignments, and statistical and phylogenetic analysis of reads or contigs (Roux, Tournayre, *et al.*, 2014). Comparisons with other metaviromes deposited on the server obtained from several environments can be performed by this server (Roux, Tournayre, *et al.*, 2014).

MG-RAST permits the QC of reads, identification of rDNA sequences, taxonomic and functional assignments of reads or contigs by BLAST comparison with multiple public databases (GenBank, SEED, IMG, UniProt, KEGG and eggNOGs) (Aziz *et al.*, 2008; Meyer *et al.*, 2008). The limitation of this MG-RAST is due to high numbers of taxonomic misannotations of viral sequences identified as bacterial in origin (Aziz *et al.*, 2008; Meyer *et al.*, 2008). However, MG-RAST is not exclusively designed for metaviromic data analysis (Aziz *et al.*, 2008; Meyer *et al.*, 2008).

1.5 Metaviromics approach to bioprospecting of nucleic acid manipulating enzymes

Advancement in metaviromics and the tools available for the discovery of enzymes with unique activities and novel functions has led to the prospect of discovering novel nucleic acid manipulating enzymes. These nucleic acid manipulating enzymes include polymerases, ligases, nucleases, phosphatases, methylases, and topoisomerases (Rittié and Perbal, 2008). These nucleic acid manipulating enzymes are used in the laboratory to modify DNA (and RNA) in defined ways. The application of these enzymes includes propagation, ligation, digestion, phosphorylation and methylation (Rittié and Perbal, 2008). Metaviromics related approaches have been utilised successfully for the isolation of novel enzymes (Niehaus *et al.*, 1999; Lorenz *et al.*, 2002; Blondal *et al.*, 2003; Ferrer *et al.*, 2009, Li *et al.*, 2012; Xiang *et al.*, 2014; Alma'abadi *et al.*, 2015; Gonçalves *et al.*, 2015; Guazzaroni *et al.*, 2015; Ferrer *et al*., 2016; Littlechild, 2016). Research by Booysen and Booysen, (2011) identified and further characterised an NAD+-dependent DNA ligase from Antarctic metagenomics

library. Blondal *et al.* (2003) discovered isolated and characterized an RNA ligase, a thermostable homolog of T4 RNA ligase 1. Moser *et al.*, (2012) identified putative thermostable viral polymerases while using metaviromic approaches. Similar approaches were used to identify family A DNA polymerases in the study by Schmidt *et al.*, (2014).

1.5.1 The commercial market for nucleic acid manipulating enzymes

The global molecular biology enzyme market is highly consolidated and competitive. The demand for these enzymes is increasingly driven by a growing need for novel types which are highly pure reagents for accurate results and facilitate rapid and efficient workflows. In addition, this enzyme market is driven by the overall increase in the research and development spending in the biotechnology sector. The total market value of nucleic acid manipulating enzymes reached $1,2 billion in 2013 and it is estimated to reach $1,8 billion by 2018, growing at a compounded annual growth rate (CAGR) of 8.33% in this period (Table 1.3). The top 4 company players who hold a combined share of over 87% of this market are New England Biolabs, Sigma-Aldrich, Takara Bio, and Thermo Fisher Scientific (The global molecular biology enzymes and kits & reagents market, 2017- 2022). The use of enzymes for PCR applications accounted for the highest sales figures for the global molecular biology enzyme portfolio of kits and reagents on the market in 2013. This market was valued at $1 million in 2013 and is expected to reach $1, 8 million by 2018, growing at a CAGR of 13.03%. Furthermore, sequencing is expected to be the fastest growing application segment valued at $0.5 million in 2013 and is estimated to reach $1,4 million by 2018, growing at a CAGR of 18.86% (Table 3) (The global molecular biology enzymes and kits & reagents market, 2017- 2022). Thus, our proposed study aims to bioprospect nucleic acid manipulating enzymes as an approach towards an entry into accessing the global molecular biology enzyme market.

Table 1.3: Overview of the total market value, highest application segments and highest product segments from the molecular biology enzyme, kit & reagent market report. This is a global forecast from 2013 to 2018. (Source: The global molecular biology enzymes and kits & reagents market, 2017- 2022, http://www.marketsandmarkets.com/Market-Reports/molecular-biology-enzymes-kits-reagents-market-164131709.html)

Market	Group	2013 ($)Million	2018 ($)Million	CAGR %
Total Market value	Total	1200	1800	8.33
Highest application segments	PCR	1.0	1.9	13.03
	Sequence	0.5	1.0	18.86
Highest product segments	Polymerase	0.3	0.5	11.99
Table	Ligases	0.2	0.3	10.00
	Nucleases			
	Restriction endonucleases	0.1	0.2	6.01
	Reverse transcriptase	0.1	0.2	6.92
	Other enzymes	0.08	0.1	5.92

1.5.2 Classification of nucleic acid manipulating enzymes and their total market value

Nucleic acid manipulating enzymes are used in recombinant DNA technology to genetically modify organisms such as bacteria, viruses, plants and animals (Rittié and Perbal, 2008; Struhl, 2009). This is carried out through the construction of a recombinant DNA molecule for cloning and propagation in the host (Struhl, 2009). Numerous enzymes are required for the manipulation process including nucleases, ligases and polymerases (Errol, 2003). The nucleic acid manipulating enzymes that perform these specific functions are further elaborated upon in this section.

1.5.2.1 Polymerases

DNA polymerases catalyse the formation of polymers made by the assembly of multiple deoxyribo-nucleotides triphosphate structural units (dNTPs) (Rittié and Perbal, 2008). These enzymes play an important role in DNA replication and repair (Rittié and Perbal, 2008). There are many different types of polymerases which offer a large degree of flexibility in terms of reaction conditions and catalytic specificity (Hamilton *et al.*, 2001). The largest product segment of the global molecular biology enzyme market portfolio includes T4 DNA polymerases, *Taq* DNA polymerases and High-Fidelity DNA polymerases (The global molecular biology enzymes and kits & reagents market, 2017- 2022). Polymerase enzymes are used for PCR, DNA sequencing, DNA labelling and other procedures that are key components in molecular biology research (Yamagami *et al.*, 2015).

DNA polymerases are grouped into 7 different families (A, B, C, D, X, Y, and RT) based on their sequence similarities and phylogenetic relationships (Yamagami *et al.*, 2015). DNA polymerase family A includes the DNA polymerase 1 (pol 1) enzyme encoded by the *polA* gene (Zhu and Ito, 1994). DNA polymerase 1 is involved in the recombination, repair and replication of DNA (Hamilton *et al.*, 2001). DNA polymerase 1 contains 3 different domains a 5'-3' exonuclease domain at the N-terminus, a central proofreading 3'-5' exonuclease domain and a polymerase domain at the C terminus of the enzyme (Zhu and Ito, 1994). The function of DNA polymerase 1 includes 4 enzymatic activities known as (1) a 5'-3' DNA-dependent DNA polymerase activity which requires a 3' primer site and a template strand, (2) a 3'-5' exonuclease activity that mediates proofreading, (3) a 5'-3' exonuclease activity which mediates nick translation during DNA repair and (4) a 5'-3' RNA-dependent DNA polymerase activity (Ricchetti and Buc, 1993). The 5'-3' exonuclease activity makes DNA polymerase 1 undesirable for many molecular biology research applications (Hamilton *et al.*, 2001).

1.5.2.2 Ligases

Ligases are enzymes that catalyse the joining of two molecules by forming a new chemical bond (Wilkinson *et al.*, 2001). The ligation reaction requires energy in the form of adenosine triphosphate (ATP) or nicotinamide adenine dinucleotide (NAD) depending on the type of ligase (Doherty and Suh, 2000). In order to join two restriction fragments, two

phosphodiester bonds must be synthesised in two DNA strands (Struhl, 2009). Many different types of ligases are used in molecular biology, thus offering a large degree of diversity in the reaction conditions and catalytic specificities (Pergolizzi *et al.*, 2016). Amongst the ligases, T4 DNA ligases and thermostable DNA ligases are the largest product segment in the global molecular biology enzyme market (The global molecular biology enzymes and kits & reagents market, 2017- 2022). A detailed description of the global molecular biology enzyme market for ligases is shown in Table 1.3 (Doherty and Suh, 2000).

1.5.2.3 Nucleases

Nucleases are enzymes that hydrolyse phosphodiester bonds located in the backbone of nucleic acids (Struhl, 2009). The two main broad classes of nucleases are exonucleases and endonucleases (Rittié and Perbal, 2008). Exonucleases cleave nucleic acid molecules at the end, while and endonucleases cleave from within the nucleic acid (Littlechild, 2016). Exonuclease (e.g. Exonuclease III) cleaves and hydrolyses the 3' or 5' terminus phosphodiester bonds of polynucleotide molecules. The most studied endonucleases are the restriction endonucleases (Fox, 1988; Fukuyo *et al.*, 2015). Restriction endonucleases are divided into types I, II, III, IV, V and artificial restriction enzymes which differs in structure, cleavage site, specificity and cofactors (Loenen *et al.*, 2014). Restriction endonucleases cut DNA molecules at specific positions identifiable by their DNA sequence, which is often palindromic (Loenen *et al.*, 2014). They play a crucial role in gene cloning and recombinant DNA technology (Loenen *et al.*, 2014). Table 1.3 details the global molecular biology enzyme market size for restriction endonucleases. The most commonly used restriction enzymes in the molecular biology laboratory are *BamHI*, *EcoRI*, and *PstI* (Spot *et al.*, 2012). The application of these enzymes in molecular biology includes genotyping by RFLP and nuclease protection analysis, the digestion of genomic DNA prior to electrophoretic separation and screening using Southern blotting, and the generation of restriction fragments that can be sub-cloned into suitable vectors (Struhl, 2009).

A relatively recent endonuclease that has been successfully used for cloning is e *cas9* gene-editing enzyme. This RNA-guided DNA endonuclease enzyme cuts two DNA molecules at specific sites in the genome to enable the addition or removal of DNA, and offers and effective way for genetic manipulation (Sander and Joung, 2014).

1.5.2.4 Other molecular biology enzymes

Other important molecular biology enzymes include polynucleotide kinases, methyltransferases, glycosylases, and modifying enzymes, amongst others (Rittié and Perbal, 2008). Modifying enzymes include terminal deoxynucleotidyl transferase, alkaline phosphatase and polynucleotide kinase (Struhl, 2009). Terminal deoxynucleotidyl transferase is used in the homopolymer tailing process to add ribonucleotides to the 3' terminus of the DNA. This enzyme can therefore be used to ligate the vector DNA and the target gene (Roychoudhury *et al.*, 1976). Alkaline phosphatase removes the phosphate group at the 5' end of the DNA molecule (Brown, 2016). The application of this group of enzymes involves the prevention of 2 adaptors from being able to ligate. Unlike alkaline phosphatase which removes the phosphate group, polynucleotide kinase adds the phosphate group to the 5' end of the DNA molecule (Brown, 2016). Therefore, alkaline phosphatase and polynucleotide kinase are used for the manipulation of nucleic acid adaptors and primers (Brown, 2016). The global molecular biology enzyme market for all of the other enzymes is detailed in Table 1.3. They are primarily used for the modification of oligonucleotides in molecular biology (Struhl, 2009).

1.6 Conclusion

Stafford *et al*. (2005) hypothesised that the presence of the unique plant species in the Kogelberg Biosphere Reserve's *fynbos* soil has broad microbial diversity. We hypothesise that the uniqueness of the plant species and microbial diversity extends to unique viral diversity and therefore to novel viral enzymes. Assessing the microbial diversity in Kogelberg Biosphere Reserve's *fynbos* soil contributes to a more comprehensive picture of the microbial community and provides an understanding of the potential impact of a particular microbial community in the *fynbos* soil (Lako, 2005; Stafford *et al.*, 2005; Slabbert *et al.*, 2010; Ramond *et al.*, 2015; Miyambo *et al.*, 2016; Moroenyane *et al.*, 2016; Postma *et al.*, 2016).

Metaviromic techniques facilitate the discovery of novel viruses and new enzymes with improved catalytic properties. However, the metaviromes in the *fynbos* soil are still unexplored, since the viral communities associated with this soil type have never been studied previously. Therefore, metaviromic approaches offer an opportunity to advance the

ability to explore viral genomes for novel genes, including those related to nucleic acid manipulating enzymes.

High-throughput sequencing together with bioinformatics analyses are crucial for improving the effectiveness of bioprospecting of soil metaviromics for the discovery of novel enzymes. The consideration of sequence-based and function-based screening of metaviromes highlights an need to use both screening methods to mine relevant enzymes for use in the biotechnology or molecular biology industries, since both methods have biases and limitations (Schmitz *et al.*, 2010).

Recent advances in biotechnology and the ongoing market demand for nucleic acid manipulating enzymes requires innovation to constantly improve the tools and techniques available for competitive and cutting-edge research. Highly pure reagents which are rapid and reliable are needed for accurate results and facilitate rapid and efficient workflows.

1.7 Hypothesis

The Kogelberg Biosphere Reserve has been suggested to harbour both novel plant and bacterial species. This study hypothesises that the unique Kogelberg Biosphere Reserve biodiversity extends to viral diversity, providing an opportunity for novel enzyme gene discovery.

1.8 Aim

To use current cutting-edge metaviromics techniques to characterise the viral diversity of the Cape Floristic Region soil and to explore this environment for isolation of novel nucleic acid manipulating enzymes

1.9 Objectives

The objectives of this study are to:

- Extract high quality metagenomic and metaviromic DNA from soil collected from the Kogelberg Biosphere Reserve, a biodiversity hotspot.
- Determine both microbial and viral diversity from the selected site using NGS techniques.
- Screen and identify targeted nucleic acid modifying enzyme genes (such as polymerases, ligases, nucleases, phosphatases, methylases, and topoisomerases)

using both direct sequence-based and functional screening metagenomics approaches.

- Design and develop recombinant expression systems for heterologous production of targeted nucleic acid manipulating enzymes.
- Purify and functionally characterise the expressed enzyme/s.

CHAPTER 2: MATERIALS AND METHODS

2.1 Chemicals and reagents

Molecular biology and analytical grade chemicals and reagents were obtained from various suppliers specified below. Bacteriological agar, tryptone, sodium chloride and yeast extract were purchased from Merck (South Africa). The Zymo Research Soil Microbe MidiPrep™ kit was purchased from Zymo Research (USA) and the GeneJET™ gel extraction kit from Fermentas, Inqaba Biotech (South Africa). The GeneJET™ Plasmid Miniprep Kit, DNA size markers, T4 DNA ligase, Protein Page Ruler molecular weight markers and restriction enzymes were obtained from Thermo Scientific (South Africa). The Kapa Hifi PCR kit was obtained from Kapa Biosystems (South Africa). Oligonucleotide primers for polymerase chain reaction (PCR) were synthesised by Inqaba Biotec (South Africa). Protino® Ni-TED kit was purchased from Macherey Nagel (Germany). Antibiotics were purchased from Sigma-Aldrich (USA). Unless stated, all molecular biology experiments were performed using MilliQ water produced using a Merck Millipore Elix® system from Merck (South Africa).

2.2 Bacterial strains, growth conditions, primers and vectors

All growth media were autoclaved at 121ºC for 20min, at 0.1MPa pressure, cooled to 55ºC to enable, where necessary, the addition of appropriate antibiotics or inducers prior to use (Table 2.1 and 2.2). Buffer and media preparations, bacterial strains and vectors used in this study are shown in Table 2.2, 2.3 and 2.4.

Table 2.1: Antibiotics and inducers used in this study

Antibiotics / inducers stock solutions	Preparation
12.5mg/mL Chloramphenicol	12.5µg of Chloramphenicol water soluble (Sigma) dissolved in 1mL of 100 % ethanol
100mg/mL Ampicillin	100µg of Ampicillin sodium salt (Sigma) dissolved in 1mL of distilled water
50mg/mL Kanamycin	50µg of Kanamycin sulfate (Sigma) dissolved in 1mL of distilled water

Antibiotics / inducers stock solutions	Preparation
10% Arabinose	1g of L-(+)- arabinose (Sigma) dissolved in 9mL of distilled water
1M Isopropyl-β-D-thiogalactopyranoside (IPTG)	2.83g Isopropyl B-D-thiogalactoside (IPTG) (Invitrogen) dissolved in 10mL distilled water

Table 2.2: Buffers solutions and media used in this study

Buffer/Medium	Composition in 1L	pH
Buffers **10× Agarose gel running buffer**	108gTris base; 55g Boric acid; 7.45g EDTA	8.0
10× PBS buffer	81.8g NaCl, 20.1g KCl, 14.2g Na_2HPO_4, 2.45 gKH_2PO_4	8.0
Tris-HCl buffer	6.1g Tris base adjust pH using 32% HCl	7.0
10× SDS-PAGE running buffer solution	30.0g of Tris base, 144.0g of glycine, and 10.0g of SDS	8.3
Coomassie staining solution I	1.25g Coomassie R-250, 225mL isopropanol and 50mL glacial acetic acid	
Coomassie staining solution III	10% (v/v) acetic acid; 0.003% (w/v) coomassie brilliant blue G	
SDS de-staining solution	10% (v/v) acetic acid and 10% (v/v) methanol	
TE buffer	1.21g Tris Base, 0.37 g EDTA	8.0
LEW buffer	7.8g NaH_2PO_4 and 17.5g NaCl. Adjust pH using in NaOH	8.0
1× elution buffer	7.8g NaH_2PO_4, 17.5 g NaCl, 480.5g Urea, 17.0g imidazole. Adjust pH using in NaOH	8.0
1× SM buffer **Media**	5.8g NaCl, 1.2g $MgSO_4$, 50mL 1M Tris-HCl 10g Tryptone, 5g Yeast extract and 10g NaCl	7.5

Buffer/Medium	Composition in 1L	pH
Luria Burtani medium (LB medium)	10g Tryptone, 5g Yeast extract, 10g NaCl and 15g agar	7.0
Luria Burtani Agar	16g Tryptone, 10 g Yeast extract and 5g NaCl	7.0
2YT Medium		7.0

Table 2.3: Bacterial strains used in this study

Strains	Genotype/Features	Selective Marker	Source
***E. coli* EPI300-T1R**	*F- mcrA Δ(mrr-hsdRMS-mcrBC) Φ80dlacZ ΔM15 ΔlacX74 recA1 endA1 araD139 Δ(ara, leu)7697 galU galK λ- rpsL nupG trfA tonA dhfr.*	N/A	Epicentre
***E. coli* DH5α**	fhuA2Δ*(argF-lacZ)U169 phoA glnV44 Φ80 Δ(lacZ)M15 gyrA96 recA1 relA1 endA1 thi-1 hsdR17*	N/A	Lucigen
***E. coli* BL21-(DE3)**	F– *ompT hsdSB (rB- mB-) gal dcm lon* λ(DE3 [*lacI lac*UV5-T7 gene 1 *ind1 sam7 nin5*])	N/A	Lucigen
***E. coli* B121 (Ai)**	BF$^-$*ompT gal dcm lon hsdS$_B$*($r_B^-m_B^-$)[*malB*$^+$]$_{K\text{-}12}$(λ^S) *araB::T7RNAP-tetA*	N/A	Thermo Fisher
***E. coli* B121 (DE3) pLysS**	BF$^-$*ompT gal dcm lon hsdS$_B$*($r_B^-m_B^-$) λ(DE3 [*lacI lacUV5-T7p07 ind1 sam7 nin5*]) [*malB*$^+$]$_{K\text{-}12}$(λ^S) pLysS[*T7p20 ori*$_{p15A}$](Cm^R)	Chloramphenicol	Lucigen
***E. coli* mutant CSH26 fcsA29**	*[F_ ara (lac-pro) thi fcsA29 met::Tn5]*	N/A	Georg-August University

Table 2.4: Vectors used in this study

Strains/Vectors	Features	Selective Marker	Source
pCC2FOS	Copy controlled vector, linearized and dephosphorylated at *Eco*72I restriction site.	Chloramphenicol	Epicentre
pUC57	pUC57 used in this study is a GenScript standard vector for synthetic genes. The vector length is 2,710bp and is isolated from *E. coli* strain DH5α by standard procedures.	Ampicillin	Novagen
pUC57_RNALig2	pUC57 derived plasmid containing 237bp RNA ligase (*RNALig2*) gene isolated from a Kogelberg Biosphere Reserve soil metavirome.	Ampicillin	This study
pUC57_HNHc	pUC57 derived plasmid containing 174bp putative HNH endonuclease (*HNHc*) gene isolated from a Kogelberg Biosphere Reserve soil metavirome.	Ampicillin	This study
pUC57_PolB1	pUC57 derived plasmid containing 210bp DNA polymerase III subunit Beta (*PolB1*) gene isolated	Ampicillin	This study

Strains/Vectors	Features	Selective Marker	Source
	from a Kogelberg Biosphere Reserve soil metavirome.		
pUC57_RNALig1	pUC57 derived plasmid containing 391bp RNA ligase T4 (*RNALig1*) gene isolated from a Kogelberg Biosphere Reserve soil metavirome.	Ampicillin	This study
pUC57_PolA1	pUC57 derived plasmid containing 738bp DNA polymerase type A family (*PolA1*) gene isolated from a Kogelberg Biosphere Reserve soil metavirome.	Ampicillin	This study
pUC57_E7	pUC57 derived plasmid containing 179bp putative endonuclease VII (*E7*) gene isolated from a Kogelberg Biosphere Reserve soil metavirome.	Ampicillin	This study
pUC57_DNAlig	pUC57 derived plasmid containing 343bp DNA ligase AM (*DNALig*) gene isolated from a Kogelberg Biosphere Reserve soil metavirome.	Ampicillin	This study
pUC57_Pol A2	pUC57 derived plasmid containing 741bp DNA	Ampicillin	This study

Strains/Vectors	Features	Selective Marker	Source
	polymerase type A family *(Pol A2*) gene isolated from a Kogelberg Biosphere Reserve soil metavirome.		
pET20b(+)	Expression vector containing T7 *lac* promoter inducible with IPTG and C-terminal HIS Tag sequence.	Ampicillin	Novagen
pET20b(+)_RNALig1	pET20b(+) derived expression vector containing 237bp RNA ligase gene (*RNALig2*) gene in frame with theT7 *lac* promoter gene and C-terminal his tag sequence.	Ampicillin	This study
pET20b(+)_HNHc	pET20b(+) derived expression vector containing 174bp Putative HNH endonuclease (*HNHc*) gene in frame with the T7 *lac* promoter gene and C-terminal his tag sequence.	Ampicillin	This study
pET20b(+)_PolB1	pET20b (+) derived expression vector containing 210bp DNA	Ampicillin	This study

Strains/Vectors	Features	Selective Marker	Source
	polymerase III subunit Beta (*PolB1*) gene in frame with the T7 *lac* promoter gene and C-terminal his tag sequence.		
pET28a(+)	Bacterial expression vector with T7 *lac* promoter inducible with IPTG and C-terminal HIS Tag sequence.	Kanamycin	Novagen
pET28a(+)_RNALig2	pET28a (+) derived expression vector containing 391 bp RNA ligase T4 (*RNALig2*) gene in frame with the T7 *lac* promoter gene and C-terminal his tag sequence.	Kanamycin	This study
pET28a(+)_PolA1	pET28a(+) derived expression vector containing 738bp DNA polymerase type A family (*PolA1*) gene in frame with the T7 *lac* promoter gene and C-terminal his tag sequence.	Kanamycin	This study
pET28a(+)_RE	pET28a(+) derived expression vector containing 248bp putative superfamily II DNA/RNA	Kanamycin	This study

Strains/Vectors	Features	Selective Marker	Source
	helicase (*RE*) gene product in frame with theT7 *lac* promoter gene and C-terminal his tag sequence.		
pET28a(+)_E7	pET28a(+) derived expression vector containing 179bp putative endonuclease VII (*E7*) gene in frame with the T7 *lac* promoter gene and C-terminal his tag sequence.	Kanamycin	This study
pET30 b(+)	Bacterial expression vector with T7 *lac* promoter inducible with IPTG and C-terminal HIS Tag sequence.	Kanamycin	Novagen
pET30b(+)_PolA2	pET30b(+) derived expression vector containing 741 bp DNA polymerase type A family (*Pol A2*) gene in frame with the T7 *lac* promoter gene and C-terminal his tag sequence.	Kanamycin	This study
pET30b(+)_DNALig1	pET30b (+) derived expression vector containing 343bp DNA ligase AM (*DNAlig*) gene in frame with the T7 *lac*	Kanamycin	This study

Strains/Vectors	Features	Selective Marker	Source
	promoter gene and C-terminal his tag sequence.		

2.3 Sample site location

Fynbos soil samples were collected from the Kogelberg Biosphere Reserve, situated in the Boland Mountains to the east of Cape Town, South Africa (GPS coordinates: 34°19'48".0 S, 18°57'21.0" E). Soil samples were collected aseptically during the winter of 2014. Approximately 20kg of soil was collected at depth of 0 - 4 cm and were stored in sterile containers at -80ºC. These samples were used for both bacterial diversity analysis using the 16S rRNA marker gene and the metavirome analysis. A flow diagram demonstrating an overview of the experimental design of this study is shown in Figure A1, Appendices.

2.4 Bacterial diversity analysis using the 16S rRNA phylogenetic markers

2.4.1 Extraction of total DNA from *fynbos* soil

Total DNA extraction was performed using the ZR soil microbe DNA Midi Prep kit (Zymo Research). In triplicate, 0.25 g of soil was dissolved in 750µl lysis solution and the process was followed according the manufacturer's instructions. The size and the quantity of DNA extracted was determined using agarose gel electrophoresis and a NanoDrop 2000C spectrophotometer, respectively.

2.4.2 Polymerase chain reaction amplification of 16S rDNA fragments

Universal primer pair E9F (5′-GAGTTTGATCCTGGCTCAG-3′) (Farrelly *et al.*, 1995) and U1510R (5′-GGTTACCTTGTTACGACTT-3′) (Reysenbach and Pace, 1995) were used for 16S rDNA amplification. The PCR reaction (50Ll) contained: 1× PCR buffer, 1.5mM $MgCl_2$, 1U of *Taq* polymerase, 0.5mM (each) E9F and U1510R primers, 200µM of each deoxynucleoside triphosphate (dNTPs), and approximately 10–15 ng of isolated DNA from Kogelberg Biosphere Reserve as a template. The following PCR conditions were used for DNA amplification: 96°C for 2min (1 cycle); followed by 30 cycles of 96°C for 1 min, 50°C

for 1min, 72°C for 1min; and a final incubation at 72°C for 10min. The amplified PCR products were separated by electrophoresis on a 1% agarose gel and purified using ZR soil microbe gel extraction kit.

2.4.3 Next-generation sequencing and analysis of 16S rRNA amplicon

The recovered amplified PCR products were sequenced using 27F-16S (5′-AGAGTTTGATCMTGGCTCAG-3′) and 518R-16S (5′-ATTACCGCGGCTGCTGG-3′) (Lu *et al.*, 2007) primers through the High-Throughput Illumina MiSeq sequencing platform (service provided by Inqaba Biotech, Pretoria, South Africa). The quality of the raw read files was checked, filtered and trimmed, to remove low quality (sequence limit of 0.05) and ambiguous reads (maximum of 2 and minimum length of 15). Sequence reads of ~500bp in length with an average quality score of 25 or higher were analysed using the QIIME (Quantitative Insights Into Microbial Ecology) software, version 1.8.0 (Caporaso *et al.*, 2012). The Usearch tool in QIIME was used to identify chimeric sequences. The Uclust approach and the open-reference OTU picking strategy of the QIIME package was used for OTU clustering and representative sequence determination based on 97% sequence similarity. Taxonomic assignment was carried out by comparing with the 97-OTUs and the 97-OTU-taxonomy files of the Greengenes database version 13_08 (McDonald *et al.*, 2012). Values of the Chao1, Shannon, and Simpson (Caporaso *et al.*, 2010) diversity indices were determined with subsamples adjusted to contain the same number of sequences. A neighbour joining tree was constructed with the MEGA6 program (Tamura *et al.*, 2013) to show relationships between strains.

2.5 Metavirome nucleic acid extraction and analysis

2.5.1 Sample processing, nucleic acid extraction

Kogelberg Biosphere Reserve soil samples were processed as previously described (Casas and Rohwer, 2007) with some modifications. Eight kilograms (8kg) of soil was resuspended in 8L of sterile 1× SM buffer (Table 2). The suspension was mixed by vigorous shaking and left overnight at 4°C to settle. The upper layer was carefully decanted into new sterile centrifuge bottles and centrifuged (10000g, 15min) to pellet any remaining soil particles and other debris. The recovered supernatant was passed through a 0.22µm filter (Millipore, streicup 500mL) using a peristaltic pump. The filtrate was treated with approximately 10

units per millilitre of DNase (NEB) to remove unwanted DNA from bacterial cells. Viral particles were precipitated by overnight incubation in 10% (w/v) polyethylene glycol (PEG) 8000 at 4°C followed by centrifugation (15min, 11000 × g). The supernatant was discarded. The recovered viral pellet was resuspended in 10mL TE buffer, pH 7.6 (Casas and Rohwer, 2007).

The absence or otherwise of bacterial and eukaryotic DNA was confirmed by PCR using the following universal primer pairs: E9F (5′-GAGTTTGATCCTGGCTCAG-3′) and U1510R (5′-GGTTACCTTGTTACGACTT-3′) targeting 16S rDNA and ITS1 (5′-TCCGTAGGTGAACCTGCGG-3′) and ITS4 (5′-TCCTCCGCTTATTGATATGC-3′) targeting 18S r DNA (Merseguel *et al.*, 2015) using the cycling conditions described in section 2.4.2 above.

2.5.2 Transmission electron microscopy

Aliquots of viral suspension isolated from soil were fixed with 2% glutaraldehyde for 3 hours at 4°C. An amount of 10µL of the phage suspension was then overlaid on a carbon coated grid of 200 Mesh (Ackermann, 2009). The suspension was allowed to dry on the grid, which was then negatively stained with 2% uranyl acetate. Excess stain was removed using filter paper and allowed to air-dry prior to examination using a Philips (FEI) CM100 TEM. TEM services were provided by CSIR (Pretoria, South Africa).

2.5.3 Nucleic acid sequencing and quantification

Extracted metaviromic DNA (section 2.5.1) was sequenced using the Illumina MiSeq platform (service provided by Inqaba Biotech) (Pretoria, South Africa). Following DNA quantification using the NanoDrop 2000C spectrophotometer (Thermo Fisher scientific), 1 ng of isolated metavirome DNA was used to prepare 4 individually indexed NexteraXT libraries. They were then sequenced using the MiSeq v3 (600 cycles) sequencing kit, generating 2 × 300bp reads. The raw reads were trimmed and demultiplexed, resulting in 4×2 fastq files.

2.5.4 Metavirome sequence assembly, analysis and screening

2.5.4.1 Sequence assembly, and analysis

The quality of the raw read files was checked with CLC genomics workbench version 6.0.1 (CLC, Denmark). The reads were then filtered and trimmed, with the removal of low quality (sequence limit of 0.05), ambiguous reads (maximum of 2 and minimum length of 15). This yielded 1 488 462 918 reads with an average length of 212bp. Quality control (QC) was done by manually investigating the reads after sequencing by CLC genomics workbench. The post-QC reads were assembled using CLC genomics workbench as paired files (4 × 1 read files). The assembly resulted in 28 511 204 contigs with a minimum length of 1 002 bases at an N50 (minimum contig length needed to cover 50% of the genome) of 2 047 and a maximum length of 47 854 bases.

DNA sequence reads and contigs were uploaded to the MetaVir (Roux *et al.*, 2014) (http://metavir-meb.univ-bpclermont.fr), VIROME (http://virome.dbi.udel.edu/) (Wommack *et al.*, 2012) and MG-RAST (http://metagenomics.anl.gov/) (Keegan *et al.*, 2016) servers for identifying sequence similarities to known genotypes and to observe the diversity of viral types. The taxonomic composition was computed from a BLAST comparison with the RefSeq complete viral genomes protein sequence database from NCBI (release of 2016-01) using BLASTp with a threshold of 50 on the BLAST bitscore. The assembled sequences were searched for ORFs and compared to the RefSeq complete protein database using MetaVir. Functional and phylogenetic assignments were based on annotations and other information obtained from the following databases: GenBank, Integrated Microbial Genomes (IMG) (http://img.jgi.doe.gov), Kyoto Encyclopaedia of Genes and Genomes (KEGG) (Kanehisa *et al.*, 2012), Pathosystems Resource Integration Center (PATRIC), RefSeq (Pruitt *et al.*, 2005), SEED (Overbeek *et al.*, 2005), Swiss-Prot (Boutet *et al.*, 2007), TrEMBLE (Bairoch and Apweiler, 2000), and eggnog (Muller *et al.*, 2010); and for the assignment of functional hierarchy, COG (clusters of orthologous groups) (Tatusov *et al.*, 2003), KEGG Orthology (KO), and NOG databases were used. The Genome relative Abundance and Average Size (GAAS) (Angly *et al.*, 2009) tools were used for normalisation of the total composition, estimation of the mean genome length and for the estimation of relative abundance and size for each taxon. The phylogenetic tree was generated by an open-source JavaScript library called jsPhyloSVG (Smits and Ouverney,

2010). The phylogenetic trees were based on the most similar viral sequences obtained from the databases similarity searchers together with the Kogelberg Biosphere Reserve virome sequences, and computed with 100 bootstraps. Further comparison analysis of the sequences was performed using METAGENassist (a web server that provides a broad range of statistical tools for comparative metagenomics) (Arndt *et al.*, 2012). Functional assignments produced by VIROME using 120 identified functional subsystems were used for the statistical analysis with METAGENassist.

Clustering analysis comparison was plotted as a clustering tree and computed with pvclust computed by MetaVir (an R package for assessing the uncertainty in hierarchical clustering) (Suzuki and Shimodaira, 2006). Hierarchical clustering using dinucleotide comparisons was used to quantify the grouping behaviour of 9 published metaviromes and the comparisons were plotted and demonstrated as clustering dendrograms. Only metaviromes containing more than 50,000 sequences and with an average sequence length of over 100bp were used to avoid sequence error, as this comparison is based on a normalised virome sub-sample. Metaviromes that did not match these criteria were not selected for nucleotide composition comparison. Hence, only 9 metaviromes were suitable for comparison using dinucleotide frequencies in the MetaVir sever. The largest contigs were analysed by MetaVir. The SEED classification clustering of the 12 metaviromes was assessed using BLASTp against the nr database of NCBI (release 2017-05) (Aziz *et al.*, 2012). Differences between the virome SEED functional components were transformed into a Bray-Curtis dissimilarity matrix using the vegan package in RStudio, clustered using the hclust algorithm (method = average), and represented as a dendrogram (Racine, 2012; Oksanen*et et al*., 2016).

2.5.4.2 Sequence-based screening

Sequence-base screening involved a combination of 3 metavirome sequencing processing platforms such as MetaVir (http://metavir-meb.univ-bpclermont.fr), VIROME (http://virome.dbi.udel.edu/) and MG-RAST (http://metagenomics.anl.gov/) to estimate viral diversity and taxonomical classification of the constructed metavirome library (Roux *et al.*, 2014) (Wommack *et al.*, 2012) (Keegan *et al.*, 2016). The assembled *de novo* CLC sequence data was used as input data for the 3 pipeline platforms coupled with the BLASTP server to screen for genes encoding novel nucleic acid manipulating enzymes. MetaVir server displays, for each contig, a map of predicted ORFs with its associated annotations and

sequences, as well as a list of related contigs with a significant BLASTp similarity. In total, 9 putative DNA modifying genes were identified from MetaVir pipeline, such as ligases (*RNALig2*, *RNALig3* and *DNAlig*), DNA polymerases (*PolB*, *PolA1* and *PolA2*), Nucleases (Restriction endonuclease (*RE*), Homing endonuclease (*HNHc*) and Endonuclease 7 (*E7*)) (it will be discussed further in details in Chapter 5).

2.5.5 Metavirome fosmid library construction

2.5.5.1 Multiple displacement amplification

Owing to low concentration, the Metavirome DNA isolated under section 2.5.1 was subjected to a MDA procedure using the REPLI-g kit (Qiagen) in order to improve the DNA concentration. In brief, 6.5ng of DNA was added to 2.5µL of buffer D1 (containing KOH & EDTA) to denature the DNA. The sample was incubated for 3 minutes at room temperature, before adding 5µL of buffer N1 (neutralising buffer). After neutralising, 40µL of Master Mix containing the REPLI-g DNA polymerase was added to 10µL of the denatured DNA solution. The mixture was incubated overnight at 30°C for amplification. After amplification, the REPLI-g DNA polymerase was heat-inactivated for 10 minutes at 65°C.

2.5.5.2 Construction of metavirome DNA fosmid library

The amplified DNA from section 2.5.5.1 was prepared for library construction using the Epicentre® fosmid library construction kit (Epicentre, Madison, WI) according to the manufacturer's protocol, with some minor modifications. The modifications included the omission of the blunt-end step since the product of the REPLI-g procedure leaves blunt-end as a result of the proofing reading activity of the Phi 29 polymerase used. Briefly, approximately 250ng/mL DNA of the amplified DNA (representing an equivalent of 561.75pmol/L of DNA ends) was purified and directly ligated into the CopyControl pCC1FOS fosmid vector (Epicentre). Purified ligation reaction was *in vitro* packaged into lambda phages and then used to infect phage resistant *Escherichia coli* EPI300-T1R strain. The packaged reaction (200µL) was then plated on LB agar supplemented with 12.5µg/mL chloramphenicol plates and incubated at 37°C overnight. The colony growth was used to determine library size and to verify insert size following growth cultivation under the following conditions: 5mL LB medium, containing 12.5µg/mL of chloramphenicol and 10µL induction solutions at 1×final concentration (EPICENTRE®). The fosmid library was

then used for functional screening of the polymerase 1 genes using complementation screening of cold-sensitive *E. coli* mutant explained in the subsequent section 2.5.5.3.

2.5.5.3 Functional screening of the library for DNA polymerase 1 using complementary assay

Function-based detection of fosmids harbouring polymerase-encoding genes was based on complementation screening of cold-sensitive *E. coli* mutant CSH26 fcsA29 [F_ ara (lac-pro) thi fcsA29 met::Tn5] developed by Nagano *et al.*, (1999). For the direct screening of DNA polymerase 1, LB agar plates containing 12.5µg/mL chloramphenicol were prepared in the 96 well plates. The screening was initiated by transformation of the recombinant fosmids library into *E. coli* mutant CSH26 fcsA29. Subsequently, the resulting transformed *E. coli* CSH26 fcsA29 clones were plated onto LB agar containing 12.5µg/mL of chloramphenicol and incubated at 18°C. Positive clones with a colony diameter of approximately 3mm visible after 48 to 72hr of incubation were selected for secondary screening. The colonies were sub-cultured onto LB agar containing 12.5µg/mL chloramphenicol and incubated at 18°C. The negative control was the *E. coli* CSH26 fcsA29 harbouring the cloning vector without an insert. The positive clones were pooled together and DNA sequencing carried out using the Illumina MiSeq platform (Inqaba Biotechnical Industries). The sequences were analysed using MG-RAST (http://metagenomics.anl.gov/) (Keegan *et al.*, 2016) and BLAST (NCBI).

2.5.6 General analysis procedures

2.5.6.1 Fluorimetry (QubitTM)

Plasmid DNA concentrations were measured using the Quant-iT^{TM} dsDNA BR Assay Kit (Invitrogen) according to the manufacturer's instructions. All reagents for DNA assays were used at room temperature. Readings were taken using a Qubit^{TM} fluorimeter.

2.5.6.2 NanoDrop analysis

Quantity and quality of the DNA extracted was determined using the NanoDrop 2000C spectrophotometer (Thermo Fisher scientific, SA). DNA concentration was measured using modified Beer-Lamberts Law (NanoDrop 2000C Spectrophotometer User Manual V), as follows: $A=\varepsilon cb$, where A= the absorbance of the sample, ε = the molar absorptivity with units L mol^{-1} cm^{-1}, b = the path length of the sample - that is, the path length of the cuvette in which the sample is contained (cm) and c = the concentration of the compound in solution,

expressed in mol L^{-1}. DNA quality was determined by the ratio of absorbance at wavelengths of 260nm over 280nm (A_{260} /A_{280}).

2.5.6.3 Agarose gel electrophoresis

Agarose gel electrophoresis was used to separate nucleic acid fragments. Genomic and plasmid DNA and PCR amplicons were visualised by the addition of 6× loading buffer (30% v/v glycerol, 0.25% w/v bromophenol blue) and subsequent electrophoresis in 1% or 0.7% (w/v) agarose gels prepared in 1×TAE buffer containing 0.5 µg/mL ethidium bromide (Sambrook and Russell, 2001). DNA molecular markers of an appropriate size distribution were used for molecular weight comparisons. Gel images were visualised and photographed using a digital imaging system (AlphaImager 2000, Alpha Innotech, and San Leadro, USA).

2.5.6.4 Restriction enzyme digestions

Restriction enzyme digestions were prepared in sterile 1.5mL microcentrifuge tubes in 50µL reaction volumes and were incubated at 37°C overnight. Approximately 1U of enzyme was used per microgram of plasmid or genomic DNA in the presence of the appropriate buffer as supplied by the manufacturer. Restriction enzymes were inactivated at 80ºC for 20mins.

2.5.6.5 DNA ligations

Ligations were carried out in 20µL volumes. In each microcentrifuge tube insert DNA and an appropriate cloning vector in a 2:1 or 3:1 ratio insert:vector molar concentrations were combined with 1U of T4 DNA ligase and 1× ligation buffer (Sambrook & Russell, 2001). Reactions were incubated at room temperature overnight. Ligation reactions were transformed directly into host cells.

2.5.6.6 Preparation of competent *E. coli* cells by $CaCl_2$ treatment

E. coli BL21 cultures from the glycerol stocks were streaked onto the surface of an LB agar plate and incubated for 24hrs at 37ºC. A single colony was then transferred into 5mL LB medium and incubated overnight at 37ºC in a shaking incubator. An amount of 500µL of the overnight culture was inoculated into 100mL LB medium in a 1L flask. The culture was incubated at 37ºC until an optical density (OD at 600nm) of 0.4-0.6 was reached. Cells were kept on ice in all subsequent steps. The cultures were centrifuged at 4ºC for 5mins at 5000rpm. The pellet was resuspended in 100mL ice cold 0.1M $CaCl_2$ and left on ice for 1min

and the supernatant was discarded. Cells were collected again and resuspended in 50mL of ice cold 0.1M $CaCl_2$ and held on ice for 9mins. The cultures were centrifuged at 4ºC at 5000rpm for 5mins and placed on ice. The pellet was again resuspended in 10mL ice cold 0.1M $CaCl_2$ and the supernatant was discarded. A volume of 10mL of ice cold sterile glycerol was added, the cells were resuspended, and aliquots were stored at -80ºC (Sambrook & Russell, 2001).

2.5.6.7 Transformation by heat shock

Approximately 10ng of purified DNA was added to 100µL of chilled chemically competent *E. coli* cells, left on ice for 10mins and heat shocked at 37°C for 5mins. The mixture was incubated on ice for 1min and then incubated in 1mL LB for 1hr at 37°C (150rpm). The cells were plated onto LB agar plates and incubated at 37°C overnight.

2.5.7 Recombinant protein production

2.5.7.1 Recombinant expression strategy

Nine genes, namely ligases (*RNALig2*, *RNALig3* and *DNAlig*), DNA polymerases (*PolB*, *PolA1* and *PolA2*), Nucleases (Restriction endonuclease (*RE*), Homing endonuclease (*HNHc*) and Endonuclease 7 (*E7*)) (discussed in details in Chapter 5), were selected from the metavirome for recombinant expression studies. The synthetic genes were synthesised through services offered by GenScript (USA Inc.). Firstly, the *de novo* synthesis of ORFs of all 9 identified genes was conducted with codon optimisation for *E. coli* codon usage. The 9 selected genes were synthesised with *Nde*I and *Xho*I restriction sites at the 5'and 3' prime of the gene sequences, respectively. All of the gene sequences were synthesised with 6× His sequence to facilitate downstream purification of the corresponding gene products. Eight of the synthetic gene constructs were provided in pUC57 cloning vector. However one gene (RE) was unstable in pUC57 and cloned directly into pET expression vector.

The gene fragments encoding the nucleic acid manipulating enzymes were excised from the pUC57 parental vector using *Nde*I and *Xho*I restriction enzymes and were ligated into the pET20b(+), pET28a(+) and pET30b(+) expression vectors pre-digested with *Nde*I/*Xho*I enzymes.

RNAlig1, *HNHc* and *Pol B* genes were cloned in to pET20b(+) expression vector, whereas *RNAlig2*, *RE*, *Pol A1* and *E7* were cloned in to pET28a(+) expression vector. Furthermore,

DNAlig and *Pol A2* were cloned in to pET30b(+) expression vector. Sequencing and restriction digestion using *Xba*I and *Xho*I for pET20b(+) constructs, *Mlu*I and *Xho*I for pET28a(+) constructs and *Nde*I and *Xho*I for pET30b(+) constructs, were used to confirm the presence of the correct inserts. The digested products were analysed using 1% agarose gel electrophoresis.

2.5.7.2 Standard protein expression conditions

Bacterial cultures were grown in LB medium, 2× YT medium and/or EnPresso® (BioSilta, Cambridgeshire, UK, and Boca Raton, FL, USA) (Krause *et al.*, 2016) *E. coli* high yield growth media (section 2.2 and Table 2). Briefly, one colony was inoculated in to 5mL LB media (containing 50µg/mL ampicillin or 50µg/mL kanamycin), followed by overnight incubation at 37ºC. The pre-inoculum (1mL of the overnight culture) was used to inoculate 50mL medium. The culture was incubated at 37 ºC until the OD_{600nm} reached 0.4 to 0.6.

Expression was induced by addition of 1 mM IPTG followed by incubation at a range of different temperatures (16 – 37ºC). After selecting the optimum temperature, IPTG concentrations between 100µM and 1mM were tested at this selected optimum temperature.

After induction of protein expression overnight, cultures were recovered by centrifugation at 13 000g for 10min. The pellets were resuspended in lysis buffer (100µL of B-PER) (Bacterial Protein Extraction Reagent, Thermo Pierce; in phosphate buffer, 50mM (pH 7.5)) following the manufacturer's instructions and sonicated using Bandelin electronic Heinrichstrasse 3-4 D-12207 (Berlin, Germany) set at: 10 × 15 s burst, with 15s cooling period between busts, to release intracellular proteins. Supernatants were centrifuged at 13 000g for 10min using a JA14 rotor (Beckman, Optima™ L-100XP) to pellet membrane proteins and the supernatant recovered. The pellet was resuspended in 8M Urea for 30min. Both the soluble and insoluble fractions were analysed by Sodium dodecyl sulphate page (SDS-PAGE) and stained with Coomassie blue to visualise proteins, according to standard procedures (Bio-Rad, SA).

2.5.7.3 Protein purification

HIS-tagged proteins were purified by immobilized metal ion affinity chromatography (IMAC) using the Protino® Ni-TED packed 2000 column (Machenery-Nagel, Germany). For larger culture volumes, the purification chromatography was monitored automatically

using an Akta Avant 150 FPLC. An XK16/20 column was packed with 25 mL Protino® Ni-TED resin. The column was equilibrated with 3 column volumes (CV) LEW buffer and 15mL of the soluble crude fraction containing a targeted HIS-tagged protein was loaded. Unbound proteins were washed with 3 CV of LEW buffer followed by elution of the targeted protein with elution buffer. Eluted targeted proteins were concentrated using a VIVASpin 10kDa cut-off membrane spin column (Sartorius Stedim, France) and the recovered proteins were resuspended in 50mM phosphate buffer (pH 8.0). Fractions containing purified protein were confirmed with 12% SDS-PAGE.

2.5.7.4 Western blotting

The proteins were first electrophoretically separated on 12% SDS-PAGE gel. After electrophoresis, proteins were transferred into 0.2µm-pore polyvinylidene difluoride (PVDF) membrane. The non-specific protein binding was blocked by incubating the membrane with 5% skim milk in Tris buffered saline with tween 20 (TBST)* for 1 hour. After washing 3 times, samples were reacted with anti-HIS-tag IgG conjugate with horse radish peroxidase (HRP from Sigma Germany; diluted 1:1000 in TBST,) for 1hr at room temperature. Protein bands were developed and detected using an enhanced chemiluminescence (ECL) kit method (Thermo Fisher scientific, SA).

2.5.8 General protein analytical procedures

2.5.8.1 Quick Start™ Bradford assay

Protein concentration was estimated using the Quick Start™ Bradford dye reagent (Bio-Rad, SA) using a bovine serum albumin (BSA) standard at concentrations of 4-20 µg/mL. An appropriately diluted volume of the protein sample (10µL) was mixed with 240µL Bradford dye reagent followed by incubation at room temperature for 5min, after which the absorbance was measured at 595 nm using a Bio-Tek® spectrophotometer operated via KC4 software.

2.5.8.2 SDS-PAGE analysis

Sodium dodecyl sulphate page gel was used to analyse protein samples at different stages of processing (Laemmli, 1970). Proteins were mixed with loading dye (2.5mL 1 M Tris-HCl pH 6.8, 0.5mL of ddH20,1.0g SDS, 0.8mL 0.1% Bromophenol Blue, 4mL 100% glycerol, 2mL 14.3M β-mercaptoethanol) and boiled at 95°C for 5 minutes before loading onto the

gel. Electrophoresis was performed in 1× SDS-PAGE running buffer at 180V for 40 minutes. Gels were stained with Coomassie brilliant blue staining solution and incubated for 30min or longer shaking (Table 2.2) and de-stained with destaining solution and incubated for an hour (Table 2.2).

2.5.8.3 Bioinformatics alignments of DNA ligase

The multiple sequence alignments combined with motif and domain search results to determine sequence features of DNAlig protein was done using CLC genomics workbench version 6.0.1 (CLC, Denmark). The deduced amino acid sequences aligned with DNAlig sequence were identified from BLAST searches (NCBI).

2.5.8.4 Activity assay of recombinant fusion protein

The activity of soluble recombinant DNA ligase was tested using ligation reactions essentially as described by Sambrook and Russell (2001). DNA ligase activity was compared with commercial preparations of T4 DNA ligase (Thermo Scientific, Invitrogen and Roche). The ability of the ligase isolated from Kogelberg Biosphere Reserve soil metavirome to ligate sticky and blunt ends was tested in ligation reaction consisting of: the test recombinant DNA ligase (1.5µg) (no DNA for the negative control) together with lambda DNA completely digested with sticky end restriction enzymes (*Hind*III, *BamH*I and *EcoR*I) or blunt-end restriction enzymes (*EcoR*V, *Sma*I, and *Pvu*III), with an appropriate buffer (250mM Tris-HCl (pH 7.6), 50mM $MgCl_2$, 5mM ATP, 5 mM DTT, 25% (w/v) polyethylene glycol-8000) and nuclease free water in a 20µl ligation reaction (Thermo Scientific) at 37°C (pre-setting) for one hour. The thermal stability of the enzyme was inferred from its ability to remain active in the ligation reaction after deactivation by heating at 60°C and 80°C for 15min. In order to demonstrate that the test recombinant DNA ligase can ligate vector to an insert, ligations were carried out by performing ligation of pUC57 with an insert digested with *Nde*I/*Xho*I. Molar ratios of 1:3 vector:insert were ligated and the products were used to transform *E. coli* DH5α and ligation products were also visualised with 1% agarose gel electrophoresis.

CHAPTER 3: 16S rRNA GENE DIVERSITY ANALYSIS

3.1 Introduction

The Kogelberg Biosphere Reserve in the Cape Floristic Region of South Africa is a centre of *fynbos* endemism with remarkably high plant diversity (Cowling, 1992). Furthermore, Cowling (1992) reported that 6252 of the world's identified plant species are exclusively found in the Kogelberg Biosphere Reserve. The soils in the Kogelberg Biosphere Reserve are characterised by an acidic pH and by a shortage of vital plant nutrients (i.e. phosphorus, potassium and nitrogen) (Richards *et al.*, 1997). The primary inhabitants of this area are slow growing *fynbos* plants, with lifespans of 12 - 20 years (Cowling, 1992; Richards *et al.*, 1997). As a result of a relatively infertile soil habitat, *fynbos* plant species have developed specialised nutrient-uptake mechanisms, which include symbiotic or endophytic relationships with bacteria, archaea and fungi (McDonald *et al.*, 1995).

For many years, knowledge of the microbiological diversity in this *fynbos* soil habitat was limited due to the difficulty and inconsistency of success of the microbial culture methods used to isolate and characterise several microbial species. Since the early 1990s, the development of the 16S ribosomal RNA (rRNA) gene as a prokaryote biomarker has provided insights into microbial community structure. Conventional techniques targeting the rRNA genes; such as Fluorescent In Situ Hybridisation (FISH) (Harmsen *et al.*, 2002), DGGE (Harmsen *et al.*, 2002), Terminal restriction Fragment Length Polymorphism (T-RFLP) of PCR amplified rRNA genes (Cancilla *et al.*, 1992; Case *et al.*, 2007) and the construction and sequencing of rRNA gene libraries (Bej *et al.*, 1990) have been used to estimate microbial diversity of various habitats, including *fynbos* soil (Easton, 2009). However, these methods have resulted in some fundamental challenges such as having inherent methodological bias, being labour intensive and time consuming (Ju and Zhang, 2015). There are several biases inherent in these methods which may select for certain genotypes or genospecies in soil. These methods have various inherent bias that come from the preferential DNA extraction, unequal amplification of target genes and the instability of certain genes upon cloning (Ju and Zhang, 2015).

Recent advances in the exploration of microbial diversity involve, amongst others, the application of high-throughput 16S rRNA gene sequencing using NGS platforms. These

approaches can result in less inherent bias since the direct sequencing of metagenome samples as well as PCR amplification, cloning and sequencing are possible. There are various advanced NGS platforms being used for detailed investigations of microbial diversity in various ecological niches (Vergin *et al.*, 2013). These include the 454 GS 20 sequencer (Margulies *et al.*, 2005), Illumina, Ion Torrent, and SOLiD (Rothberg *et al.*, 2011) sequencing platforms.

In addition to high-throughput NGS platforms, the availability of bioinformatics tools and statistical analysis has greatly improved the effectiveness of characterising various microbial communities. These bioinformatics analysis tools include web-based and locally installed platforms, reference databases and ecological matrices (Ju and Zhang, 2015). The local or web-based software packages such as QIIME (Caporaso *et al.*, 2010), Mothur (Schloss *et al.*, 2009), RDP (Cole *et al.*, 2009), VAMPS (Huse *et al.*, 2014) (http://vamps.mbl.edu/) and MEGAN (Huson *et al.*, 2007) are commonly used to filter, analyse and visualise large amplicon sequence datasets from NGS (Ju and Zhang, 2015). Open-source or freely available software packages such as R (e.g., vegan, ade4), PAST (http://folk.uio.no/ohammer/ past/), and STAMP (http://kiwi.cs.dal.ca/Software/STAMP), and commercial graphical software or programs such as CANOCO, PRIMER-E, SPSS and Microsoft EXCEL, are commonly applied for statistics and visualisation of data (Ju and Zhang, 2015; Vierheilig *et al.*, 2015).

Stafford *et al.* (2005) analysed Kogelberg Biosphere Reserve soil samples using DGGE and revealed previously unexplored bacterial diversity, with a relative abundances of phyla such as *Proteobacteria, Firmicutes, Actinobacteria* and *Acidobacteria*. In addition, Slabbert *et al.* (2010) studied microbial diversity in *fynbos* soils. However, most of the studies on *fynbos* soils to date has focused on the relationship between the vegetation types (e.g. rhizosphere, mycorrhizal) with the microbial diversity (Allsopp and Stock, 1995; Caravaca *et al.*, 2002; Spriggs *et al.*, 2003; Lako, 2005; Stafford *et al.*, 2005; Slabbert *et al.*, 2010; Ramond *et al.*, 2015; Miyambo *et al.*, 2016; Moroenyane *et al.*, 2016; Postma *et al.*, 2016). Hence, the main objective of this study was to obtain a comprehensive view of the *fynbos* soil (free soil not associated with plant roots) bacterial communities by using the 16S rRNA amplicon NGS techniques with the direct sequencing of PCR amplicons. The characterisation of the microbial composition found in the Kogelberg Biosphere Reserve *fynbos* soil will contribute

to an in-depth knowledge of the microbial community in this area of study, and further lead to the understanding of the possible influence of certain bacterial community members on *fynbos* soil function and dynamics.

3.2 Results and discussion

The area selected for sampling was the Kogelberg Biosphere Reserve (coordinates: 34° 19'48" .0S; 18° 57'21.0" E). The bacterial community structures of the *fynbos* soil were evaluated using extraction of metagenomics DNA from soil and 16S rRNA analysis. Three samples were pooled from three sites within the Reserve and analysed in triplicate to minimise experimental bias or random errors.

3.2.1 Chemical properties of the soil samples

To evaluate the general factors that influence bacterial communities in the *fynbos* soil, the chemical properties such as nutrient characteristics and pH were measured. The average concentration of minerals measured from the soil sample were as follows (Table B1, Appendices): potassium (K) ≤20; sodium (Na) ≤10; calcium (Ca) ≤5; magnesium (Mg) ≤1; sulphates (SO4) ≤5; chloride (Cl) ≤10; aluminium (Al) ~0.4; iron (Fe) ≤0.1; and manganese (Mn) ≤0.1 (units: ppm). The average soil moisture content measured was 11.3%, and the pH was acidic with an average of 5.3 (Table B1, Appendices). Low moisture content in various natural environments has been reported by Haynes and Swift (1989) to have an influence on general microbial growth and activity in soils (Hastings *et al.*, 2000; Dzikiti *et al.*, 2014). The sampling for this study was conducted as a once-off event, during a specific time of the day. It is acknowledged that the moisture content may vary during the day and during the year, and that this could specifically change dramatically during rainy seasons. Soil pH influence bacterial diversity as it affects nutrient availability and microbial activity (Rousk *et al.*, 2010; Griffiths *et al.*, 2011; Centeno *et al.*, 2012; Wang, 2016). However, the soil analysis in this study revealed that the *fynbos* soil was acidic and was characterised by a low level of inorganic nutrients. This observation is consistent with results obtained in previous studies conducted by Richards *et al.*, (1997) and McDonald *et al.*, (1995) on the *fynbos* soil.

3.2.2 16S rRNA gene analysis and amplicon sequence analysis and species richness estimation

3.2.2.1 Sequence assembly and analysis

To assess the prokaryotic diversity in the Kogelberg Biosphere Reserve *fynbos* soil, the universal primer pair E9F (Farrelly *et al.*, 1995) and U1510R (Reysenbach *et al.*, 1995) were used to target the 5′-GAGTTTGATCCTGGCTCAG-3′ and 5′-GGTTACC TTGTTACGACTT-3′ conserved region of the 16S rRNA gene. The expected 1.5kb PCR product was amplified using community DNA isolated from the sample sites as a template. The isolated PCR product was sequenced using Illumina MiSeq sequencing, using 27F-16S (5′-AGAGTTTGATCMTGGCTCAG-3′) and 518R-16S (5′-ATTACCGCGGCTGCTGG-3′) primers which target the V1‑V3 regions of the 16S rRNA gene (Lu *et al.*, 2007). The output was 2 × 500GB of data. Illumina MiSeq sequencing yielded an average of 16743 raw sequence reads over the 3 replicate samples. The quality of the raw read files was checked, filtered and trimmed to remove low quality (sequence limit of 0.05 to remove low quality sequence) and ambiguous reads (maximum of 2 to remove ambiguous nucleotides and minimum length of 15 to remove sequence based on length). This resulted in an average of 42091 high quality sequences. The processed reads were *de novo* assembled and the reads with an average quality score of 25 or higher, were analysed using the QIIME software, version 1.8.0 (Caporaso *et al.*, 2012). The Usearch, Uclust and open-reference OTU picking strategy features of QIIME were utilized to identify chimeric sequences, as well as for OTU clustering and representative sequence analysis based on 97% sequence similarity. Sequence comparison with the 97-OTU-taxonomy files of the Greengenes database version 13_08 were used to carry out taxonomic assignment (McDonald *et al.*, 2012). The values of the Chao1, Shannon, and Simpson diversity indices were then determined.

3.2.2.2 Diversity analyses and OTU clustering of sequences

Microbial diversity of the 16S rRNA gene from the Kogelberg Biosphere Reserve *fynbos* soil samples was analysed using rarefaction. In total, approximately 10,000 sequences were analysed without reaching a rarefaction plateau. In terms of the principles of the method, a rarefaction plateau represents the entire bacterial diversity (Fierer *et al.*, 2007) (Figure 3.1). However, estimation of species richness showed that only a portion of the richness in the bacterial communities (at the ≥97% sequence similarity level) was surveyed. Extrapolation

suggested that approximately 78.2% of the bacterial diversity was covered by the sequence dataset. Therefore, a fraction of the species diversity in this sample remains to be identified.

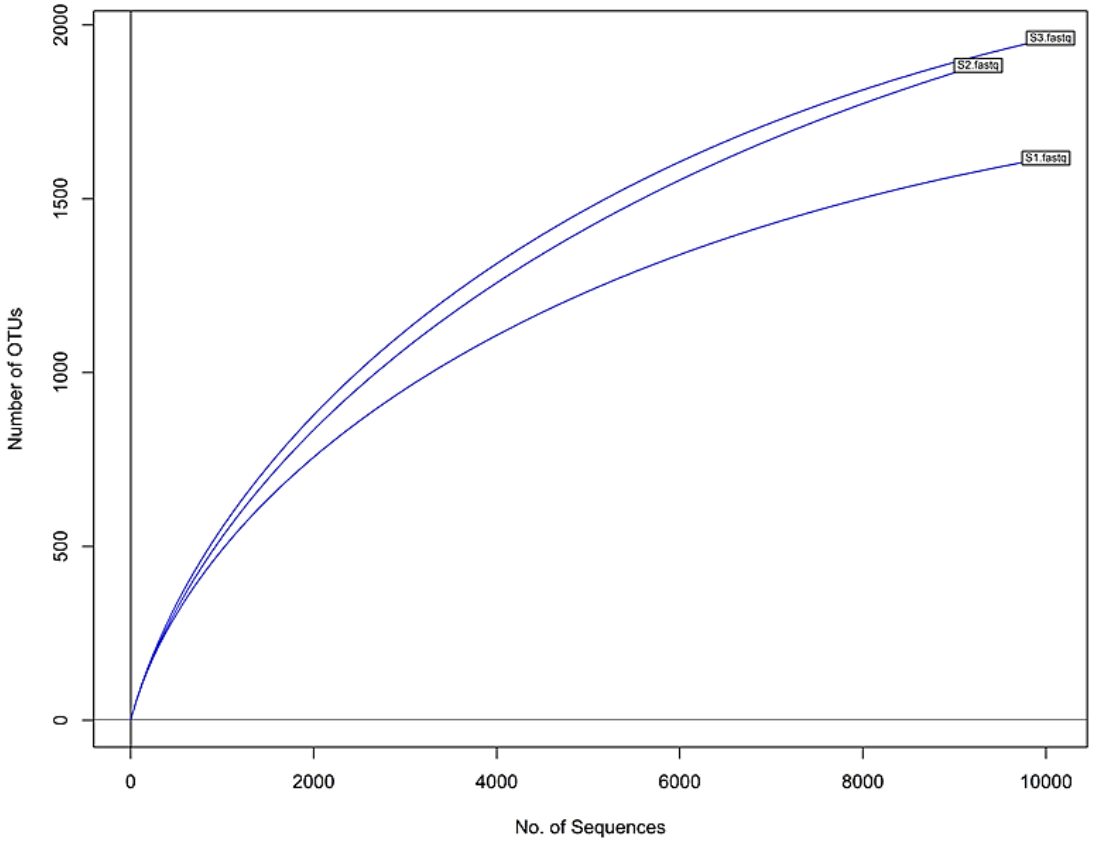


Figure 3.1: **Rarefaction curves for the bacterial libraries from the 3 Kogelberg Biosphere Reserve soil samples. Rarefaction curves were generated using EstimateS (version 7.5; R. K. Colwell, http://purl.oclc.org/estimates). OTU was defined at the ≥97% sequence similarity level.**

The average length of the assembled sequences filtered for quality and examined for chimeric sequences for the Kogelberg Biosphere Reserve *fynbos* soil samples was 251bp, while the average number of reads in the three samples was 41979. The sequences were clustered into OTUs at 97% similarity levels, and sequences with less than this level of similarity to known sequences were termed unclassified. In total, 1821 genospecies or Operational Taxonomic Units (OTUs) were observed that had sequence similarity to existing database entries, and overall 3% (of the total bacteria) sequences were unclassified at 97% sequence similarity with rRNA genes in public databases and therefore may be considered

novel. The rarefaction curves suggest that only a portion of the total diversity has been analysed. Extrapolation of the curves indicates that the total diversity may be 2328. An estimated bacterial richness was reported by Naveed *et al.*, (2016) to vary depending on the location in the field, with the number of OUT reported ranged from 845 to 1675 at a 97% similarity threshold. Bacterial richness was determined in all 2,173 soils samples at 95% of sequence similarity with OUT numbers ranging from 555 to 2,007 and 1,170 and 1,424 (Terrat *et al.*, 2017). Therefore, the total diversity in this study is in the same order of magnitude with other studies of different soil environments which used comparable sequencing technology and sequencing depth (Naveed *et al.*, 2016; Siles and Margesin, 2016; Terrat *et al.*, 2017).

The Chao1 value obtained for the Kogelberg Biosphere Reserve *fynbos* soil samples were 2191 on average (Table 3.1). The Chao1 can indicate the species richness, since the distribution of rare taxa detected in a soil sample is used to extrapolate the total number of taxa present in the soil (Borneman *et al.*, 1996). It explains the abundance of singleton species which results in the greater species richness. The Chao1 value therefore, represents the level of bacterial diversity (Park *et al.*, 2018). The Chao value ranging between 2301 to 3312 were observed, when Siles and Margesin, (2016) investigated soil bacterial diversity and abundance at four Alpine forest sites using Illumina sequencing. Additionally, the soils of the Desert of Maine were explored using PCR amplified 16S rDNA genes to assess bacterial diversity, community structure and the relative abundance of bacterial taxa, where 1394 total number of OTUs at 97% sequence identity and 1145 – 1693 Chao 1 values were observed in the two samples. Compared to other studies of different soil samples (Kaiser *et al.*, 2016; Siles and Margesin, 2016; Wang, 2016), this Chao1 value observed in *fynbos* soils is frequent in studies of soil bacterial communities using high-throughput DNA sequencing techniques.

Table 3.1: **Summary of the number of sequences and diversity indices for the KBR sample, with the OTUs clustered at 97% sequence identity**

Sample	Number of reads	Average length(bp)	chao1	Total_OTUs
KBR 1	42513	251	1858	1617
KBR 2	38782	251	2456	1884
KBR 3	44644	252	2259	1962
Average	41979	251	2191	1821

3.2.3 Taxonomic analysis

The sequences of the 3 samples were classified at 6 taxonomic levels with QIIME in addition to the OUTs described above that may represent genotypes or genospecies. This taxonomy yielded 24 phyla, 70 classes, 120 orders, 189 families and 269 genera, plus a number of unclassified sequences at various taxonomic levels. Detailed descriptions of the 5 most frequent phyla (*Actinobacteria*, *Proteobacteria*, *Acidobacteria*, *Planctomycetes* and *Bacteroidetes)* for the Kogelberg Biosphere Reserve *fynbos* soil samples are presented in Table 3.2. Other phyla included *WPS-2*, *Cyanobacteria*, *Elusimicrobia*, *AD3*, *Armatimonadetes*, *Chlorobi*, *Chloroflexi*, *FBP*, *FCPU426 Fibrobacteres*, *Firmicutes*, *GAL15*, *Gemmatimonadetes*, *OP3*, *Spirochaetes*, *TM7*, *TM6* and *Verrucomicrobia*.

Table 3.2: **Representation of the most dominant bacterial taxonomic groups**

Phylum	Class	Order	Family	Genus
Actinobacteria	*Acidimicrobiia*	*Acidimicrobiales*	*EB1017*	
	Actinobacteria	*Actinomycetales*	*Actinosynnemataceae*	*Kibdelosporangium*
			Frankiaceae	*Frankia*
			Geodermatophilaceae	*Geodermatophilus*
			Intrasporangiaceae	
			Microbacteriaceae	*Mycobacterium*
			Micromonosporaceae	*Micromonospora*
			Nocardiaceae	*Nocardia*
			Nocardioidaceae	
			Propionibacteriaceae	*Propionibacterium*
			Pseudonocardiaceae	*Amycolatopsis*
				Pseudonocardia
			Sporichthyaceae	
			Streptomycetaceae	*Streptomyces*
			Streptosporangiaceae	
	MB-A2-108	*0319-7L14*	*Thermomonosporaceae*	
	Thermoleophilia	*Gaiellales*	*AK1AB1_02E*	

Phylum	Class	Order	Family	Genus
			Gaiellaceae	
		Solirubrobacterales	*Conexibacteraceae*	
			Patulibacteraceae	
Proteobacteria	*Alphaproteobacteria*	*Caulobacterales*	*Caulobacteraceae*	*Mycoplana*
				Phenylobacterium
		Rhizobiales	*Beijerinckiaceae*	
			Bradyrhizobiaceae	*Balneimonas*
			Brucellaceae	
			Hyphomicrobiaceae	*Devosia*
				Rhodoplanes
			Methylobacteriaceae	*Methylobacterium*
			Methylocystaceae	*Methylosinus*
			Phyllobacteriaceae	*Mesorhizobium*
			Rhizobiaceae	*Agrobacterium*
		Rhodobacterales	*Rhodobacteraceae*	*Rhodobacter*
				Rubellimicrobium
		Rhodospirillales	*Acetobacteraceae*	*Acidocella*
			Acetobacteraceae	*Gluconacetobacter*
			Rhodospirillaceae	*Telmatospirillum*

Phylum	Class	Order	Family	Genus
				Azospirillum
		Rickettsiales	*Rickettsiaceae*	
		Sphingomonadales	*Erythrobacteraceae*	
			Sphingomonadaceae	*Kaistobacter*
				Novosphingobium
				Sphingobium
				Sphingomonas
	Betaproteobacteria	*Burkholderiales*	*Burkholderiaceae*	*Burkholderia*
			Comamonadaceae	*Acidovorax*
				Comamonas
				Methylibium
				Pelomonas
				Ramlibacter
			Oxalobacteraceae	*Janthinobacterium*
		Neisseriales	*Neisseriaceae*	*Chitinimonas*
		Rhodocyclales	*Rhodocyclaceae*	*Dechloromonas*
				Zoogloea
	Deltaproteobacteria	*Bdellovibrionales*	*Bacteriovoracaceae*	
		Myxococcales	*0319-6G20*	

Phylum	Class	Order	Family	Genus
			Haliangiaceae	*Haliangium*
			Myxococcaceae	
			Polyangiaceae	*Sorangium*
	Gammaproteobacteria	*Enterobacteriales*	*Enterobacteriaceae*	
		Legionellales	*Coxiellaceae*	*Aquicella*
		Pseudomonadales	*Moraxellaceae*	*Perlucidibaca*
			Pseudomonadaceae	*Pseudomonas*
		Xanthomonadales	*Sinobacteraceae*	*Hydrocarboniphaga*
				Nevskia
				Steroidobacter
			Xanthomonadaceae	*Dokdonella*
				Lysobacter
	TA18	*PHOS-HD29*		*Pseudoxanthomonas*
Acidobacteria	*Acidobacteria-5*			
	Acidobacteria-6			
	Acidobacteriia	*Acidobacteriales*	*Acidobacteriaceae*	
			Koribacteraceae	*Candidatus Koribacter*
	DA052	*Ellin6513*		
	EC1113			

Phylum	Class	Order	Family	Genus
	Solibacteres	*Solibacterales*	*AKIW659*	
			Solibacteraceae	*Candidatus Solibacter*
	TM1			
	[Chloracidobacteria]	*RB41*	*Ellin6075*	
	iii1-8	*DS-18*		
	Phycisphaerae	*CPla-3*		
		WD2101		
	Pla4			
Planctomycetes	*Planctomycetia*	*Gemmatales*	*Gemmataceae*	*Gemmata*
			Isosphaeraceae	
		Pirellulales	*Pirellulaceae*	*A17*
		Planctomycetales	*Planctomycetaceae*	*Planctomyces*
	vadinHA49	*DH61*		
	Cytophagia	*Cytophagales*	*Cytophagaceae*	*Cytophaga*
				Larkinella
Bacteroidetes				*Rudanella*
				Spirosoma
				Sporocytophaga

Phylum	Class	Order	Family	Genus
	Flavobacteriia	*Flavobacteriales*	*Cryomorphaceae*	*Fluviicola*
			Flavobacteriaceae	*Flavobacterium*
			Weeksellaceae	
	Sphingobacteriia	*Sphingobacteriales*	*Sphingobacteriaceae*	*Pedobacter*
	Saprospirae	*Saprospirales*	*Chitinophagaceae*	*Flavisolibacter*

3.2.4 Taxonomic distribution of phylogenetic groups at the phylum level

The distribution of sequences at the phylum level is shown in Figure 3.2. Unclassified sequences at the phylum level represent an average of 3% of the sequences in Kogelberg Biosphere Reserve *fynbos* soil samples. The 5 dominant bacterial taxonomic groups were *Actinobacteria* (34.6%), *Proteobacteria* (32.9%), *Acidobacteria* (15.4%), *Planctomycetes* (3.0%) and *Bacteroidetes* (2.4%). The less abundant phyla included *WPS-2* (2.1%), *Cyanobacteria* (1.9%), *Elusimicrobia* (0.6%), *AD3*, *Armatimonadetes, Chlorobi, Chloroflexi, FBP, Fibrobacteres, Firmicutes, GAL15, Gemmatimonadetes, OP3, Spirochaetes, TM7* and *Verrucomicrobia*. Members of the phylum *Proteobacteria, Actinobacteria, Acidobacteria, Cyanobacteria, Bacteroidetes* and *Chloroflexi* phylotypes, were observed in our study and are generally amongst the most common inhabitants of soils (Lako, 2005; Stafford *et al.*, 2005; Slabbert *et al.*, 2010; Miyambo *et al.*, 2016; Postma *et al.*, 2016). These major phyla have also been detected in soil environments with different above-ground environmental conditions all over the world, as reported by Miyashita *et al.*, (2013). Our results, together with previous soil community studies, suggest that the compositions of soil bacterial communities, at higher taxonomic levels, are relatively similar across different soil communities.

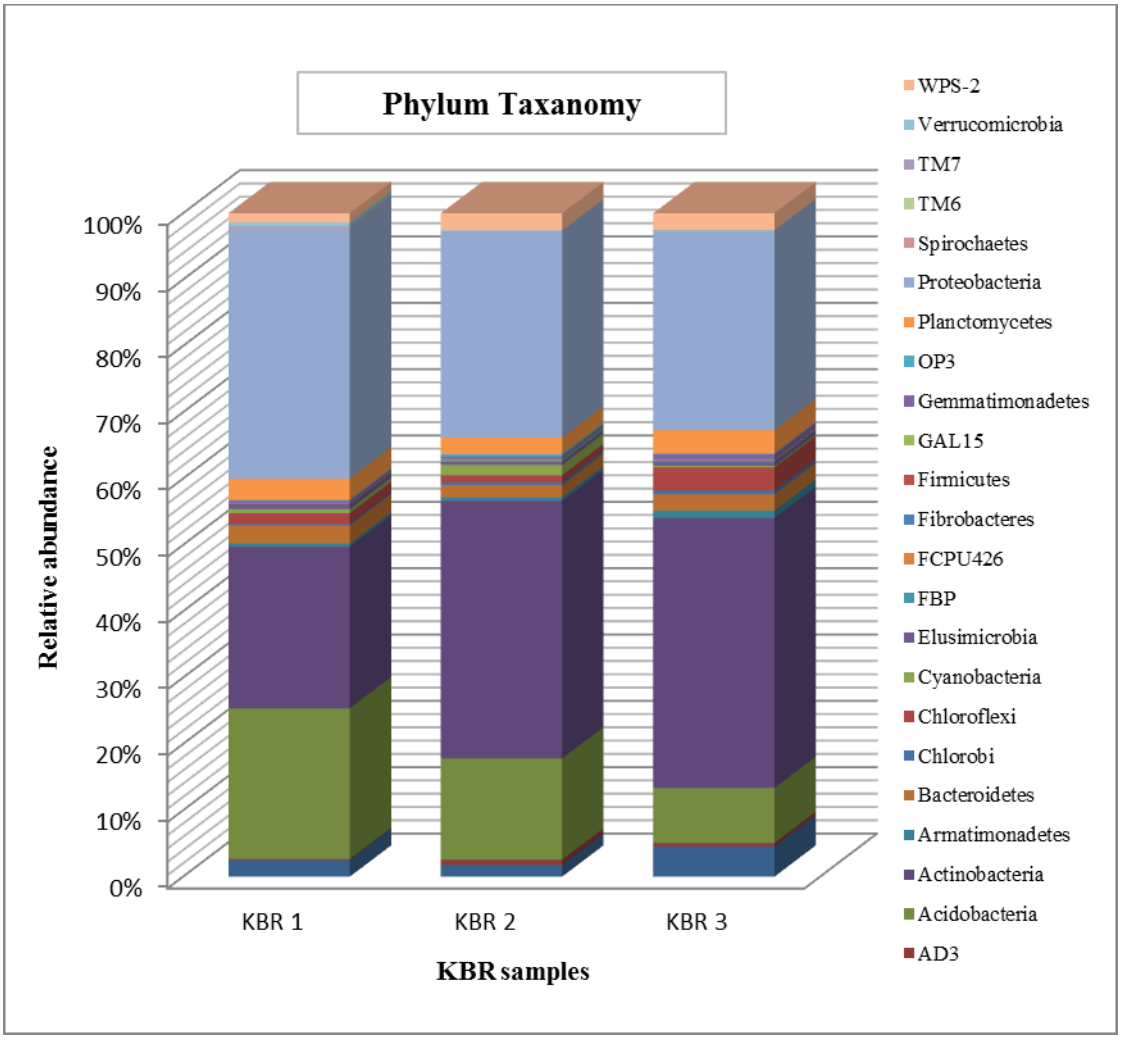


Figure 3.2: **The taxonomic distribution of phylogenetic groups at the phylum level. The candidate division phylum are : WPS-2 (candidate division wittenberg polluted soil-2), candidate division TM7 (Saccharibacteria), candidate division TM6 (belongs the microbial dark matter that gathers uncultivated bacteria detected only via DNA sequencing), OP3 (candidate phyla recovered from Obsidian Pool), GAL15 (candidate phylum), FCPU426 (unclassified bacterial candidate division), FBP (candidate phylum widespread in extreme environments), AD3 (unclassified bacterial candidate division).**

3.2.5 Phylogenetic analysis

Figure 3.3 shows a 16S rRNA phylogenetic tree depicting the taxonomic identification at a phylum level, as well as the relative abundance for the 3 KBR samples combined. The 5 most abundant taxonomic groups were *Actinobacteria*, *Proteobacteria*, *Acidobacteria*, *Planctomycetes* and *Bacteroidetes*, consistent with the taxonomic distribution of the phylogenetic groups at the phylum level (Figure 3.2 and Table 3.2). A large proportion of

unclassified sequences at the phylum level are also shown in Figure 3.2 and in the phylogenetic tree (Figure 3.3, shown by black shade). The phylogeny-based taxonomy assignment approach has proven to be the most efficient method to study taxonomic distribution of bacteria and to discover the unknown taxa (Holovachov *et al.*, 2017).

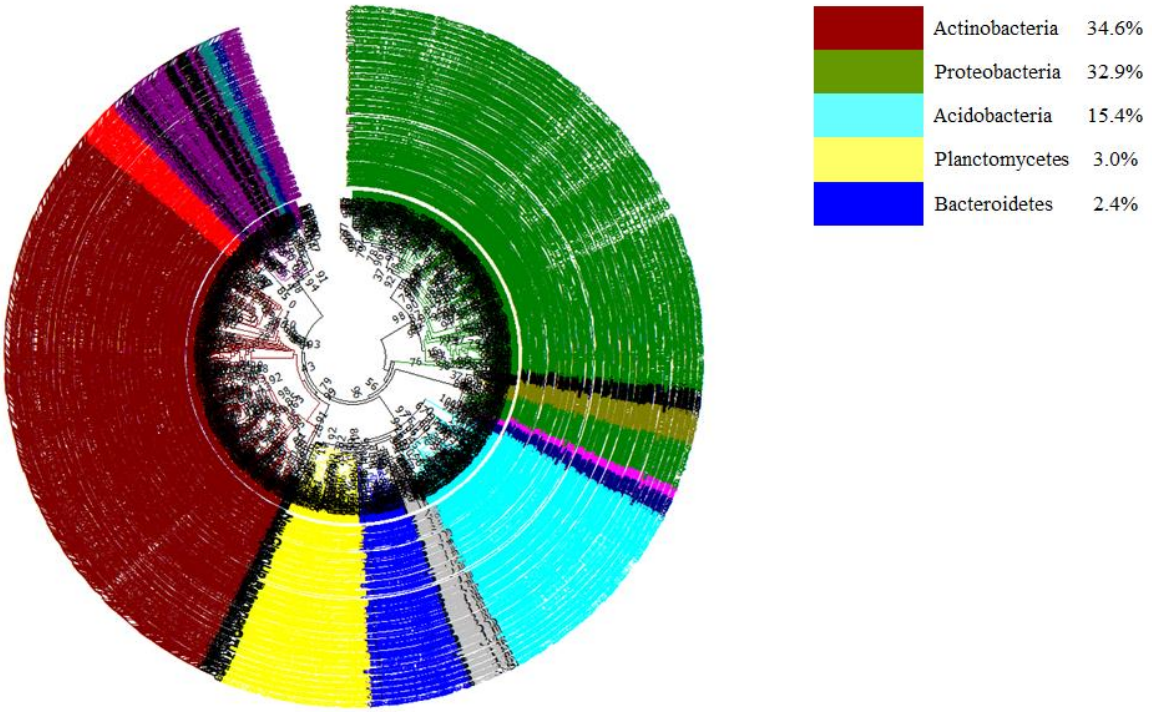


Figure 3.3: **A phylogenetic map of the microbial community in the Kogelberg Biosphere Reserve *fynbos* soil sample. The clade colours represent the taxonomic identification at a phylum level and the relative abundance for the combined library. The clade with the top 5 abundant phylum are represented by maroon shade for *Actinobacteria* (34.6%), green for *Proteobacteria* (32.9%), aqua for *Acidobacteria* (15.4%), yellow for *Plantomycetes* (3.0%) and blue for *Bacteroidetes* (2.4%). Black shading represents the unclassified sequences (3%) found in the KBR sample. The other colours in the tree represented the lower abundant bacterial phylum including *WPS-2* (2.1%), *Cyanobacteria* (1.9%), and <0% bacterial phyla (*Elusimicrobia* (0.6%), *AD3, Armatimonadetes, Chlorobi, Chloroflexi, FBP, Fibrobacteres, Firmicutes, FCPU426, TM6, GAL15, Gemmatimonadetes, OP3, Spirochaetes, TM7* and *Verrucomicrobia)*.**

3.2.6 Taxonomic distribution of phylogenetic groups at lower levels

The distribution of sequences at the class and order level for the members of the abundant phylum are shown in Figure 3.4, whereas the family and genus level are shown in Figures

B1 to B7 in the Appendices section B. Taxonomic distribution of phylogenetic groups at the phyla and class level revealed that *Actinobacteria* (class: *Thermoleophilia, Acidimicrobiia, Thermoleophilia* and *MB-A2-108*), *Proteobacteria* (class: *Alphaproteobacteria, Gammaproteobacteria, Betaproteobacteria* and *Deltaproteobacteria*), *Acidobacteria* (class: *Acidobacteria* and *DA052*), *Plantomycetes* (class: *Plantomycetia*) and class *Bacteroidetes* were observed in our sample (Table 3.2). On the order, family and genus level of *Actinobacteria*, the orders *Solirubrobacterales, Gaiellales, Actinomycetales* and *Acidimicrobiales*, the families *Actinosynnemataceae Conexibacteraceae* and *Frankiaceae* and the genera *Kibdelosporangium, Frankia* and *Micromonospora* were mostly predominant. Whereas for *Proteobacteria*, the orders *Rhodospirillales, Rhizobiales, Burkholderiales, Xanthomonadales, Caulobacterales* and *Ellin329*, the families *Rhodospirillaceae, Bradyrhizobiaceae Oxalobacteriaceae* and *Comamonadaceae* and the genus *Rhodoplanes, Janthobacterium, Burkholderiales* and *Acidovorax* dominated (Table 3.2). Furthermore, for *Acidobacteria*, the orders *Acidobacteriales* and *Solibacterales*, the families *Acidobacteriaceae, Koribacteraceae, Solibacteraceae* and the genera *Candidatus Koribacter* and *Candidatus Solibacter* were most abundant (Table 3.2). For the phylum *Plantomycetes*, the most dominating order was *Planctomycetale* and the family *Gemmataceae* (Table 3.2).

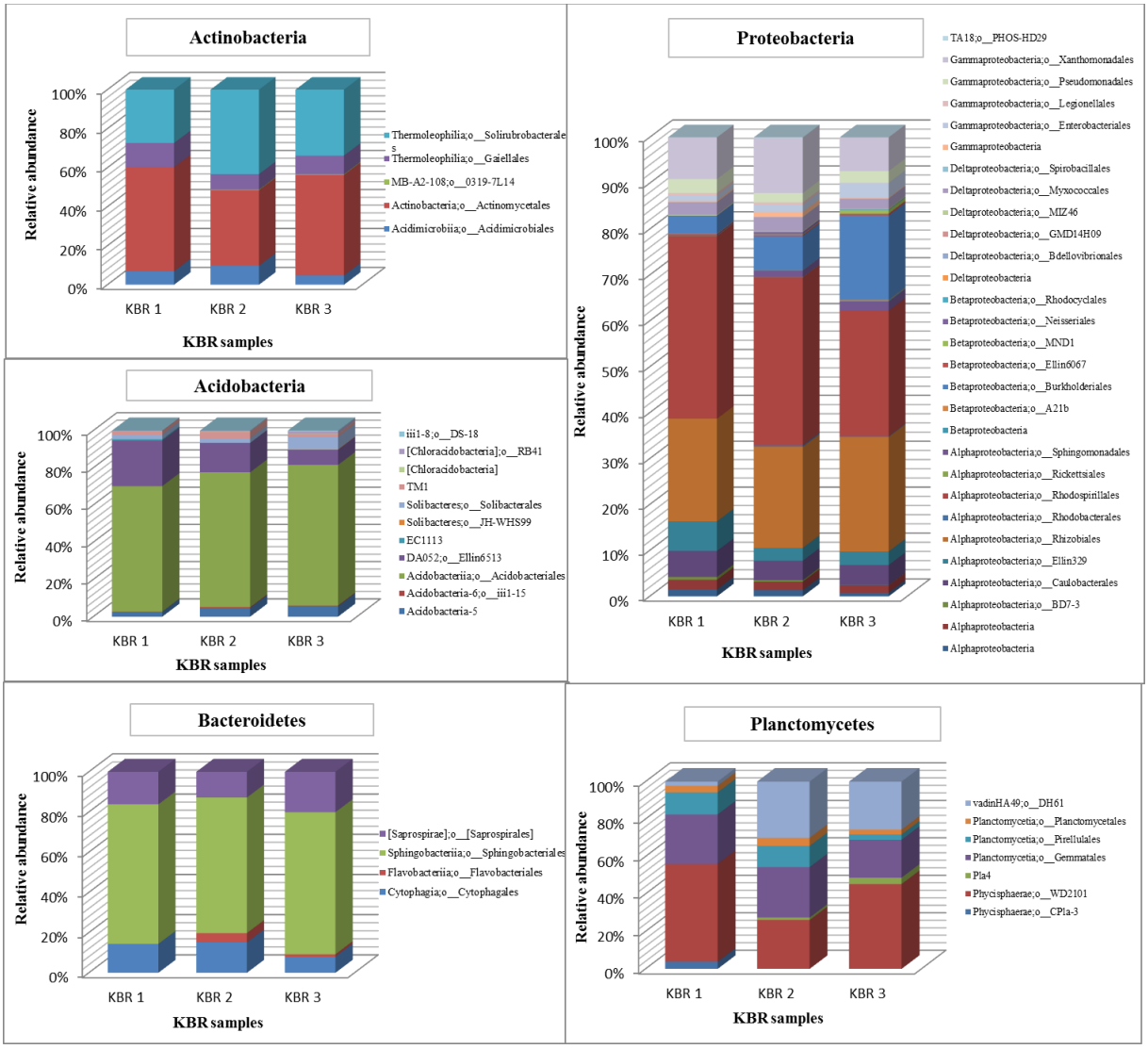


Figure 3.4: **Bar charts representing the taxonomic distribution of phylogenetic groups at the Class and Order level (A) *Proteobacteria*, (B) *Actinobacteria*, (C) *Acidobacteria* and (D) *Plantomycetes* and *Bacteriodetes*.**

3.2.7 Taxonomic abundance in different environmental samples

In order to confirm the ability of 16S rRNA NGS to reveal a snapshot of diversity of the dominant bacterial populations from Kogelberg Biosphere Reserve *fynbos* soil samples, the microbial community in the *fynbos* soil samples was compared to microbial communities in different types of biomes available in the public database accessed through the MG-RAST server. Unlike QIIME, MG-RAST allows comparison and investigations of other publicly available datasets, with no requirement for access to a powerful computer (Keegan *et al.*, 2016). MG-RAST left more than 50% of the reads unclassified at the phylum level for the Kogelberg Biosphere Reserve soil sample, significantly a lot more than QIIME.

Furthermore, a different taxonomic abundance was observed when using MG-RAST, however, *Actinobacteria, Proteobacteria, Acidobacteria* and *Bacteriodetes* were still dominating Kogelberg Biosphere Reserve *fynbos* soil sample communities. Comparison of taxonomic compositions and diversity measures analysed by QIIME and MG-RAST was conducted by D'Argenio *et al.*, (2014) and Plummer *et al.*, (2015) and they concluded that MG-RAST produce high number of reads that are unable to be classified, whereas, QIIME produced more accurate assignments (Plummer *et al.*, 2015). These studies highlight the impact of bioinformatics pipeline for 16S rRNA gene sequencing data analysis.

Figure 3.5 shows the bacterial phyla abundances of the Kogelberg Biosphere Reserve *fynbos* soil sample compared to different sample types, where the samples were prepared by 16S rRNA NGS techniques. *Proteobacteria* and *Bacteriodetes* were detected in all the different samples. *Proteobacteria, Actinobacteria, Bacteriodetes* and *Firmicutes* dominated the soil samples and the wastewater samples. Water samples were dominated by *Proteobacteria, Bacteriodetes* and *Firmicutes*, whereas *Actinobacteria, Bacteriodetes* and *Firmicutes* dominated the faeces sample. Different biomes were evidently dominated by distinct bacterial populations; while similar biomes were dominated by similar bacterial populations (Figures 3.5).

It is evident that 16S rRNA NGS amplicon data allow for the identification of drivers of bacterial community composition in the environmental samples (Vierheilig *et al.*, 2015). Alternative methods for the identification of bacterial community structures such as DGGE and T-RFLP were used previously to highlight the dominant bacterial populations (Stafford *et al.*, 2005; Makhalanyane *et al.*, 2013). However, these techniques have much lower resolution and do not provide sequence information directly. Furthermore, 16S rRNA NGS approaches provides depth and much more information density. In addition, the information obtained from this study provides an indication of the abundance and diversity of bacterial communities in different environmental samples using the 16S rRNA gene.

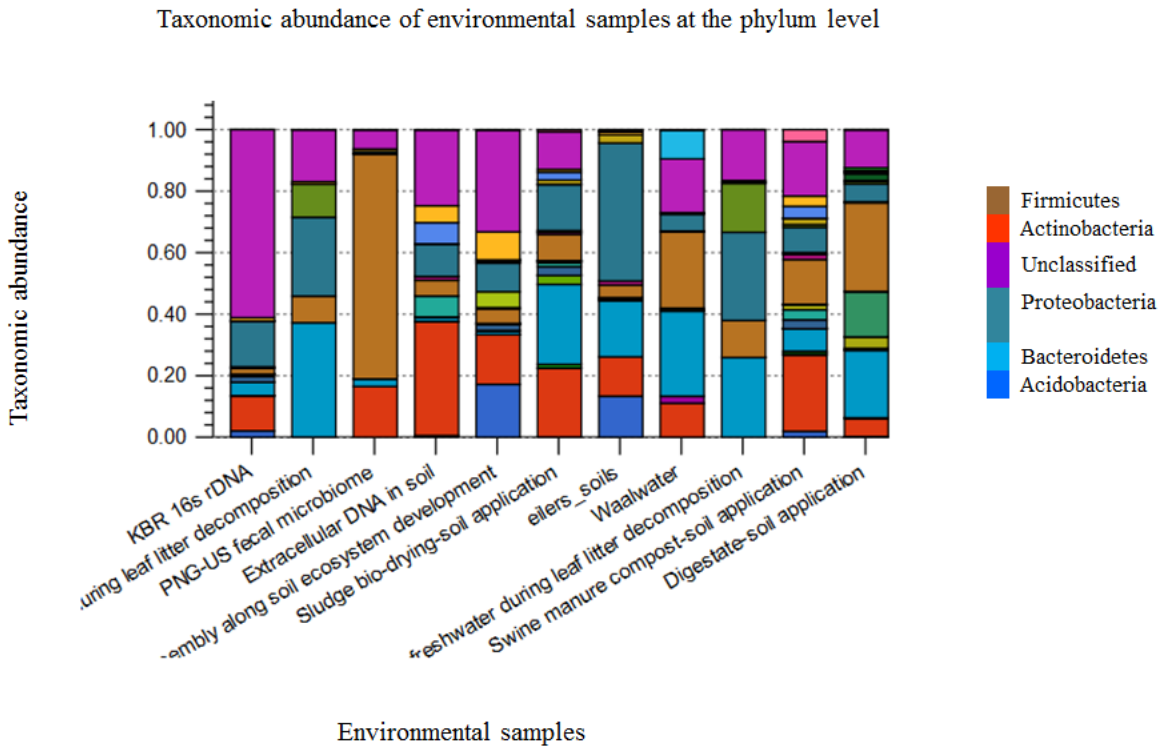


Figure 3.5: **Bacterial phyla found in the Kogelberg Biosphere Reserve Biome compared to publicly available bacterial phyla of different samples from different biomes. Sample 1= soil sample from Kogelberg Biosphere Reserve, Sample 2 and 9 = water samples from Wet Beaver creek (USA), Sample 3 = Faeces samples from Asaro Valley (Papua New Guinea), Sample 4 = soil sample from Flagstaff (USA), Sample 5 = soil sample from Svalbard (Norway), Sample 6, 10 and 11 = wastewater/sludge samples from Beijing (China), Sample 7 = soil sample from Cedar Creek Natural History Area in Minnesota (USA) and Sample 8 = water sample from Nijmegen (Netherlands).**

3.2.8 Principal coordinate analysis (PCoA)

The PCoA plot with Bray-Curtis was used to compare dissimilarity of the microbial community of different environments compared to the Kogelberg Biosphere Reserve *fynbos* soil samples. Irrespective of taxonomic assignment, Bray-Curtis can provide a measure of differences of community composition between samples based on OTU counts. PCoA was used for cluster analysis to visualise which communities are more closely related and which are more distinct (Figure 3.6). Results revealed that Kogelberg Biosphere Reserve *fynbos* soil samples clustered separately from water, wasterwater/sludge and feces samples. However, wastewater samples did not cluster together and with the other samples. Fresh water samples clustered together, at a distant from other samples. Feces sample is detected

far from all the samples. Clustering of similar samples indicate similar taxonomic abundance between the samples. The results demonstrates the similarity of microbial communities' distribution in the soil samples and their significant difference from the other environmental samples. One significant factor that influences the microorganisms' biological diversity could be the nature and origin of the samples. From this analysis, the *fynbos* soils are observed to be more similar to the diversity of other soils previously studied, compared to fresh water, wastewater/sludge and humangut samples.

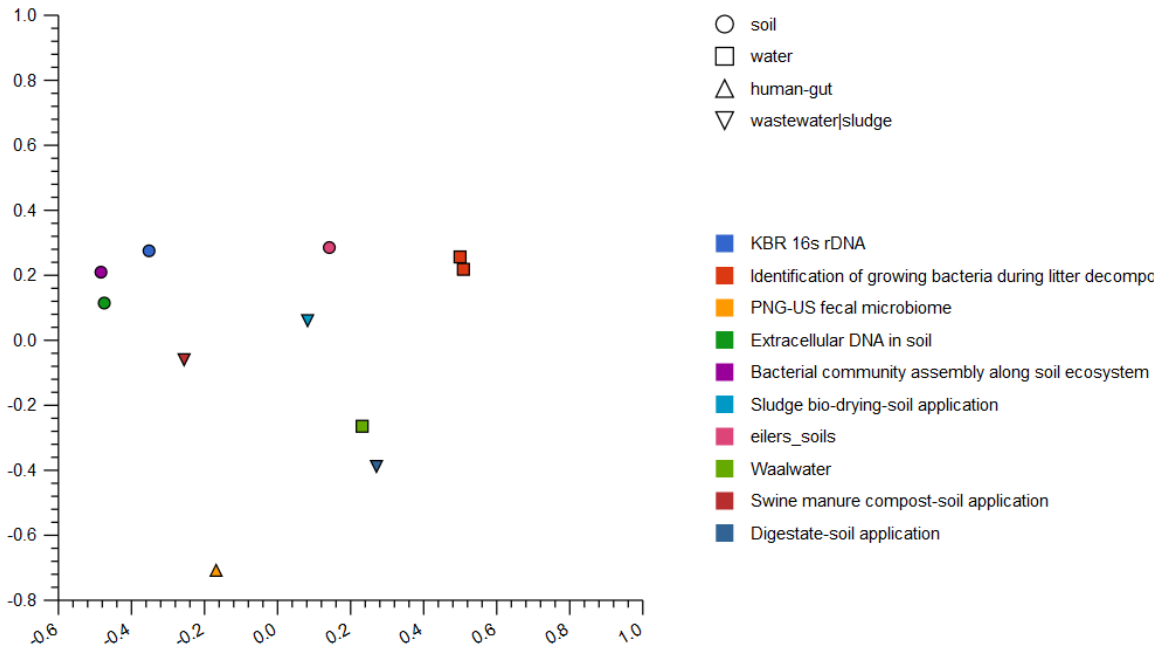


Figure 3.6: **PCoA plot of differences between the microbial communities amongst the 5 biomes based on OTU relative abundance at the phylum level using Bray-Curtis method. Blue = soil sample from Kogelberg Biosphere Reserve, Red = water samples from Wet Beaver creek (USA), Yellow = Faeces samples from Asaro Valley (Papua New Guinea), Dark green = soil sample from Flagstaff (USA), Purple= soil sample from Svalbard (Norway), Turquoise, maroon and dark blue = wastewater/sludge samples from Beijing (China), pink = soil sample from Cedar Creek Natural History Area in Minnesota (USA) and light green = water sample from Nijmegen (Netherlands).**

Conclusion

The present study used 16S rRNA amplicon and NGS techniques to examine the taxonomical abundance and diversity of bacterial communities present in the *fynbos* soil

from Kogelberg Biosphere Reserve. The Kogelberg Biosphere Reserve harbours unique plant biodiversity and may contain distinct and unique microbial communities.

With 10 000 rRNA gene sequences analysed, 1821 genospecies or Operational Taxonomic Units (OTUs) were observed that had sequence similarity to existing database entries. Overall, 3% sequences were unclassified and had <97% sequence similarity with rRNA genes in public databases and therefore may be considered novel genospecies. The rarefaction analysis revealed that there is under-sampling of the *fynbos* soil metagenome. From the extrapolated rarefaction analysis, the total diversity may be 2328 in terms of genospecies. This study is therefore a snapshot of the diversity of bacteria in *fynbos* soils and there is considerably greater diversity in *fynbos* soil that was revealed by this study. Nonetheless, the soil samples appeared to be dominated by *Actinobacteria, Proteobacteria, Acidobacteria, Plantomycetes* and *Bacteriodetes*. Comparative analysis of the Kogelberg Biosphere Reserve *fynbos* soil sample with samples from different global environments enhances our understanding of microbial diversity and further provides an opportunity to understand how microbial diversity is reflective of the characteristics of environment from which the samples was obtained, and may enhance our understanding of microbial ecology.

CHAPTER 4: EXPLORING VIRAL DIVERSITY IN A UNIQUE SOUTH AFRICAN SOIL HABITAT

4.1 Introduction

The Cape Floristic Region situated in the Western Cape province of South Africa is one of five Mediterranean-type ecosystems in the world (Bergh and Compton, 2015) and is recognized as one of the world's biodiversity hotspots (Slabbert *et al.*, 2010). *Fynbos* (fine bush) is the main vegetation type of this region with the *Proteaceae*, *Ericaceae* and *Restionaceae* families dominating Kogelberg Biosphere Reserve *fynbos* vegetation. Within this region, the *fynbos* comprises approximately 9000 plant species of which 70% are endemic to the region (Van Wyk and Smith, 2001; Bergh and Compton, 2015). *Fynbos* vegetation types survive on highly heterogeneous, acidic, sandy, well-leached and infertile soils. The *fynbos* plants also survive invasions by foreign plants (Sprent and Parsons, 2000) and seasonal drought conditions (Mucina and Wardell-Johnson, 2011).

Microorganisms make up a great proportion of the living population in the biosphere. They provide important ecosystem services in edaphic habitats (Jeanbille *et al.*, 2016) and form complex symbiotic relationships with plants (Rappé and Giovannoni, 2003). Plant-associated microorganism studies have shown high microbial diversity in *fynbos* soils (Slabbert *et al.*, 2010), where they play a role in sustaining plant communities (van der Heijden *et al.*, 2008). A study focusing on the linkage between *fynbos* soil microbial diversity and plant diversity showed the presence of novel taxa and of bacteria specifically associated with the rhizospheric zone (Stafford *et al.*, 2005). Studies on ammonium-oxidizing bacteria demonstrated that plant-species specific and monophyletic ammonium oxidizing bacterial clades were present in *fynbos* soils (Ramond *et al.*, 2015), where abundance might be driven by the acidic and oligotrophic nature of these soils (Prosser and Nicol, 2008). There is evidence that above-ground floral communities are implicated in shaping microbial communities (Nüsslein and Tiedje, 1999; Hamilton and Frank, 2001), and that some microbial clades show a high level of plant–host specificity (Ramond *et al.*, 2015). This is consistent with the general concept of the mutualistic relationships between the plants and the microbial communities in *fynbos* soils (Keluskar *et al.*, 2013).

CHAPTER 4 EXPLORING VIRAL DIVERSITY IN A UNIQUE SOUTH AFRICAN SOIL HABITAT

Soil-borne viruses, including phages, are of great importance in edaphic habitats due to their ability to transfer genes from host to host and as a potential cause of microbial mortality (leading to changes in turnover and concentration of nutrients and gases), processes that can profoundly influence the ecology of soil biological communities (Kimura *et al.*, 2008). Virus diversity associated with *fynbos* plants from Kogelberg Biosphere Reserve *fynbos* soil has never been thoroughly investigated (Cowan *et al.*, 2013). The difficulty of culturing viruses, which are absolutely dependent on a cell host to provide the apparatus for replication and production of progeny virions, presents a barrier to fully accessing viral biodiversity. This is a particular issue in poorly studied habitats, such as *fynbos* soil, where the true microbial (host) diversity is largely unknown and most microbial phylotypes have never been cultured (Schoenfeld *et al.*, 2010). The biodiversity and ecology of viruses in many soils therefore remain poorly investigated and poorly understood (Zablocki *et al.*, 2016).

Metaviromic surveys of terrestrial environments such as hot desert soil (Zablocki *et al.*, 2016), rice paddy soil (Kim *et al.*, 2008, 2013), Antarctic cold desert soil (Srinivasiah *et al.*, 2013; Zablocki *et al.*, 2014) and hot desert hypolithic niche communities (Adriaenssens *et al.*, 2015) have been reported in recent years and have significantly advanced the field of soil viral ecology (Fancello *et al.*, 2013; Kim *et al.*, 2013). These studies have also facilitated the discovery of novel virus genomes (Kim *et al.*, 2013; Zablocki *et al.*, 2014; Adriaenssens *et al.*, 2015) and novel viral enzymes (Gudbergsdóttir *et al.*, 2015).

However, surveys of viral diversity using NGS sequencing techniques in conjunction with metaviromic databases have focused principally on aquatic environments (Rodriguez-Brito *et al.*, 2010; Breitbart, 2012; Roux *et al.*, 2012). Studies on taxonomic composition using public metaviromic databases for viral diversity estimations have shown that a majority of environmental virus sequences are unknown (Kim *et al.*, 2008): ~70% of sequences have no homologs in public databases and are therefore typically labelled “viral dark matter” (Hatfull, 2015; Simon Roux *et al.*, 2015). Bacteriophages constitute the largest known group of viruses found in both aquatic (Alhamlan *et al.*, 2013; Fancello *et al.*, 2013) and soil environments (Williamson *et al.*, 2005; Reavy *et al.*, 2015).

Here we report the first investigation of virus diversity in a unique soil type (*fynbos* soil) using metaviromic approaches. The metavirome of Kogelberg Biosphere Reserve *fynbos* soil was

characterised in terms of diversity and functional composition and adds a new level of understanding to the exceptional biodiversity of this habitat.

4.2 Results and Discussion

4.2.1 Viral morphology

Analysis of the morphology of viruses identified in Kogelberg Biosphere Reserve *fynbos* soil was carried out by transmission electron microscopy (TEM). TEM analysis of the virus preparations showed that the majority of the isolated virus particles were morphologically similar to known virus taxonomic groups (Ackermann, 2007). The isolated virus particles from the *fynbos* soil were tailed, spherical or filamentous (Appendices C, Figure S1). Various particles with head-tail morphology, typically belonging to the families *Myoviridae*, *Siphoviridae* or *Podoviridae*, were observed.

These results are in a good agreement with previously published findings showing the high dominance of tailed phages in soils from various geographic areas (Fancello *et al.*, 2013; Reavy *et al.*, 2015; Zablocki *et al.*, 2015). The undetermined spherical or filamentous morphologies in TEM micrographs could be *bona fide* but uncharacterised viral structures. Spherical particles resembling capsid structures could be members of the *Leviviridae, Partitiviridae, Chrysoviridae, Totiviridae* or *Tectiviridae* families, or small plant viruses (Ackermann, 2007). Filamentous particles may possibly correspond to the virus structures of the *Inovirus* genus, the members of which contain circular ssDNA within flexible filamentous virions. The presence of spherical types and filamentous type of virus particles was also reported for Delaware soils (Williamson *et al.*, 2005). The aggressive extraction procedure used in the current study may have resulted in a high incidence of phage tail breakage and the generation of tailless phages (Ackermann, 2006).

4.2.2 Metavirome assembly

Assembly of the DNA sequence reads yielded 13,595 contigs larger than 500bp, with an average length of 2,098bp, accounting for a total of 28,526,478bp (Table 1). Two different metagenomics pipelines; MetaVir (Roux *et al.*, 2014) and VIROME (Wommack *et al.*, 2012), were used for analysis of the contigs, while MG-RAST (Meyer *et al.*, 2008) was used for the analysis of the uploaded reads (Table 2). The MetaVir pipeline predicted 51,274 genes, with

5,338 affiliated contigs (i.e., contigs with at least one BLAST hit) and 7880 unaffiliated contigs (Table 2). MetaVir compares reads/contigs to complete viral genomes from the Refseq database and is specifically designed for the analysis of environmental viral communities (Roux *et al.*, 2014). The VIROME pipeline (Wommack *et al.*, 2012) predicted 51,242 protein coding regions. Of these, 9555 were assigned as functional proteins, and 31,109 were unassigned (Table 2). Comparisons of functional and taxonomic analysis between VIROME and MetaVir indicate that many of the predicted genes were overlapping between the two pipelines with MetaVir on average having a higher predictive potential (Appendices C Table S1). The MG-RAST pipeline predicted 2,555,524 protein coding regions. Of these predicted protein features, 119,220 were assigned a functional annotation using protein databases (M5NR) (Wilke *et al.*, 2012) and 2,362,076 had no significant similarities to sequences in the protein databases (ORFans). MG-RAST core analysis and annotation depends heavily on the SEED database which is largely comprised of bacterial and archaeal genomes (Overbeek *et al.*, 2004). The majority of the annotated sequences in MG-RAST were mapped to bacterial genomes. This high percentage of bacterial sequences in metaviromes may be due to the presence of unknown prophages in bacterial genomes, phages carrying host genes, relatively large size of bacterial genomes compared to viral genomes and larger size of the microbial genome database which is statistically increasing the chance of matching bacterial sequences. The MG-RAST pipeline was used to analyse the reads, not the contigs and shows, therefore a higher number of predicted features, including more partial CDSs (Mohiuddin and Schellhorn, 2015). No rDNA sequences were found with the MG-RAST and VIROME pipelines, confirming the viral origins of the DNA. The fact that more than 80% of the hits in this study, consistent with previous viral metagenomics studies (Breitbart *et al.*, 2002; Cann *et al.*, 2005; Alhamlan *et al.*, 2013), were assigned as hypothetical proteins derived from unknown viruses suggests the presence of a substantial pool of novel viruses.

Table 4.1: Next Generation sequencing data analysis. Representation of the assembly, annotation, and diversity statistics produced by CLC Genomics

Features	CLC
#Pre-QC Sequence reads	7,019,527
#Pre-QC sequence in base pairs	1,488,462,918
#post-QC average read length	212.05
#contigs	13,595
#contigs/reads in bp	28,526,478bp

Table 4.2: Comparison of the automated pipelines; such as MetaVir (contigs), VIROME (contigs) and MG-RAST (reads), used to characterize the Kogelberg Biosphere Reserve.* Affiliated CDS are CDS with homologues in at least one of the databases used, while ORFans are predicted ORFs which have no database homologue.

Features	MetaVir	MG-RAST	VIROME
#predicted CDS	51,274	2,555,524	51,242
#affiliated CDS*	5,868	119,220	9,555
#ORFans*	45,406	2,362,076	31,109
#rRNAs	NA	0	0
Database used for CDS annotation	RefSeq virus, pfam	GenBank, IMG, KEGG, PATRIC, RefSeq, SEED, SwissProt, TrEMBL, eggNOG, COG, NOG, KOG,	KEGG, SEED, COG, GO, UniRef100, PHGSEED, MgOI, ACLAME

4.2.3 Viral diversity estimation and taxonomic composition

The rarefaction curve computed by MG-RAST showed 3952 species clusters at 90% sequence identity for the 3,095,000 reads. The curve did not reach an asymptote (Fig 1), although extrapolation suggested that approximately 78% of the viral diversity was covered by the metavirome sequence dataset.

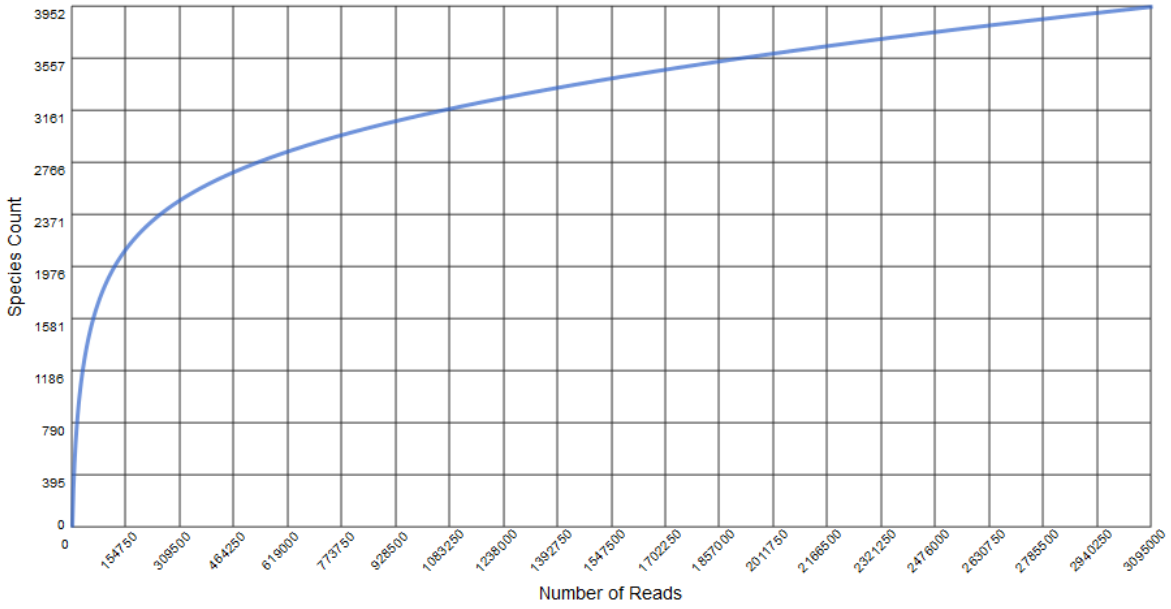


Figure 4.1: Rarefaction curve of the Kogelberg Biosphere Reserve *fynbos* soil metavirome. Clustering was set at 90% similarity.

MetaVir was used for viral taxonomic composition analysis of the contigs. The taxonomic composition was computed from a BLASTp comparison of the predicted proteins in the contigs with the Viral Refseq protein database (release of 2016-01-19). The results revealed that 37.6% of the contigs represented a significant hit (threshold of 50 on the BLAST bit score). MetaVir identified 18 virus families, in which prokaryotic viruses were the most abundant and dominated by the order *Caudovirales*, consistent with the TEM observations. The relative abundance ranking of the different families was as follows: tailed bacteriophage families *Siphoviridae* > *Myoviridae* > *Podoviridae*, followed by the algae-infecting family *Phycodnaviridae*, the archaeal virus family *Ampullaviridae* and the amoeba-infecting family *Mimiviridae* (Table 3). Surprisingly, large viruses belonging to the families *Phycodnaviridae* and *Mimiviridae* were detected, which should have been removed during the filtration process due to the use of a 0.22µm filtration step to remove bacterial cells. The identification of

Mimiviridae suggests that this filtration process allowed partial mimivirus particles or free-floating DNA to pass through the membrane. Mimiviruses appear to infect only species of *Acanthamoeba*, which are ubiquitous in nature and have been isolated from diverse environments including freshwater lakes, river waters, salt water lakes, sea waters, soils and the atmosphere (Gascuel *et al.*, 1997)(Ghedin and Claverie, 2005; Short and Short, 2008; Zablocki *et al.*, 2015). This suggests the existence of Mimivirus relatives in the KBR soil.

Other viral families and unclassified viruses (dsDNA and ssDNA) were found in low numbers. Putative contamination of *Enterobacteria* phage phiX174 was also detected in our metavirome sequences. This phage is used for quality control in sample preparation for high-throughput sequencing. Seven sequences from this dataset are similar to the phiX174 genome and were thus disregarded in the taxonomic composition as an artefact of sample processing. Plant viruses were not identified in the dataset, most probably because the majority of plant viruses are RNA viruses which were not sampled in this study.

Table 4.3: Taxonomic abundance. Representation of taxonomic abundance of identified viral ORFs BLASTp with threshold of E value10-5 identified by MetaVir.

Virus Order and family	Hosts	Relative abundance of taxa
***C**audovirales*		
Myoviridae	Bacteria, Archaea	29
Podoviridae	Bacteria	23
Siphoviridae	Bacteria, Archaea	45
Herpesvirales		
Herpeviridae	Vertebrates	0.04
Virus Family and groups not assigned in to Order		
Phycodnaviridae	Algae	2
Ampullaviridae	Archaea	0.9
Mimiviridae	Amoebae	0.8

Salterprovirus	Archaea	0.7
Tectiviridae	Bacteria, Archaea	0.5
Iridoviridae	Vertebrates (Amphibians, Fishes), Invertebrates	0.1
Marseilleviridae	Amoeba	0.04
Nudiviridae	Arthropods	0.04
Poxviridae	Human, Arthropods, Vertebrates	0.02
Baculoviridae	Invertebrates	0.02
Bicaudaviridae	Archaea	0.02
Turriviridae	Archaea	0.02
Asfarviridae	Swine	0.02
Retroviridae	Vertebrates	0.02
Virus not assigned into Family		
Unclassified dsDNA phages	Bacteria	2
Unclassified dsDNA virus	NA	4
Unclassified ssDNA Viruses	NA	0.07
Unclassified phages	Bacteria	2

The viral composition of Kogelberg Biosphere Reserve *fynbos* soil was compared to 12 previously published metaviromes from both similar and dissimilar environments, including fresh water (Roux *et al.*, 2012), soil and hypolithic niche communities (Zablocki *et al.*, 2014; Adriaenssens *et al.*, 2015), pond water (Rodriguez-Brito *et al.*, 2010) and sea water (Angly *et al.*, 2006) (Fig 2). A comparative metaviromics approach was used to investigate the assumption that certain environments will select for specific viruses (de Wit and Bouvier, 2006; Dinsdale *et al.*, 2008).

CHAPTER 4 EXPLORING VIRAL DIVERSITY IN A UNIQUE SOUTH AFRICAN SOIL HABITAT

SAMPLE TYPES	SOIL		HYPOLITH		DEEP SEA		FRESH WATER						POND
Taxonomy	KBR (s)	AOS (s)	AH	NH	ALOHA (ds)	B47 (ds)	LB (fw)	LP (fw)	57th-1 (fw)	M1 (fw)	M2 (fw)	Far (fw)	SP (p)
Viruses	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
Retro-transcribing viruses	0.000	0.000	0.020	0.000	0.300	0.010	0.000	0.000	0.010	0.030	0.080	0.000	0.000
Satellites	0.000	0.000	0.000	0.000	0.000	0.000	0.020	0.040	0.000	0.000	0.000	0.000	0.000
dsDNA viruses, no RNA stage	97.00	96.00	97.00	85.00	96.00	98.00	95.00	94.00	98.00	98.00	98.00	100.00	100.00
Caudovirales	89.00	80.00	89.00	81.00	52.00	81.00	77.00	79.00	61.00	66.00	65.00	72.00	24.00
Siphoviridae	39.00	37.00	59.00	47.00	16.00	28.00	27.00	28.00	19.00	20.00	20.00	9.00	19.00
Myoviridae	26.00	30.00	18.00	16.00	52.00	27.00	26.00	28.00	33.00	34.00	27.00	56.00	5.00
Podoviridae	20.00	11.00	10.00	5.00	7.00	24.00	21.00	20.00	7.00	11.00	16.00	6.00	0.000
Ampullavirus	1.00	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Mimiviridae	0.800	0.000	1.00	0.000	0.000	1.00	0.080	1.00	0.000	0.000	0.000	0.000	0.000
Salteprovirus	0.700	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.005	0.000	0.000
Tectivirus	0.500	0.070	0.030	0.080	0.000	0.010	0.040	0.000	0.020	0.004	0.040	0.000	0.000
Iridoviridae	0.100	0.400	0.090	0.080	0.600	0.200	0.200	0.300	0.500	0.600	0.600	0.000	0.000
Marseilleviridae	0.040	0.300	0.050	0.040	0.090	0.060	0.050	0.040	0.300	0.400	0.300	0.000	0.000
Herpesvirales	0.040	0.100	0.100	0.000	0.500	0.040	0.050	0.040	0.500	0.400	0.300	0.000	0.000
Nudiviridae	0.020	0.070	0.000	0.000	0.000	0.000	0.000	0.000	0.300	0.200	0.100	0.000	0.000
Poxviridae	0.020	0.400	0.060	0.040	2.00	0.200	0.070	0.100	0.000	1.00	1.00	0.000	0.000
Baculoviridae	0.020	0.100	0.080	0.000	0.200	0.070	0.020	0.000	0.200	0.200	0.300	0.000	0.000
Asfivirus	0.020	0.070	0.020	0.000	0.000	0.010	0.000	0.000	0.040	0.020	0.300	0.000	0.000
Polydnaviridae	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.008	0.000	0.000	0.000
Adenoviridae	0.000	0.000	0.020	0.000	0.000	0.000	0.000	0.000	0.000	0.004	0.005	0.000	0.000
White spot syndrome virus	0.000	0.070	0.000	0.000	0.000	0.000	0.020	0.000	0.020	0.020	0.010	0.000	0.000
Fuselloviridae	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.005	0.000	0.000
Ascovirus	0.000	0.300	0.030	0.000	0.000	0.100	0.000	0.000	0.800	0.020	0.500	0.000	0.000
Polyomaviridae	0.000	0.000	0.000	0.000	0.000	0.010	0.000	0.000	0.000	0.000	0.009	0.000	0.000
Bicaudaviridae	0.000	0.000	0.050	0.000	0.300	0.000	0.000	0.000	0.070	0.040	0.070	0.000	0.000
Corticovirus	0.000	0.000	0.000	0.000	0.000	0.010	0.020	0.000	0.000	0.000	0.005	0.000	0.000
Ligamenvirales	0.000	0.000	0.030	0.000	0.000	0.040	0.020	0.040	0.030	0.020	0.020	0.000	0.000
Plasmavirus	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.010	0.000	0.000	0.000
dsRNA viruses	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.040	0.005	0.008	0.009	0.000	0.000
Environmental samples	0.000	0.000	0.000	0.000	0.000	0.000	0.030	0.100	0.000	0.000	0.000	0.000	0.000
ssDNA viruses	0.080	0.700	0.800	0.500	0.090	0.100	2.00	3.00	0.400	0.300	0.200	0.000	0.000
ssRNA viruses	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Unassigned viruses	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Unclassified archaeal viruses	0.000	0.000	0.000	0.000	0.090	0.000	0.000	0.000	0.000	0.000	0.009	0.000	0.000
Unclassified phages	3.00	3.00	3.00	0.000	3.00	2.00	3.00	2.00	2.00	2.00	2.00	0.000	0.000
Unclassified virophages	0.000	0.400	0.020	0.000	0.000	0.500	0.300	0.800	0.020	0.030	0.040	0.000	0.000
Unclassified viruses	0.000	0.000	0.000	0.000	0.000	0.020	0.000	0.000	0.000	0.000	0.000	0.000	0.000

Figure 4.2: Comparison of the Kogelberg Biosphere Reserve metavirome taxonomic composition with selected publicly available metaviromes. Abundances normalized according to predicted genome size with the GAAS tool. Blue colour represents 0.000 taxon, yellow represents 0.01 – 19.00, mustard represents 20.00 – 29.00, light red represents 30.00 – 49.00, and red represents 50.00 – 100.00 taxon. More details on the description of metaviromes are described in Supplementary Table S3 online.

The *Caudovirales* taxon dominated all metaviromes. In particular, members of the family *Siphoviridae* were dominant in most metaviromes except for some of the freshwater samples, in which myoviruses were dominant. Within the dsDNA viruses, members of rare taxonomic groupings such as the genera *Tectivirus, Asfivirus* and *Salterprovirus*, the families *Mimiviridae, Iridoviridae, Marselleviridae, Nudiviridae, Poxviridae* and *Baculoviridae* and the order *Herpesvirales* were detected in soil samples as well as in hypolith, deep sea, and freshwater metaviromes. Archaeal virus signatures belonging to the family *Ampullaviridae* have been observed only in the Kogelberg Biosphere Reserve *fynbos* soil. This family contains viruses with pleomorphic morphologies and a dsDNA genome, and the type species infects the thermoacidophile *Acidianus convivator,* isolated from Italian hot springs (Häring *et al.*, 2005). Fresh Water Lake, Antarctic soil and coral metaviromes showed a high abundance of ssDNA viruses, results possibly biased by the use of phi29 polymerase amplification (MDA) of the metaviromic DNA during library construction. The amplification of metaviromic DNA using phi29 polymerase amplification (Multiple Displacement Amplification) has been reported to be biased towards ssDNA templates (Kim *et al.*, 2008). It is notable, however, that a high abundance of ssDNA viruses has been observed in beach freshwater samples (Watkins *et al.*, 2016), where amplification was not used in the preparation of metagenomic DNA. However, in general, other metaviromes which were not amplified using MDA showed a very low number of ssDNA viruses. In general, soils or soil-associated habitats seem to harbour relatively fewer ssDNA viruses and more tailed phages than aquatic ecosystems.

Consistent with other data (Breitbart *et al.*, 2002; Fancello *et al.*, 2013; Zablocki *et al.*, 2014), it was found that bacteriophage sequences in Kogelberg Biosphere Reserve *fynbos* soil made

up the majority of the virus fraction. Bacteriophages are common in the environment and are the dominant viral type recovered from metaviromics analyses in soil environments (Kim *et al.*, 2013; Adriaenssens *et al.*, 2015; Hatfull, 2015; Zablocki *et al.*, 2016). This finding was not surprising, given the observations from previous studies (Ashelford *et al.*, 2003; Zablocki *et al.*, 2015) which showed high prokaryotic abundances in the Kogelberg soil environment. Nevertheless, signature sequences from large dsDNA eukaryotic virus families such as *Mimiviridae* (Raoult *et al.*, 2007) were represented in the Kogelberg Biosphere Reserve library despite the use of small pore size filters in sample preparation. Mimivirus signatures have been reported previously in other soil habitats (Zablocki *et al.*, 2014). Sequences that were found to be most similar to mimivirus ORFs were also obtained from Sargasso sea water samples, suggesting that these viruses, and their hosts, have a rather cosmopolitan distribution (Ghedin and Claverie, 2005).

4.2.4 Phylogeny of the Kogelberg Biosphere Reserve *fynbos* soil metavirome

Specific markers targeting virus families or species were used to analyse the taxonomic affiliations of the annotated ORFs and analyse the diversity within the group (reviewed in (Adriaenssens and Cowan, 2014). Phylogenetic trees were drawn from metavirome sequences on the basis of homology to marker gene reference sequences from the PFAM database. Sequences homologous to the marker genes (*polB*, *polB2*, *T7gp17* and *terL* (Appendices C Fig S2, S3, S4 and S5 online) and reference sequences were used to draw phylogenetic trees.

Using the DNA polymerase family B (polB) marker gene, conserved in all dsDNA viruses, Kogelberg Biosphere Reserve sequences appeared to be distantly related to *Rhodothermus* phage RM378 (order *Caudovirales,* family *Myoviridae).* This phage is the only sequenced representative of the "Far T4" group of myoviruses (i.e., distantly related to *Escherichia virus T4*) found in a previous diversity analysis of sequences from French lakes (Roux *et al.*, 2012) . The Kogelberg polB sequences from this study as well as the gp23 and gp20 marker gene sequences from the French lake study contribute to the expansion of the "Far T4"-like phages dataset.

A DNA polymerase family B *(polB2)* marker gene, which is conserved in members of *Adenoviridae, Salterprovirus*, and *Ampullaviridae* and *Podoviridae* family viral groups, was

analysed. The analysis showed a separate clade of sequences from the Kogelberg Biospheres reserve soil samples. Other *polB2* sequences from our dataset were found to be distantly related to members of the *Adenoviridae* family (isolated from a wide range of animal sources), the *Podoviridae* family (such as *Mycoplasma* phage *P1*, *Clostridium* phage *phi24R*, *Bacillus* phages *B103, phi39, Ga1*), the *Ampullaviridae* family (such as *Acidianus*-bottle-shaped virus) and the *Tectiviridae* family (such as *Bacillus* phages *G1L16C, Bam35C and AP50*).

Analysis of the metavirome sequence database using the marker gene *T7gp17* showed the presence of members of the *Podoviridae* family, subfamily *Autographivirinae* and genus *Phikmvvirus* and *T7virus*. Members of the genus *phikmvvirus* such as *Pseudomonas* phage LKA1, and unclassified phiKMV phages such as *Ralstonia* phage RSB1, were found to be closely related to the Kogelberg Biosphere Reserve sequences. Currently unclassified members of the genus *T7virus*, such as *Klebsiella* phage K11 and *Yersinia* phage φYeO3-12, were also found to be closely related to sequences in the Kogelberg Biosphere Reserve metavirome. The phages in the subfamily *Autographivirinae* are known to infect a wide range of environmentally important bacteria (Adriaenssens *et al.*, 2011).

Tailed phages of the order *Caudovirales* were the most commonly observed DNA viruses in the Kogelberg Biosphere Reserve sequences, consistent with other environmental samples (Adriaenssens *et al.*, 2015; Reavy *et al.*, 2015; S. Roux *et al.*, 2015). A phylogenetic tree built from a *Caudovirales*-specific terminase large subunit marker gene (*terL*) was used to visualise the diversity of the Kogelberg Biosphere Reserve *fynbos* soil *Caudovirales* (Fig 3). The Kogelberg Biosphere Reserve sequences clustered with all three families of tailed phages, indicating high phage richness in our sample set. These results were consistent with the taxonomic affiliations of contigs in the virus families shown in Table 3.

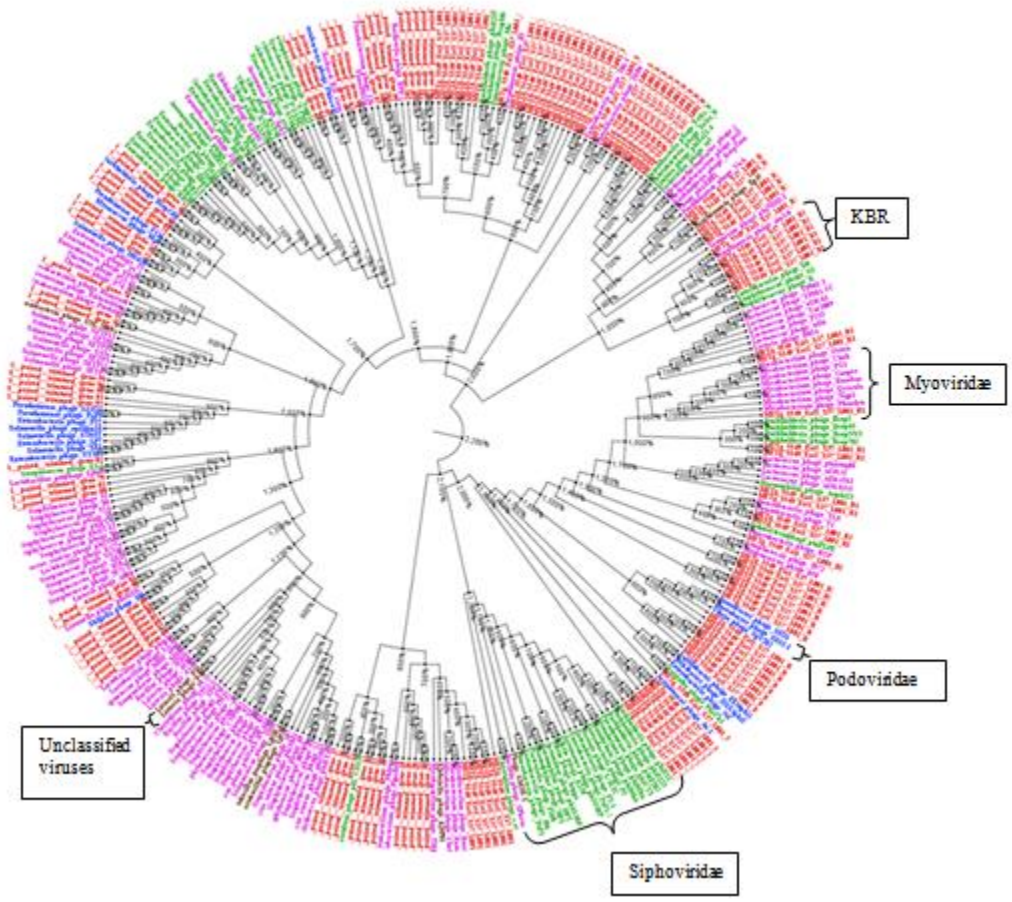


Figure 4.3: terL phylogenetic tree. Viral sequence origin of Caudovirales indicated with different colours on the contigs names. Kogelberg Biosphere Reserve *fynbos* soil - Red, Siphoviridae – green, Myoviridae – purple, Podoviridae - blue, unclassified viruses – grey

4.2.5 Analysis of a near-complete phage genome

MetaVir assemblies predicted 352 genes from the 6 contigs larger than 40kb, as well as 758 genes predicted from 19 contigs of between 20kb and 40kb. The 6 largest contigs were predicted to be linear, double stranded genomes. The sizes of the genomes were predicted to be 47kb long with 63 genes for the largest contig (Fig 4), followed by 44kb with 58 genes, 42kb with 61 genes, 42kb with 53genes, 40kb with 68 genes and 40kb with 49 genes. The genes in these contigs were predicted to show similarity to members of the order *Caudovirales*.

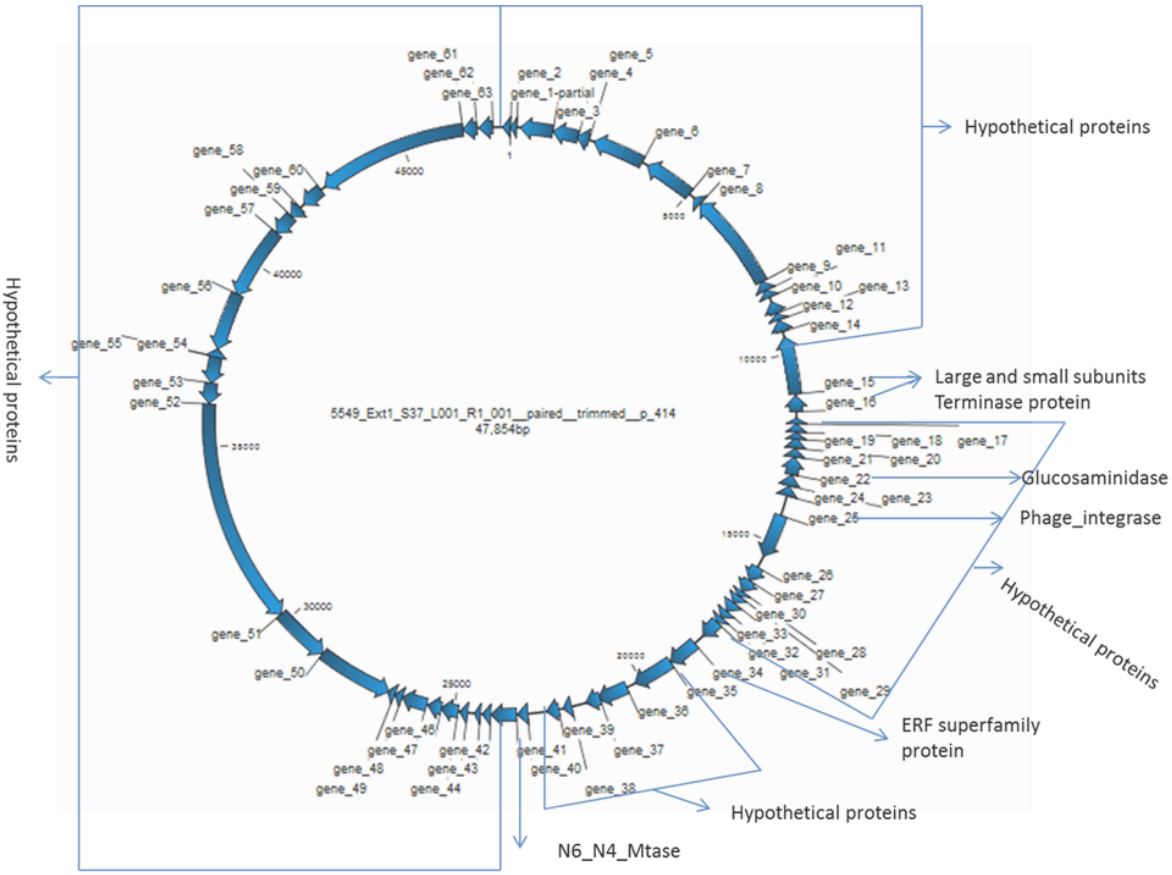


Figure 4.4: Gene annotation of contig 414. Arrowed blocks are open reading frames (ORFs), showing their orientation. Numbers within the contiguous genome are nucleotide positions, starting within gene number 1 and onwards in a clockwise orientation.

The largest contig represents a near-complete phage genome in the family *Podoviridae.* Members of this family typically contain double stranded and linear genomes of around 40 - 45kb in length with approximately 55 genes (Hulo *et al.*, 2011). Four of the genes in this assembled genome (genes 15, 16, 34 and 41) showed similarity to members of both *Podoviridae* and *Siphoviridae* families. The translated products of two of these genes (15 and 16) were identified as putative terminase large subunit (gene 15) and terminase small subunit (gene 16) genes, with 88% and 89% amino acid identity to *Puniceispirillum* phage HMO-2011 and *Pseudomonas* phage vB_PaeP_Tr60_Ab31, respectively. Both *Puniceispirillum* phage HMO-2011 and *Pseudomonas* phage vB_PaeP_Tr60_Ab31 belong to the family *Podoviridae*. The *terL* phylogenetic tree (Appendices C Fig S4 online) showed

a distant relatedness to members of the *Podoviridae* clade. Both terminase large and small subunits, together termed the terminase complex, are involved in the cleavage and packaging of concatemeric phage dsDNA (Penadés *et al.*, 2015). The large terminase subunit is involved in DNA cleavage and translocation into the procapsid while the small terminase subunit is involved in packaging initiation and stimulation of the ATPase activity of the large terminase. These DNA packaging mechanisms are used by most members of the *Caudovirales*.

The translated product of gene 34 was identified as a putative ERF superfamily protein and showed 55% amino acid identity to a homologue encoded by the unclassified *Clostridium* phage phiCP34O (order *Caudovirales,* family *Siphoviridae).* The ERF superfamily proteins are involved in the recombination of phage genomes (Wittmann *et al.*, 2011). The translated product of gene 41 was identified as a putative gp77 and showed 95% amino acid similarity to a homologue encoded by *Mycobacterium* phage Che9d (order *Caudovirales*, family *Siphoviridae*, genus *Che8likevirus).* gp77 proteins are known to function as shut-off genes during early stages of phage replication (Rybniker *et al.*, 2008).

Fifty-nine of the translated products of genes in the assembled phage genome showed identity to hypothetical proteins. Of these hypothetical proteins, 56 showed no sequence similarity to known virus families in BLASTp comparison to the RefseqVirus protein database. Three of the genes were predicted to encode glucosaminidase (a hydrolytic enzyme), Phage integrase (a site-specific recombinase that mediates controlled DNA integration and excision) and PDDEXK_1 (nuclease superfamily). Members of this PDDEXK_1 family belong to the PD-(D/E) XK nuclease superfamily. The PD-(D/E)XK nuclease superfamily contains type II restriction endonucleases and many other enzymes involved in DNA recombination and repair (Letunic *et al.*, 2004).

The protein sequences identified in this analysis indicated the presence of a putative ERF superfamily protein, Phage integrase and PDDEXK_1 family; all proteins implicated in DNA recombination. The ERF superfamily protein encoded by gene 34, whose sequences are expressed during recombination of temperate phages, catalyses annealing of single-stranded DNA chains and pairing of ssDNA with homologous dsDNA, which may function

in RecA-dependent and RecA-independent DNA recombination pathways (Dziewit *et al.*, 2014).

A few large contigs contained some predicted ORFs with similarities to phage sequences and coding for specific conserved phage proteins, including terminases, structural proteins (mainly related to Caudovirales tail structures) and phage DNA polymerases (Supplementary Table S2 online).

4.2.6 Cluster analysis

Contig datasets from nine metaviromes from various aquatic and soil habitats were selected for dinucleotide frequency comparisons (Willner *et al.*, 2009).

A comparison of the dinucleotide frequencies of the 9 metaviromes shows a clear bimodal clustering (Fig 5). Group 1, composed of soil-associated habitat and deep-sea sediment metaviromes, is further subdivided into soil, hypolith and sediments clades. Group 2 was restricted to freshwater habitats. The Arctic and Atlantic deep sea sediment and freshwater lake (Roux *et al.*, 2012) metaviromes clustered in single independent nodes. Such clustering reflects significant genetic similarity between these metaviromes, despite the geographical distances between sample locations.

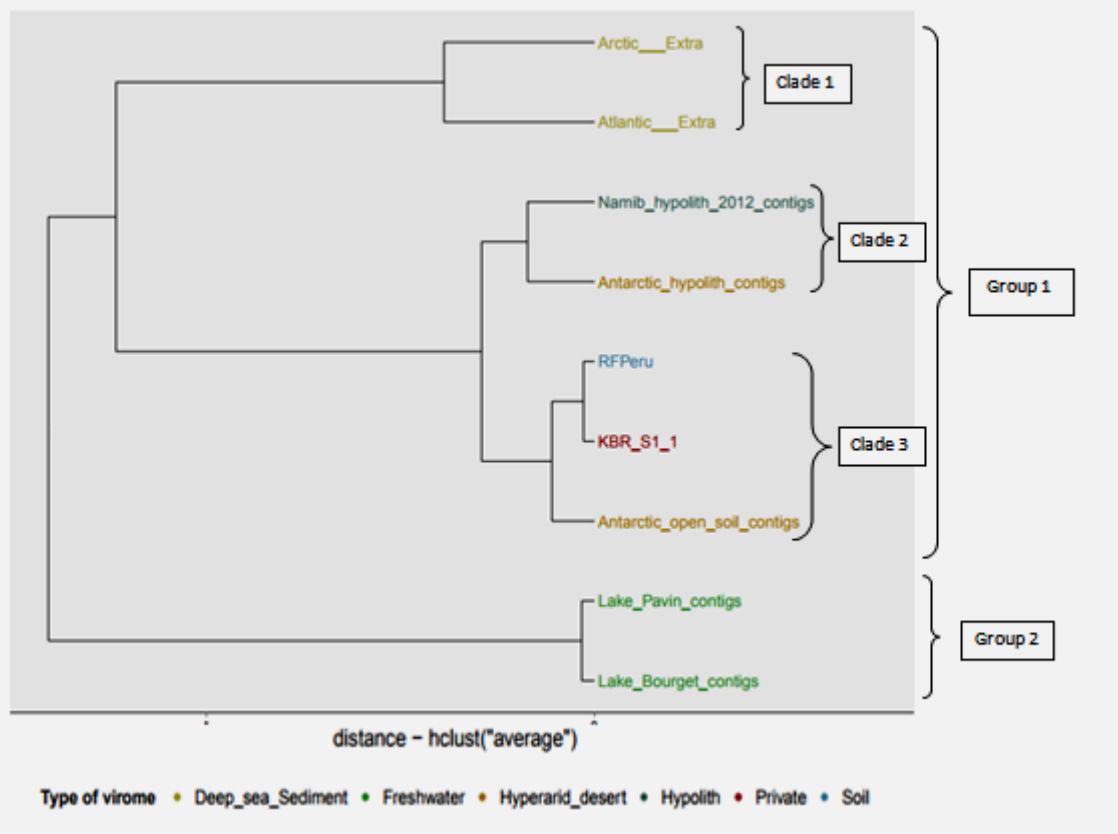


Figure 4.5: Hierarchical clustering of nine metaviromes (assembled into contigs) based on dinucleotide frequencies. The types of biome are differentiated by colour with Kogelberg Biosphere Reserve – red, freshwater – dark green, hyper-arid desert – light blue, hype hypersaline – yellow, hypolith – dark blue, seawater – light green and unknown biomes – gold. The x-axis denotes eigenvalues distances. The tree was constructed using MetaVir server pipeline according to the method in (Willner et al., 2009).More details on sample names are described in supplementary Table S3 online.

Both hypolithic metaviromes (i.e., cold Antarctic and hot Namib Desert hypolithic biomass samples) clustered as a single node, despite their widely differing habitat-associated environmental characteristics (dominated by an est. 50^{o}C mean annual temperature difference) and substantial spatial separation (approx. 55 degrees of latitude), suggesting that aridity and not temperature may be the dominant driver of host and viral diversity (Zablocki O., van Zyl L., Adriaenssens EM., Rubagotti E., Tuffin M., Cary SC., 2014)(Zablocki *et al.*, 2014). Interestingly, soil related metaviromes (from Kogelberg Biosphere Reserve *fynbos*

soil, Peruvian rainforest soil and Antarctic Dry Valley desert soil) clustered together and were clearly distinct from soils which were geographically much closer.

The Kogelberg Biosphere Reserve soil metavirome clustered at a single sub-node with the Peruvian rainforest soil metavirome. Both of these habitats experience high annual rainfall and warm temperatures and are characterised by heavily leached and low nutrient status soils, suggesting that soil composition and/or nutrient status may be the strong driver of the host and viral diversity (Leigh, 1975; Leger), 1987). These observations suggest a niche-dependent pattern, where spatially distinct niche environments cluster together and separate from their geographically closer soil counterparts (Zablocki O., van Zyl L., Adriaenssens EM., Rubagotti E., Tuffin M., Cary SC., 2014).

Previous study reported that cluster analysis of hypolith and open soil metaviromes from Antarctic and Namib Desert soil has shown that both hypolith metaviromes clustered at a single node and also that both open soil metaviromes displayed an identical pattern (Zablocki O., van Zyl L., Adriaenssens EM., Rubagotti E., Tuffin M., Cary SC., 2014). Similarly, to our study, related habitat types harboured more closely related viral communities, despite the great geographic distances or differing environmental conditions. The common factor in these hyperarid environments is water scarcity, which may be a key driver of community speciation and recruitment in these environments. We conclude that these adaptations and the nature of soil habitat compared to the 'refuge' habitat of quartz stones for hypolithic communities, may be the driving force between both communities not to cluster together.

4.2.7 Functional properties of the Kogelberg Biosphere Reserve *fynbos* soil metavirome

The functional implication of the reads was explored using MG-RAST. The Kogelberg Biosphere Reserve metavirome sequences exhibited a high proportion of uncharacterized ORFs, with 2,362,076 sequences showing no significant similarities to proteins in the databases (ORFans). Twelve functional categories were annotated by MG-RAST, each subdivided into distinct subsystems (Fig 6). The database searches against SEED in the MG-RAST subsystem resulted in 9360 hits. The highest percentage hits (20.3%) in the functional annotation belonged to the "Phage, prophages, transposable elements and plasmids"

subsystem category, with r1t-like streptococcal phages, phage packaging machinery and phage replication annotations most commonly identified.

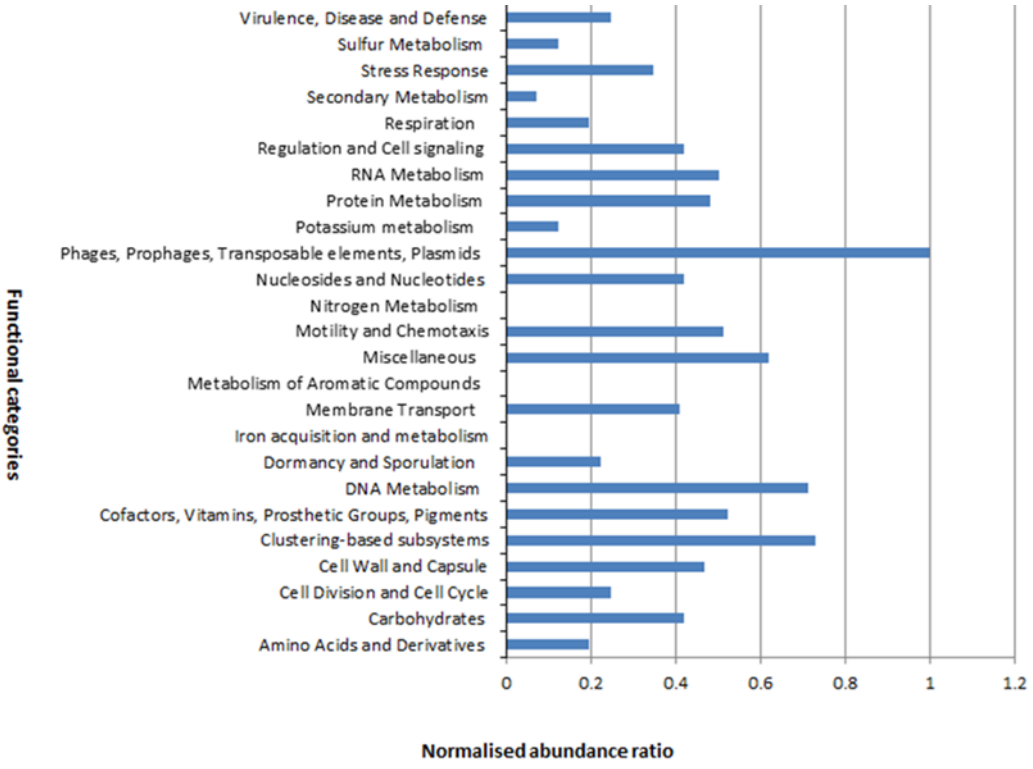


Figure 4.6: Functional assignment of predicted ORFs. Functional annotation was performed at 60% similarity cut-off as predicted by MG-RAST.

The other functional subsystem categories showed "Clustering-based subsystems (e.g., biosynthesis of galactoglycans and related lipopolysaccharides; catabolism of an unclassified compound etc., and other clusters identified as unclassified). The "Protein metabolism" and "DNA metabolism" functional categories were also dominant annotations. Many proteins in these functional categories, such as terminases, HNH homing endonucleases, DNA helicases, DNA polymerases and DNA primases, could potentially be of phage origin. These functional groups have also been found to be highly represented in previous metaviromic datasets (Roux *et al.*, 2013; Adriaenssens *et al.*, 2015; Cai *et al.*, 2016).

Analysis of the metavirome reads using the KEGG Orthology (KO) database showed metabolism protein families (carbohydrate metabolism, amino acid metabolism and nucleotide metabolism) to be the most commonly identified. Members of the genetic information procession protein family, including replication and repair, transcription and translation proteins, were also commonly identified. Deeper analysis of a subset of annotated contigs identified genes encoding numerous virus structures (e.g., phage capsid, terminase, tail fibre protein etc.) and DNA manipulating enzymes (e.g., endonuclease, DNA methylase, primase-polymerase, DNA primase/helicase, DNA polymerase I, integrase, ssDNA annealing protein, exonuclease, transferase, site-specific DNA methylase, ligase, recombinase etc.).

From this analysis, we demonstrate that phage-related genes and metabolic genes are highly represented. The virome displayed a strong enrichment in phage-like genes (e.g. phages, prophages, transposable elements, plasmids) and lacked typical cellular categories rarely observed in sequenced phages (e.g. 'cofactors, vitamins, prosthetic groups, pigments'). Cellular categories commonly identified in known phages were retrieved (e.g. 'nucleosides and nucleotides', 'DNA metabolism'). The highly abundance of virome-associated metabolic genes shows that the phages may have the potential to interfere with the metabolism of their hosts. Our virome analysis, consistent with other virome studies, demonstrate the unexpected picture of global 'viral' metabolism, suggesting that viruses might actively dictate the metabolism of infected cells on a global scale (Roux *et al.*, 2013).

The functional assignments from the SEED database of Kogelberg Biosphere Reserve *fynbos* soil was clustered with SEED database functional assignments of the 12 previously published metaviromes from both similar and dissimilar environments (fresh water (Roux *et al.*, 2012), soil and hypolithic niche communities (Zablocki *et al.*, 2014; Adriaenssens *et al.*, 2015), pond water (Rodriguez-Brito *et al.*, 2010) and sea water (Angly *et al.*, 2006) mentioned in Fig 2. A cluster analysis of the SEED database subsystem classification revealed different functional patterns between the metaviromes and no clear soil clustering (Fig 7). The sequences from Kogelberg Biosphere Reserve clustered amongst the sequences from three of the fresh water lakes and the Namib hypolith metaviromes. Antarctic samples (Antarctic open soil and Antarctic hypolith) were more distinct and formed a heterogeneous

clade with the other fresh water samples. This can potentially be explained by the larger number of cellular contamination in some of these metaviromes. This finding suggests that different biomes can share similar functional patterns and, conversely, that taxonomically similar viromes can encode different functional genes. It may also indicate that certain phage groups are more prevalent in certain biogeographic regions.

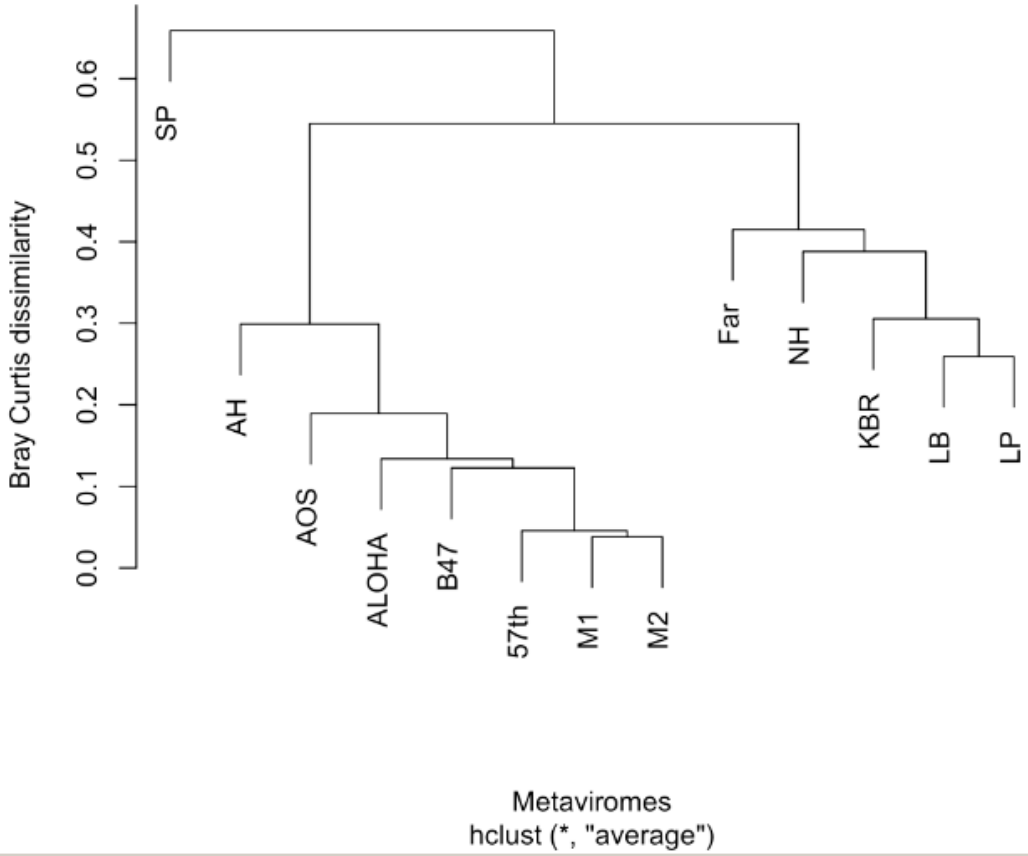


Figure 4.7: Cluster analysis of functional assignment of predicted ORFs. Viromes were clustered with the hclust algorithm in R according to the abundance of SEED database functional categories present. SEED categories were assigned using Megan6 after blastp-based comparison with the non-redundant protein database of NCBI. More details on the description of metaviromes are described in Supplementary Table 2 online.

This study is not without limitations. The major limitation to this study is the use of only a single virome that includes only double stranded DNA viruses.

4.3 Conclusion

We have successfully used the metaviromics approach to explore the diversity and functional composition of a previously unexplored Kogelberg Biosphere Reserve *fynbos* soil virome. Our quantitative comparison of taxonomic and functional composition of the Kogelberg soil metavirome with other published viromes is a valuable and novel contribution that will enhance the repertoire of publicly available datasets and advance our understanding of viral ecology. Furthermore, contigs corresponding to novel virus genomes were assembled in the current work; this presents an opportunity for future studies aimed at targeting these novel genetic resources for applied biotechnology.

CHAPTER 5: SCREENING, EXPRESSION, PURIFICATION AND ACTIVITY ASSAY OF NUCLEIC ACID MANIPULATING ENZYMES

5.1 Introduction

Metaviromics offer tremendous potential for the bioprospecting of novel nucleic acid manipulating enzymes in any given natural environment (Schoenfeld *et al.*, 2010, 2011; Choi, 2012; Moser *et al.*, 2012). The global molecular biology enzymes market was valued at $4 928.3 million in 2015 and is predicted to grow by 17.2% (CAGR) to 2022 (The global molecular biology enzymes and kits & reagents market, 2017- 2022). Applications of these enzymes in PCR, sequencing, cloning, restriction digestion and synthetic biology are already well established. The major factors contributing to the surge in demand of new nucleic acid manipulating enzymes includes the increase in research activities by end users, the increase of investments by biotechnology companies in molecular biology research, and the rising awareness and prevalence of genetic disorders and technology advancements that facilitate new applications (The global molecular biology enzymes and kits & reagents market, 2017 -2022). Examples of some of these enzymes are T4 DNA ligase (Murray *et al.*, 1979), reverse transcriptase (Moser *et al.*, 2012), recombinases (Lopes *et al.*, 2010), thermostable T4 RNA ligase (Blondal *et al.*, 2003), Reverse transcriptase (DNA directed RNA polymerase), thermostable DNA polymerases such as *Thermo aquaticus* (Taq) DNA polymerase and *Pyrococcus furiosus* (Pfu) DNA polymerase (Moser *et al.*, 2012, Mathur, 1996 Pluthero, 1993), *cas9* gene-editing enzyme (Sander and Joung, 2014), phage DNA methyltransferase (Dziewit *et al.*, 2014) and T7 DNA polymerase (Tabor and Richardson, 1985).

High-throughput sequence-based and function-based screening approaches have recently been used to access genetic information of these enzymes contained in a variety of microbial and viral communities from the environment (Ferrer *et al.*, 2009, 2015; Leemhuis *et al.*, 2009; Simon *et al.*, 2009; Uchiyama and Miyazaki, 2009; Moser *et al.*, 2012; DeCastro *et al.*, 2016; Martínez-Martínez *et al.*, 2016; Madhavan and Sindhu, 2017; Tripathi *et al.*, 2018). Unknown environmental genetic resources can be accessed by directly sequencing the metavirome DNA extracted from the environment or the construction of a metavirome library and the sequencing of all the metavirome recombinant clones. These sequencing approaches are limited by the reliance on similarities to known gene sequences in public databases of nucleic acid sequences

(Fantle *et al.*, 2003). Alternatively, the genes function can be analysed through functional screening the metavirome clones for novel activities and phenotypes produced by the bacterial host (Uchiyama and Miyazaki, 2009). However, previous studies showed that many genes from metavirome libraries were unable to be expressed in the selected host bacterium (Andraos *et al.*, 2004; Goff, 2004), and there is a need to improve the ability of bacterial hosts to harbour and express viral recombinant DNA.

In order to demonstrate functional utility and techno-economic viability of the enzymes discovered, one of the critical factors is to develop a robust protein expression system. Over the years a number of protein expression systems have been developed, including eukaryotic (e.g. mammalian and insect) (Trill *et al.*, 2001), yeast (e.g. *Saccharomyces*, *Pichia*, *Kluyveromyces*, *Hansenula* and *Yarrowia*) (Grisshammer and Tate, 1995) and prokaryotes (e.g. *Bacillus* and *E. coli*) (Terpe, 2006). Of these systems, the *E. coli* expression system remains one of the most studied and widely used expression systems in the production of recombinant proteins (Terpe, 2006). Some of the major advantages of *E. coli* as an expression host include extensive knowledge of the bacteria's genetics, the availability of versatile vector systems and host strains, easy transformation and the relatively low associated costs (Georgiou and Segatori, 2005). However, like any other expression system, *E. coli* suffers some serious drawbacks; including occasional low level protein expression as a result of codon bias and the low availability of specific tRNAs in *E. coli*, poor cell viability due to gene product expression and toxicity, formation of mRNA secondary structure or RNA instability that prevents proper expression, and low solubility of expressed proteins as result of the formation of inclusion bodies (Fahnert *et al.*, 2004; Vallejo and Rinas, 2004; Singh and Panda, 2005). Over the years different approaches have been developed to enhance recombinant protein production and rapid purification of recombinant proteins in *E. coli* (Bashiri *et al.*, 2014). Such approaches include *de novo* synthesis of the codon-optimised gene fragments (Gustafsson *et al.*, 2004), hydrophilic large fusion partners (e.g. maltose-binding protein, MBP (Kapust and Waugh, 1999) and glutathione-*S*-transferase, GST (Sohoni *et al.*, 2015)) and the fusion to PelB and OmpA leader peptides (Georgiou and Segatori, 2005)). Additional approaches to improve protein expression through agents that facilitate the correct folding of intracellular recombinant proteins include the use of various growth media (e.g. adding sorbitol, sucrose, raffinose), lowering of expression temperature, co-expression of chaperones and changing genotype of

the *E. coli* host. Notwithstanding these drawbacks outlined above, the *E. coli* expression system was chosen in this study as an expression system for recombinant nucleic acid manipulation enzymes identified in metavirome samples.

One of the major disadvantages of high-throughput genome sequencing programmes from a bioprospecting point of view is that many genes discovered remain functionally uncharacterised with many of these assigned as putative proteins, making it difficult to exploit the huge genetic resource for industrial application. The objective of this chapter is thus:

- To demonstrate the application of both the sequence and function-based screening approaches in the discovery of novel nucleic acid manipulating enzymes.
- To express, purify and characterise the identified nucleic acid manipulating proteins.

5.2 Results and discussion

5.2.1 Sequence-Based Screening of a Kogelberg Biosphere Reserve soil Metavirome library

In the previous chapter (Chapter 4), a combination of MetaVir (http://metavir-meb.univ-bpclermont.fr), VIROME (http://virome.dbi.udel.edu/) and MG-RAST (http://metagenomics.anl.gov/) metavirome sequence processing platforms was used to estimate viral diversity and taxonomical classification from the direct sequencing of the metavirome from the Kogelberg Biosphere Reserve soil samples (Roux *et al.*, 2014) (Wommack *et al.*, 2012) (Keegan *et al.*, 2016). In this chapter, the *de novo* assembled CLC sequence data were used as input data for the MetaVir pipeline coupled with the BLASTP server to screen for genes encoding novel nucleic acid manipulating enzymes.

The MetaVir sequence output revealed a total of 591 contigs (≥ 1kb in size). In order to avoid spurious sequencing products with low coverage, contigs with length sizes of less than 1kb were disregarded from the screening of putative ORFs encoding nucleic acid manipulating enzymes. About 13004 contigs were lost from using the 1kb as a cut-off for ORF analysis, which reduced the sequence information by 84.87%. Based on the length, the 591 contigs were distributed as follows: 58.2% were 1-5kb length contigs; 27.5% were 10-20kb length contigs; 6.1% were 20-30kb length contigs and 3.7% were 30-40kb length contigs (Figure 5.1).

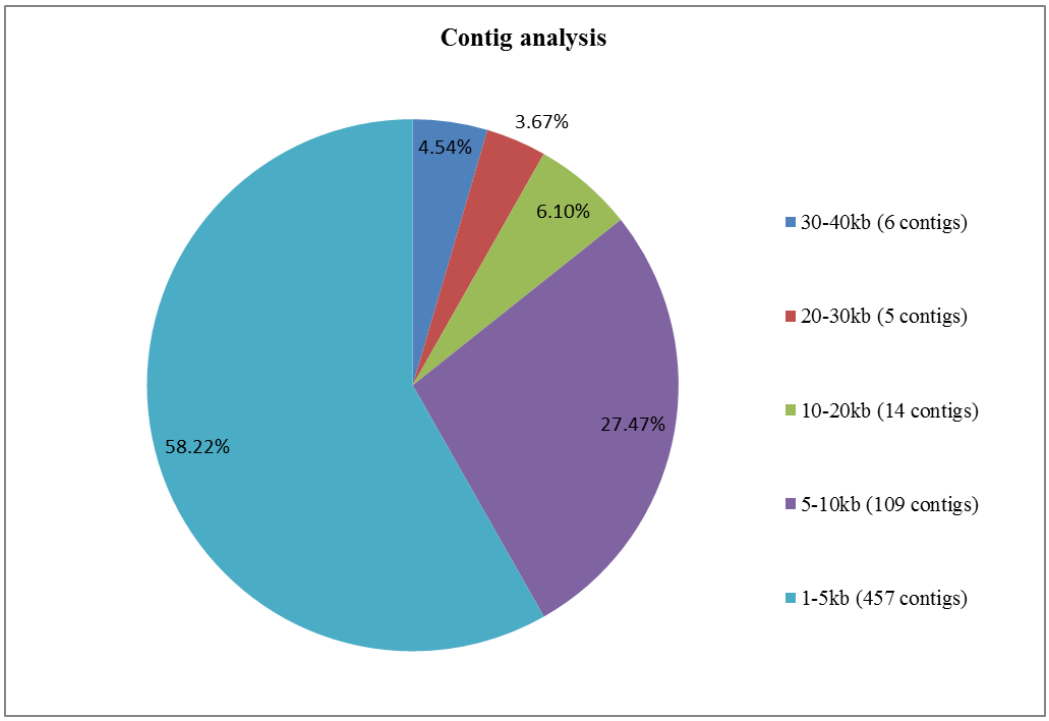


Figure 5.1: **Distribution pattern of the contigs.**

5.2.1.1 Sequence analysis of putative nucleic manipulating enzymes encoding ORFs

A combination of the MetaVir pipeline platform (Roux *et al.*, 2014) assessed by BLASTP (Altschul *et al.*, 1990) comparison resulted in a total of 907 putative ORFs encoding nucleic acid manipulating enzyme from 591 contigs identified. The ORFs belonged to different classes of nucleic acid manipulating enzymes (i.e. 316 polymerases, 323 nucleases, 55 ligases, 182 methylases, 29 phosphatases and 2 topoisomerases). Nine contigs were selected based on the length of the contig (≥ 1kb). Fifteen complete ORFs encoding nucleic acid manipulating enzymes were identified and were distributed across 9 identified contigs (Contigs 1-9) (Table 5.1). Physical maps of the 9 contigs showing the location and directional organisation of the 15 putative ORFs are shown in Figure D1 (A, B, C, D, E, F, G, H, I) in the Appendices section D.

The coding nucleotide sequence lengths and GC composition ratios of the 15 identified putative nucleic acid manipulating enzyme encoding ORFs spanned from 237 to 2217bp and 40 to 66% respectively (Table 5.1). The translated coding sequences encoded polypeptides of between 79 to 793 amino acid range and with the corresponding predicted subunit molecular masses of 8 and 83 kDa range. A global amino acid alignment of the 15 translated amino acid sequences

against the GenBank (BLASTp) revealed highest sequence identities of between 30 - 97 % to a number of nucleic acid manipulating enzyme of bacteriophage origin (Table 5.1). A total of 4 enzyme classes were identified including five polymerases (ORF 3, 6, 12, 14 and 15), three ligases (ORF 10, 11 and 13), six nucleases (ORF 2, 4, 5, 7, 8 and 9) and one methylases (ORF 1) (Table 5.1). Signal peptide prediction analysis using SignalP4.0 (Petersen *et al.*, 2011) revealed that none of the ORFs encoded leader (secretion signal) peptide (Table 5.1), suggesting that the encoded putative nucleic acid manipulating enzymes were intracellular proteins potentially secreted in the cytoplasm (Terpe, 2006). Using 60% or less sequence identity cut-off as a novelty threshold and one ORF per contig criteria, the following nine ORFs (ORF 3, 4, 7, 9, 10, 12, 13, 14 and 15) were selected for recombinant production studies.

Although the sequence-based approach resulted in the identification of 15 ORFs encoding putative nucleic acid manipulation enzymes, a number of genes are likely to remain inaccessible through this approach. This argument is strengthened by the observation that majority of the ORFs identified during this study remain unannotated (contig maps Figure D1 (A, B, C, D, E, F, G, H, I) Appendices section C). These unannotated ORFs present potentially novel gene sequences that remain functionally uncharacterised. In this regard, Culligan *et al.*, (2014) advocated for a combination of both the gene sequence and functional targeting approach in order to circumvent the problem of unannotated gene sequences and to maximize the value of the NGS methods as a tool for novel enzyme gene discovery (Culligan *et al.*, 2014). Furthermore, Teeling and Glockner, (2012) are of the view that the absence of a comprehensive tool for metagenome data analysis that incorporates all types of analysis (biodiversity analysis, taxonomic classification and binning, functional annotation and metabolic reconstruction) result in a number of genes that lacks dedicated known protein functions and are hypothetical. Consequently, the same authors advocate for an integrated approach that should adopt a combination of sequence interpretation tools and/or different screening approaches (Teeling and Glockner, 2012). Having recognised the shortcoming associated with the sequence-based method as an approach to novel gene discovery, a decision was taken in the current study to also explore a function-based screening approach to search for novel DNA manipulating enzymes.

Table 5.1: **Summary of BLASTp search using the MetaVir analysis for the selected nucleic acid manipulating contigs and ORFs**

Contigs	Length (bp)	Number of ORFs per Contig	Putative nucleic acid manipulating enzyme encoding ORFs	ORF number within the contig	ORF re-name	Nucleotide position within the contig	ORF length (bp) (aa)[a]	Predicted pI (M_w in KDa)[b]	GC contents (%)	Identity (%)	Signal peptide [c]	Predicted function (PFAM and BLASTp) [d]
Contig 1	23982	43	3	ORF 23	ORF1	14310-14892	582/194	8.97 (22)	48	41	No	putative DNA methylase (*Mycobacterium* phage *Papyrus)*
				ORF 28	ORF 2	16194-16557	363/121	9.41 (13)	52	37	No	putative endonuclease (*Mycobacterium* phage Papyrus) (PF05866.6 RusA)
				ORF 40	ORF 3	21274-21907	633/211	5.99 / 23	45	32	No	putative DNA polymerase III subunit beta (*Rhizobium* phage) (PF02767.11 DNA_pol3_beta_2)
Contig 2	23401	31	3	ORF 1	ORF 4	13-1786	1773/591	6.77 (65)	42	51	No	putative DNA helicase (*Pseudomonas* phage) (PF04851.10 ResIII)
				ORF 2	ORF 5	1789-2704	915/305	8.58 (34)	42	55	No	putative exonuclease *(Pseudomonas* phage NP1)
				ORF 27	ORF 6	21102-21573	471/157	6.84 (18)	40	30	No	putative DNA-directed RNA polymerase specialised

Contigs	Length (bp)	Number of ORFs per Contig	Putative nucleic acid manipulating enzyme encoding ORFs	ORF number within the contig	ORF re-name	Nucleotide position within the contig	ORF length (bp) (aa)[a]	Predicted pI (M_w in KDa)[b]	GC contents (%)	Identity (%)	Signal peptide [c]	Predicted function (PFAM and BLASTp) [d]
												sigmasubunit (*Sinorhizobium* phage phiM9) (PF04542.9 Sigma70_r20)
Contig 3	20911	32	2	ORF 3	ORF 7	2394-2919	526/175	8.97 (19)	66	44	No	putative HNH endonuclease (*Gordonia* phage Smoothie) (PF13392.1 HNH_3)
				ORF 6	ORF 8	3969-4206	237/79	9.30 (8)	64	97	No	putative predicted homing endonuclease (*Autographivirinae Citrobacter* phage CR44b)
Contig 4	19506	27	1	ORF 12	ORF 9	10668-11208	541/180	9.74 (19)	57	47	No	putative endonuclease VII (*Cronobacter* phage) (PF02945.10 Endonuclease_7)
Contig 5	15488	28	1	ORF 24	ORF 10	12914-13622	7089/236	6.54(26)	62	40	No	RNA ligase (PF09414.5 RNA_ligase) (Unknown)
Contig 6	11510	18	2	ORF 9	ORF 11	4844-6020	1176/392	6.39 (44)	57	38	No	putative polynucleotide kinase/ligase (*iridescent* virus) (PF09511.5 RNA_lig_T4_1)

Contigs	Length (bp)	Number of ORFs per Contig	Putative nucleic acid manipulating enzyme encoding ORFs	ORF number within the contig	ORF re-name	Nucleotide position within the contig	ORF length (bp) (aa)[a]	Predicted pI (M_w in KDa)[b]	GC contents (%)	Identity (%)	Signal peptide [c]	Predicted function (PFAM and BLASTp) [d]
				ORF 11	ORF 12	6394-8395	2001/667	5.89 (73)	63	38	No	putative DNA polymerase I (*Thermus* phage P2345) (PF00476.15 DNA_pol_A)
Contig 7	10638	22	1	ORF 2	ORF 13	181-1213	1032/344	6.39 (37)	61	47	No	putative DNA ligase *(Yersinia* phage) (PF01068.16 DNA_ligase_A_M)
Contig 8	5631	5	1	ORF 1	ORF 14	163-2380	2217/739	8.99 (83)	40	44	No	putative DNA polymerase *(Thermus* phage) (PF00476.15 DNA_pol_A)
Contig 9	1447	1	1	ORF 1	ORF 15	1-1447	14446/482	8.75 (54)	51	37	No	putative DNA polymerase I (*Vibrio* phage VpV262) (PF00476.15 DNA_pol_A)

[a] (aa= amino acid; bp =base-pair)

[b] The isoelectric point (pI) and Molecular weight (M_w) were predicted using Expasy server (https://web.expasy.org/compute_pi/)

[c] Signal peptide coding sequences were predicted using SignalP 4.1 server (http://www.cbs.dtu.dk/services/SignalP/).

[d] Best hits were determined using BLASTP server (http://blast.ncbi.nlm.nih.gov/Blast)

5.2.2 Metavirome library construction and functional screening for DNA polymerase 1 enzyme

The second approach applied in this study was to screen metavirome library for nucleic acid manipulating enzyme using a function-based approach. As part of this strategy, metaviromic DNA was extracted and a metavirome library was then created and screened using a DNA polymerase complementation assay. Thus, the function-based screening assay described in this section targets only DNA polymerase, as opposed to other nucleic acid manipulating enzyme classes.

5.2.2.1 Metavirome DNA isolation

A total of 6.5ng metavirome DNA was isolated from Kogelberg Biosphere Reserve *fynbos* soil sample. While the extracted metavirome DNA was of low amounts, it was nonetheless of high quality, as shown by the $A_{260/230nm}$ and $A_{260/280nm}$ absorbance ratios of 2.17 and 1.84, respectively (Gillespie *et al.*, 2005; Osborn, 2005)

The isolated DNA was screened for prokaryote and eukaryote DNA contamination using PCR based on the 27F and 1492R primer pair targeting the 16S rRNA gene marker (Lu *et al.*, 2007) and the NS1 and NS8 primer targeting 18S rRNA marker (Innis *et al.*, 1990) . There were no PCR products corresponding to the expected 1300pb (16S) and 1600bp (18 S) rRNA gene markers, confirming the absence of prokaryotic and eukaryotic contaminating DNA in the isolated DNA sample (data not shown). Due to a low concentration of the extracted metavirome DNA, the MDA technique was used to generate enough DNA for library construction. The MDA resulted in an increased amount of DNA from 6.5ng to 2.7µg (Figure 5.2). The proof reading capacity of the *Phi* 29 DNA polymerase coupled with the random (hexamer) primers lead to the generation of high molecular DNA fragments as observed in Figure 5.2b (Shoaib *et al.*, 2008).

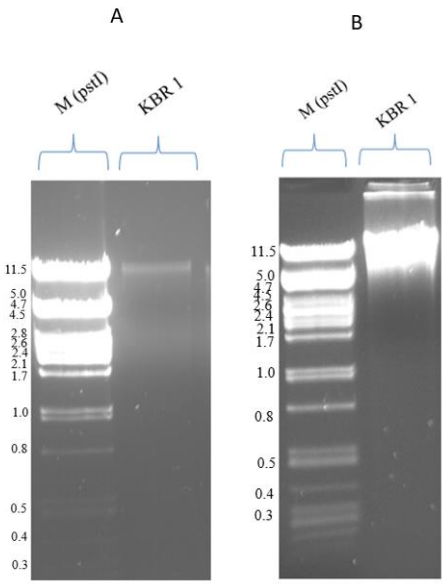


Figure 5.2: **Agarose gel electrophoresis analysis of metavirome DNA (A) Extracted metavirome DNA directly from soil. Lane 1 (M): λ *Pst*I DNA marker. Lane 2 (KBR 1): Metavirome DNA extracted directly from KBR sample. (B) Lane 1 (M): λ *Pst*I DNA marker, Lane 2 (KBR 1): MDA amplified metavirome DNA.**

5.2.2.2 Library construction

Construction of a fosmid library was carried out using a CopyControl pCC2FOS™ vector which resulted in a library size of approximately 5.7×10^6 colony forming units (cfu/mL). The restriction fragment length polymorphisms of 20 randomly selected clones using *BamH*I and *Hind*III restriction enzymes showed non-redundant patterns and average insert sizes of between 35 - 40kb (Figure 5.3). The use of fosmid vectors for metaviromics library preparation for the investigation of novel enzymes was previously suggested by (Béjà, 2004; Schoenfeld *et al.*, 2010) and used successfully to functionally screen for enzymes (Kennedy *et al.*, 2008; Terrón-González *et al.*, 2013; Gonçalves *et al.*, 2015), to investigate the diversity of viruses (Santos *et al.*, 2010; Adriaenssens *et al.*, 2016; Mizuno *et al.*, 2016) and to overcome the difficulties encountered when assembling short viral sequences of environmental origin (Santos *et al.*, 2010).

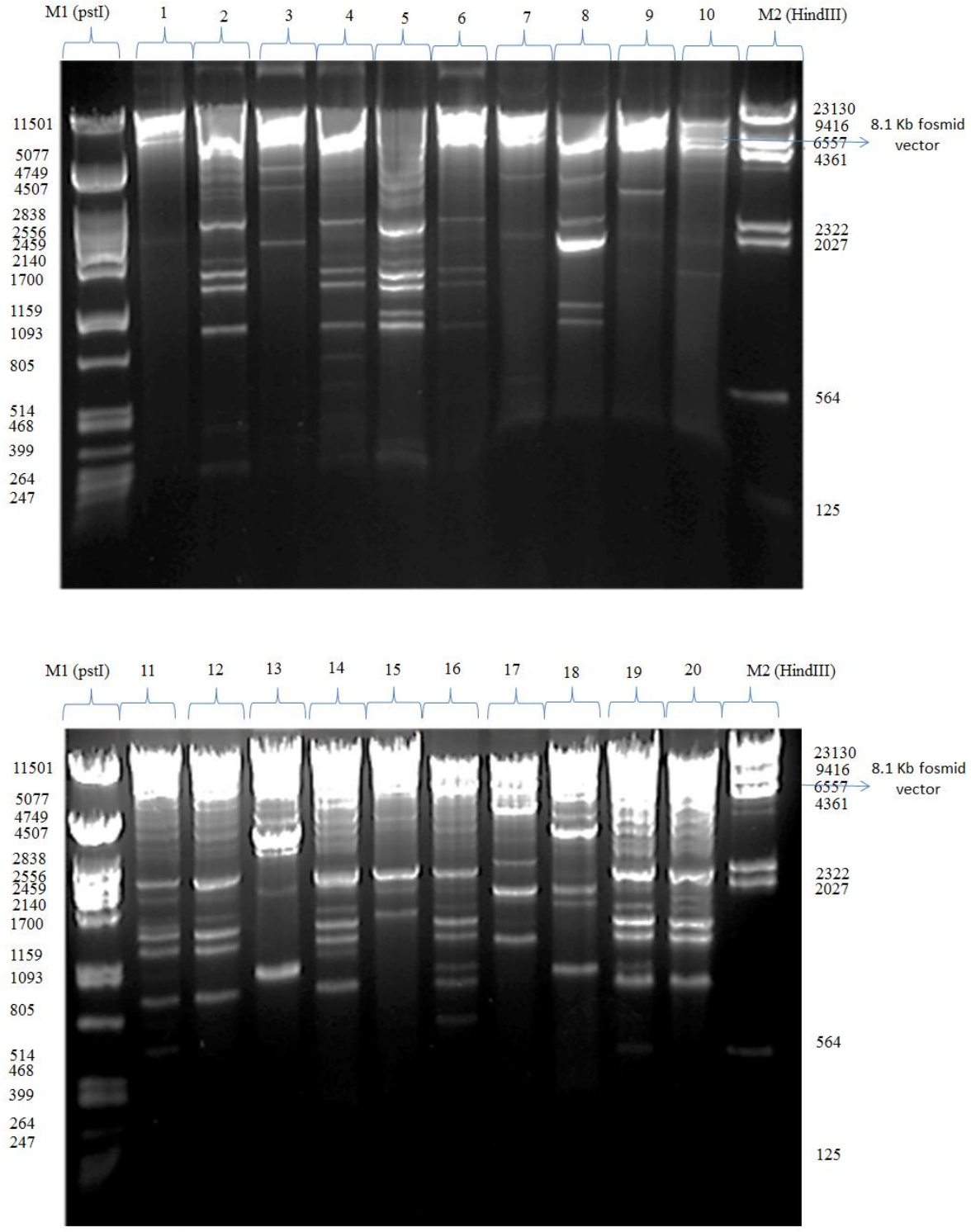


Figure 5.3: **Agarose gel electrophoresis analysis of *BamH*I and *Hind*III restricted randomly selected fosmid clones. M1 and M2: *Pst*I and *Hind*III DNA markers; lane1-20 represents fosmid DNA restriction (from 20 randomly selected clones) with *BamH*I and *Hind*III restriction endonucleases.**

5.2.2.3 Function screening of DNA polymerase 1 from Kogelberg Biosphere Reserve *fynbos* soil sample

Functional screening of the metavirome library was carried out using DNA polymerase complementation assay. The principle of the assay involves the use of the cold-sensitive *E. coli* mutant CSH26 fcsA29 [F$^-$ara (lac-pro)thi fcsA29 met::Tn5] (Nagano *et al.*, 1999) strain as a host. The temperature sensitive lethal mutation found in the 5'-3' exonuclease domain of DNA polymerase I enzyme of the *E. coli* mutant leads to filamentation of the cells with dispersed nuclei, and consequently the suppression of growth at lower temperatures, i.e. below 20°C (Nagano *et al.*, 1999). The expectation is that only recombinant *E. coli* strains harbouring a fosmid vector with a gene insert conferring polymerase activity, particularly with the 5'-3' exonuclease, could grow at the lower temperature conditions employed; and therefore complement the DNA polymerase activity of the cold-sensitive *E. coli* mutant CSH26 fcsA29.

Approximately 200 positive colonies of >3mm diameter in size were observed after incubation of the transformants (*E. coli* CSH26 fcsA29 harbouring the pCC2FOS^{TM} vector and inserts) for 72 h at 18°C. The *E. coli* CSH26 fcsA29 harbouring the pCC2FOS^{TM} vector was included as a negative control and did not grow after 72 h at 18°C. The fosmid DNA of the clones that consistently showed growth phenotypes at 18°C were then purified and were used to retransform the cold-sensitive *E. coli* mutant CSH26 fcsA29 [F-ara (lac-pro) thi fcsA29 met:Tn5] to confirm initial observations. The second round of re-screening only led to 20 positive-phenotype clones, indicating that many of the clones picked in the initial screen were false positives. The fosmid DNA of the 20 positive clones from the second round of re-screening was pooled and sequenced by NGS using Illumina platform. The pooling of multiple positive fosmids before NGS was done to increase throughput and to reduce sequencing costs. Similar approaches have been employed elsewhere by Wylie *et al.*, (2015), where samples were pooled together before sequencing to reduce sequence costs. Vester *et al.* (2014) also adopted the High-throughput Illumina sequencing approach on pooled clones in order to confirm the identity of genes initially identified through a function-based expression library.

5.2.2.4 Sequence analysis of positive fosmid clones

High-throughput Illumina sequencing on a pool of all 20 fosmid clones generated 1GB of sequence data, which was assembled using *de novo* assembly parameters of the CLC genomics workbench version 6.0.1 software. The final assembly consisted of 597 contigs with a total

metavirome size of 4657197bp and an average length of 7801bp. The predicted full-length and truncated coding sequences in the resulting metavirome contigs annotated by RAST (Aziz *et al.*, 2008), VIROME (Wommack *et al.*, 2012) and BLASTp (NCBI) (Altschul *et al.*, 1990) databases. Furthermore, 3049 sequences containing proteins with known functions were detected. There was no ribosomal RNA genes were predicted by the databases.

Five contigs encoding putative 5'-3' exonuclease domain from DNA polymerase 1 enzyme were identified. The other contigs possibly encoded housekeeping genes that are responsible for the basic function of the cell. Across all 5 contigs, 26 ORFs were identified, where only 5 were encoding putative 5'-3' exonuclease domains (i.e. one ORF per contig) (Table 5.2). The coding nucleotide sequence lengths and GC composition ratios of the 5 identified ORFs encoding putative 5'-3' exonuclease domain ranged from 198 to 2787bp and 48 to 56% respectively (Table 5.2). The translated coding sequences encoded polypeptides ranging from 929 to 66 amino acids (partially translated). The predicted subunit molecular masses of the translated polypeptides ranged between 15 and 78kDa. Sequence similarity searches indicated that these enzymes were very similar, with an amino acid identity of 98 to 100% to *Salmonella enterica, Escherichia coli and Shigella dysenteriae*, suggesting that these were probably originating from bacteria. Physical maps of the 5 contigs are shown in Figure D2 (A, B, C, D, E, F, G, H, I) in the Appendices section D, showing the location and directional organisation of the 5 putative ORFs. Function-based screening, in this study, was focused mainly on screening for DNA polymerase genes with 5'-3' exonuclease activity. Although, the complementation of the cold- sensitive *E. coli* mutant successfully allowed the screening of the metavirome library, as has been previously reported (Nagano *et al.*, 1999; Simon *et al.*, 2009), it was not ideal in the current study as it did not yield novel DNA polymerase 1 genes or novel 5'-3' exonuclease domain. As a result of high sequence identity with known genes in the database (90 to 100%), the identified ORFs from the functional complementation screening approach was not considered for recombinant expression.

Table 5.2: **Representation of contigs and ORFs sequences showing homology to known DNA polymerase sequences from BLASTp searches (NCBI)**

Contigs	Length(bp)	Number of ORFs per contig	Putative nucleic acid manipulating enzyme encoding ORFs	ORF number within the contig	ORF re-name	Nucleotide Position within the contig	ORF length(bp) (aa)[a]	predicted pI (M_w in kDa)[b]	GC contents (%)	Identity (%)	Signal peptide [c]	predicted function (PFAM and BLASTp) [d]
Contig 213	16415	9	1	2	1	520-718	198/66	4.86 / 8	48	98	No	Phage exonuclease *Escherichia coli K12(*fig|83333.1.peg.535)
Contig 126	13235	8	1	8	2	11067-12432	1365/455	9.57 / 52	56	90	No	Poly(A) polymerase *Salmonella enterica subsp. enterica serovar Choleraesuis str. SC-B67(fig|321314.4.peg.289*)
Contig 539	1127	1	1	1	3	153-1122	969/323	5.05 / 78	54	95%	No	Retron-type RNA-directed DNA polymerase *Escherichia coli K12 (fig|83333.1.peg.253)*
Contig 101	4371	4	1	2	4	339-3126	2787/929	4.85 / 22	51	99%	No	DNA polymerase I (*Shigella dysenteriae)*
Contig 70	5081	4	1	3	5	3162-3918	756/252	5.10 / 62	52	99%	No	DNA polymerase 1 (*Escherichia coli*)

a = (aa= amino acid); bp =base-pair;

[b] The *isoelectric point (pI), Molecular weight (M_w) were predicted using* Expasy server (https://web.expasy.org/compute_pi/*)*

[c] Signal peptide coding sequences were predicted using SignalP 4.1 server (http://www.cbs.dtu.dk/services/SignalP/).

[d] Best hits were determined using BLASTP server (http://blast.ncbi.nlm.nih.gov/BLAST)

5.2.3 Expression, purification and enzyme activity assay of selected nucleic acid manipulating enzymes isolated using the sequence-based screening Approach

Due to low sequence novelty of the genes identified using the functional complementation screening approach (90 to 100% identity to known genes), a decision was taken to restrict the recombinant expression studies to those ORFs identified as part of the sequence-based screening approach. Of the 15 ORFs encoding putative nucleic acid manipulating enzymes that were identified using the sequence-based approach, one ORF per contig were selected and the following 9 ORFs were selected for recombinant expression studies: ORF 3, 4, 7, 9, 10, 12, 13, 14 and 15 (Table 5.1). The 9 ORFs were selected based on the complete length of the translated coding sequence (longer than 100 amino acids) and identity percentage (less than 60% identity to known genes).

5.2.3.1 Recombinant expression strategy

Sequence identity studies revealed that all 9 selected ORFs were closely homologous to ORF sequences encoding nucleic acid manipulating enzymes from different bacteriophages, namely *Rhizobium* phage, *Pseudomonas* phage, *Gordonia* phage *Smoothie*, *Cronobacter* phage, *iridescent virus*, *Yersinia* phage, *Thermus* phage and *Vibrio* phage *VpV262* as well as one unknown phage (Table 5.1). This observation led to an assumption that a direct cloning of the 9 putative genes derived from bacteriophage origin may not be efficiently translated and transcribed in *E. coli* the standard heterologous prokaryotic expression system due to different codon usage. This assumption is based on the fact that a number of studies indicated that the differences between tRNA levels and incompatible codon usage bias between species and organism domains (virus, prokaryotes, eukaryotes) can affect heterologous expression levels (Makrides, 1996; Gustafsson *et al.*, 2004; Burgess-Brown *et al.*, 2008). Codon optimisation as a strategy to improve high level recombinant protein production has been used successfully for decades to improve translation efficiency of heterologous genes (Gustafsson *et al.*, 2004; Burgess-Brown *et al.*, 2008; Uchiyama and Miyazaki, 2009; Culligan *et al.*, 2014; Khalili *et al.*, 2015; Strazzulli *et al.*, 2017). Burgess-Brown *et al.* (2008) successfully reported the improvement in heterologous protein expression levels in *E. coli* through the elimination of rare codons by a codon optimisation strategy (Burgess-Brown *et al.*, 2008). Similarly, Willner *et al.* (2009) reported on the difficulty associated with codon bias that led to inefficient translation of proteins derived from phage genomes in *E. coli*, which was subsequently

overcome by the codon optimisation approach (Willner *et al.*, 2009). Given this background, the devised expression strategy adopted in this study involved the *de novo* synthesis of all 9 identified ORFs using codons optimised for expression in *E. coli*. The 9 selected genes (*PolB* encoding putative DNA polymerase III subunit beta; *HNHc* encoding putative HNH endonuclease; *Lig1* encoding RNA ligase; *E7* encoding putative endonuclease VII; *Lig 2* encoding putative polynucleotide kinase/ligase; *Pol A1* encoding putative DNA polymerase; *Pol A2* encoding putative DNA polymerase I; *DNAlig* encoding putative DNA ligase and *RE* encoding restriction enzyme III) were synthesised with *Nde*I and *Xho*I restriction sites at the 5'and 3' prime of the gene sequences. All of the gene sequences lacked a stop codon at the 3'end of the genes to enable the in frame fusion with 6x His sequence in pET vectors series utilised, in order to facilitate downstream purification of the corresponding gene products. The 8 synthetic gene constructs were cloned in pUC57 cloning vector for stable maintenance of the genes at the *Nde*I and *Xho*I restriction sites ((Figure 5.4) and were accordingly named pUC57_PolB, pUC57_HNHc, pUC57_RNALig1, pUC57_RNALig2, pUC57_E7, pUC57_PolA1, pUC57_Pol A2 and pUC57_DNALig. However, one gene (*RE)* encoding a putative restriction enzyme III was unstable in pUC57 vector and the decision was taken to clone it directly in pET28 expression vector and named pET28_RE.

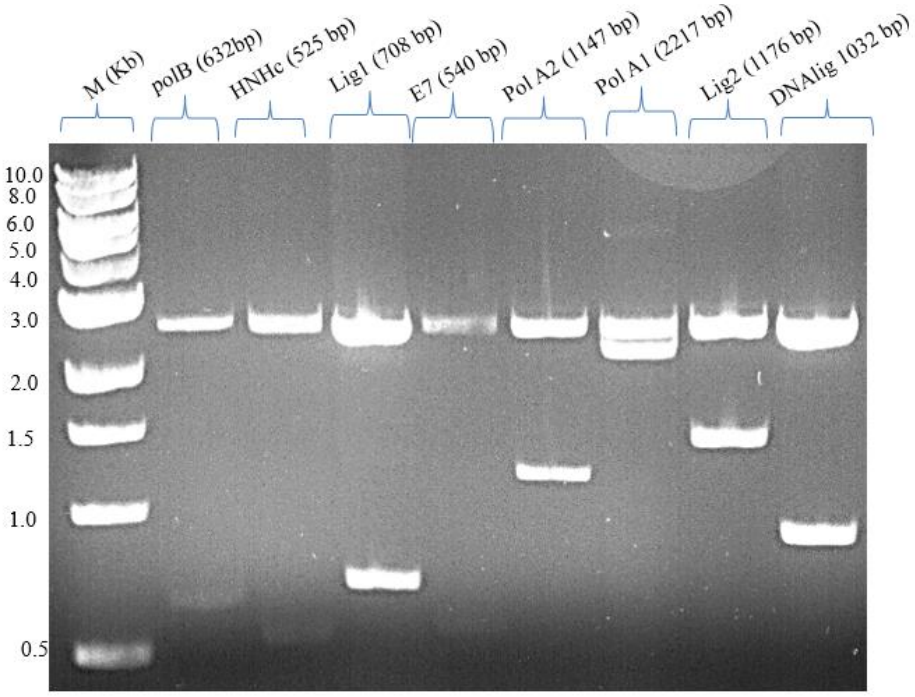


Figure 5.4: **Agarose (1% w/v) gel electrophoresis showing gene constructs provided in pUC57 cloning vector and digested with *Nde*I and *Xho*I restriction enzymes. Lane 1 = Marker, Lane 2-9 = pUC57 vector + gene inserts.**

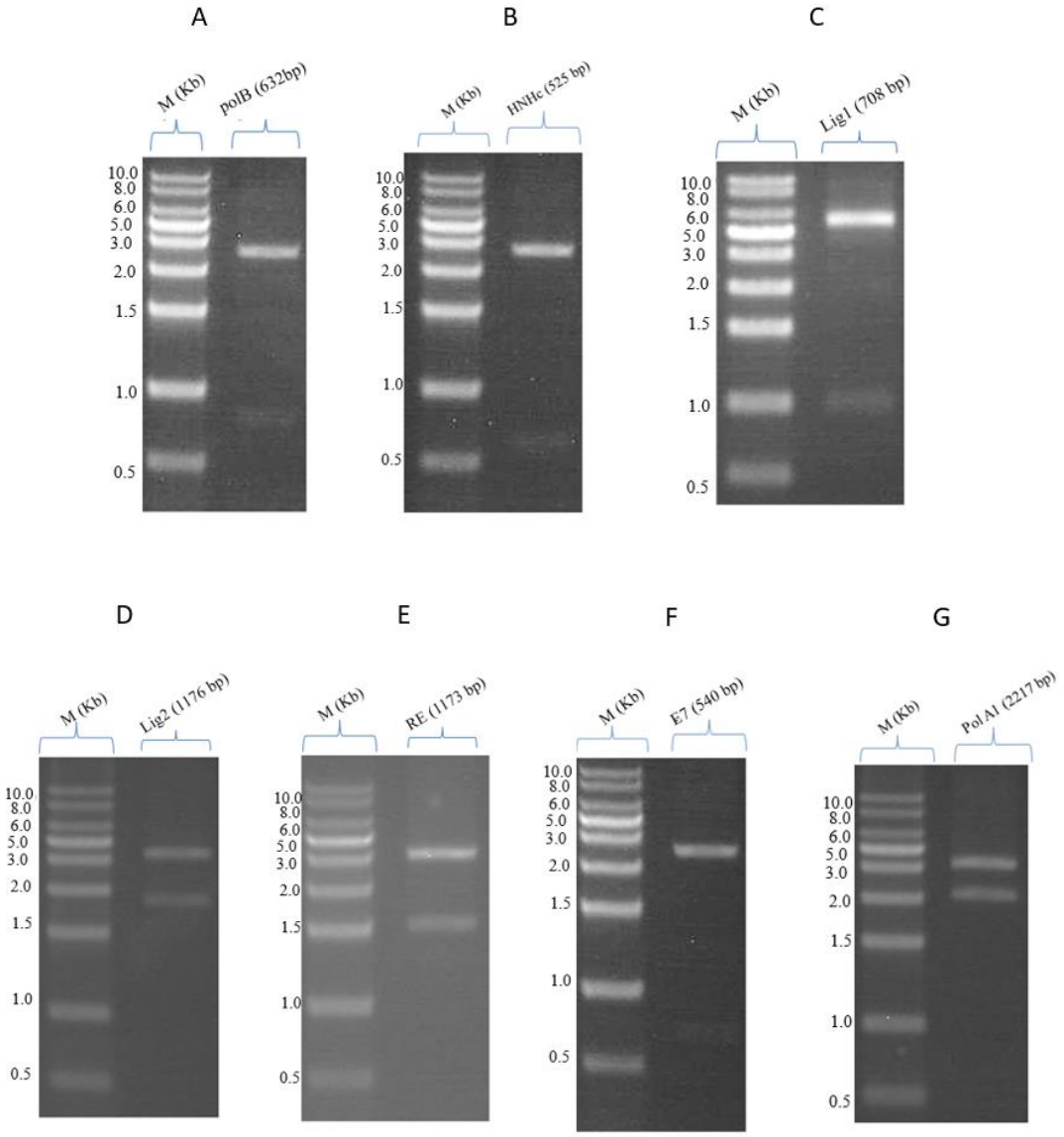
5.2.3.2 Development of expression systems

For the purpose of this study, pET expression system was initially chosen to express 9 identified genes encoding nucleic acid manipulating enzymes. This pET system was originally developed by Studier and Moffatt, (1986) and uses the bacteriophage T7 promoter to direct the expression of target genes. Two stringent promoter options are offered in this T7 promoter dependent system, such as plain T7 promoter and T7 *lac* promoter. T7 *lac* promoter is 25bp in length and is situated downstream from the promoter region (Mierendorf *et al.*, 1998).

The initial intent was to express all the genes in pET20b(+) expression vector which contains the T7 promoter and ampicillin selection marker. However, due to weak selection pressure offered by ampicillin selection marker (perhaps due to ampicillin resistance) and instability experienced with some gene products other expression vectors from the series, namely pET28a(+) and pET30b(+) were also tested. The pET28a(+) and pET30b(+) contain strong kanamycin selection marker and the T7 *lac* operator which, in the absence of *lac* repressor protein, reduces basal (background) expression levels. Furthermore, both vectors (pET28a(+) and pET30b(+)) offer an option for fusion the product of interest with either N or C terminus 6x histidine affinity tag to facilitate downstream purification (Mierendorf *et al.*, 1998).

The pUC57 derived gene fragments that encoded the nucleic acid manipulating enzymes were excised from the pUC57 parental vector using *Nde*I and *Xho*I restriction enzymes. The recovered DNA fragments were then ligated into the following pET expression vectors pre-restricted with *Nde*I/*Xho*I enzymes: RNA ligase 1, HNH endonuclease and DNA polymerase III subunit beta were cloned in pET20b(+) to yield pET20_RNALig1, pET20_HNHc and pET20_PolB, respectively; while the genes encoding RNA ligase 2, type III restriction enzyme, DNA polymerase A1 and endonuclease VII were cloned in pET28a(+) to produce pET28B_RNALig2, pET28_RE, pET28_PolA1 and pET28_E7. The genes encoding putative DNA ligase and DNA polymerase A2 were ligated in pET30b(+) to respectively yield pET30_DNAlig and pET30_PolA2 recombinant expression plasmids. This resulted in *E. coli*/pET expression hosts listed in Table 5.3. The presence of the correct gene fragments were checked by restriction digestion (Figure 5.5) and the correct in-frame, directional cloning of

the DNA inserts were confirmed by Sanger sequencing method using T7 promoter and T7 terminator primer pair (data not shown).



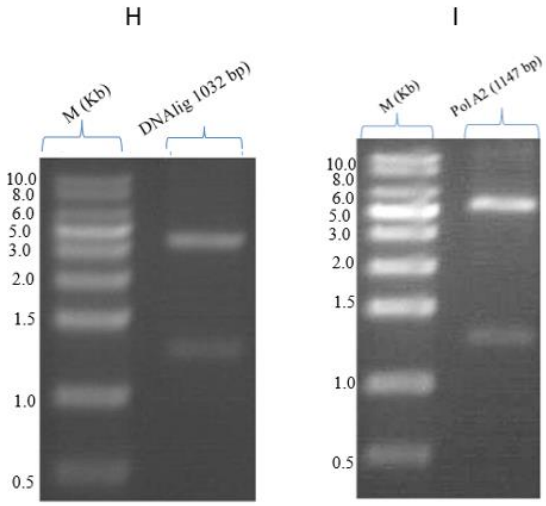


Figure 5.5: **Agarose gel electrophoresis of gene constructs provided in pET expression vectors. A, B and C represent gene (A = *pol B*, B= *HNHc* and C = *RNALig 1*) constructs in pET20 b (+) digested with *Xba*I and *Xho*I. D, E, F and G represent gene (D = *RNALig*, E = *RE*, F = *E7* and G = *Pol A1*) constructs in pET28a(+) digested with *Mlu*I and *Xho*I. H and I represent gene (H = *DNAlig* and I = *Pol A2*) a constructs in pET30b(+) digested with *Nde*I and *Xho*I. Lane 1 = Marker, Lane 2 = pET expression vectors + gene inserts.**

The nine developed recombinant plasmids were then used to transform the *E. coli* Bl21 strain to yield expression hosts: *E. coli* BL21 /pET20_*RNALig1*, pET20_*HNHc*, pET20_*PolB*; *E. coli* Bl21/ pET28_*RNALig2*, pET28_*RE*, pET28_*PolA1*, pET28_*E7* pET28_RNALig2, pET28_RE, pET28_PolA1, pET28_E7 and *E. coli* Bl21/ pET30_*DNALig1*, pET30_*PolA2* expression hosts (Table 5.3).

Table 5.3: **Nucleic acid manipulating enzymes constructs, with their size, restriction sites, and vectors**

Expression plasmids	Gene name	Description	Expected protein product (pI/Mw) (kDa)	Restriction sites	Marker gene	Expression *E. coli* host strain	Maintenance *E .coli* host strain
pET20_PolB1	*Pol B* (ORF 3)	pET20b(+) derived expression vector containing 210bp DNA polymerase III subunit Beta (*PolB1*) gene in frame with the T7 lac promoter gene and C-terminal his tag sequence.	5.99 / 23	*Nde*I: 5' C A ↓ T A T G 3' *Xho*I: 5' C ↓ T C G A G 3	Ampicillin	BL21 DE3, BL21 AiBL21 DE3 pLysS	DH5α strain
pET20_HNHc	*HNHc* (ORF 7)	pET20b(+) derived expression vector containing 174bp Putative HNH endonuclease (*HNHc*) gene in frame with the T7 lac promoter gene and C-terminal his tag sequence.	8.97 / 19	*Nde*I: 5' C A ↓ T A T G 3' *Xho*I: 5' C ↓ T C G A G 3	Ampicillin	BL21 DE3, BL21 AiBL21 DE3 pLysS	DH5α strain
pET20_RNALig1	*RNALig1* (ORF 10)	pET20b(+) derived expression vector containing 237bp RNA ligase gene (*RNALig2*) gene in frame with theT7 lac promoter gene and C-terminal his tag sequence.	6.54 / 26	*Nde*I: 5' C A ↓ T A T G 3' *Xho*I: 5' C ↓ T C G A G 3'	Ampicillin	BL21 DE3, BL21 AiBL21 DE3 pLysS	DH5α strain
pET28_RNALig2	*RNALig2* (ORF 11)	pET28a(+) derived expression vector containing 391bp RNA ligase T4 (RNALig2) gene in frame with the T7 lac promoter gene and C-terminal his tag sequence.	6.39 / 44	*Nde*I: 5' C A ↓ T A T G 3' *Xho*I: 5' C ↓ T C G A G 3'	Kanamycin	BL21 DE3, BL21 AiBL21 DE3 pLysS	DH5α strain
pET28_RE	*RE* (ORF 4)	pET28a(+) derived expression vector containing 248bp putative superfamily II DNA/RNA helicase (*RE*) gene product in frame with theT7 lac promoter gene and C-terminal his tag sequence.	6.77 / 65	*Nde*I: 5' C A ↓ T A T G 3' *Xho*I: 5' C ↓ T C G A G 3	Kanamycin	BL21 DE3, BL21 AiBL21 DE3 pLysS	DH5α strain
pET28B_E7	*E7* (ORF 9)		9.74 / 19	*Nde*I: 5' C A ↓ T A T G 3'	Kanamycin	BL21 DE3, BL21 AiBL21 DE3 pLysS	DH5α strain

Expression plasmids	Gene name	Description	Expected protein product (pI/Mw) (kDa)	Restriction sites	Marker gene	Expression *E. coli* host strain	Maintenance *E .coli* host strain
		pET28a(+) derived expression vector containing 179bp putative endonuclease VII (*E7*) gene in frame with the T7 lac promoter gene and C-terminal his tag sequence.		*Xho*I: 5’ C ↓ T C G A G 3			
pET28_PolA1	*Pol A1* (ORF 14)	pET28a(+) derived expression vector containing 738bp DNA polymerase type A family (*PolA1*) gene in frame with the T7 lac promoter gene and C-terminal his tag sequence.	8.99 / 83	*Nde*I: 5' C A ↓ T A T G 3' *Xho*I: 5’ C ↓ T C G A G 3	Kanamycin	BL21 DE3, BL21 AiBL21 DE3 pLysS	DH5α strain
pET30_PolA2	*Pol A2* (ORF 15)	pET30b(+) derived expression vector containing 741bp DNA polymerase type A family (*Pol A2*) gene in frame with the T7 lac promoter gene and C-terminal his tag sequence.	6.70 / 56	*Nde*I: 5' C A ↓ T A T G 3' *Xho*I: 5’ C ↓ T C G A G 3	Kanamycin	BL21 DE3, BL21 AiBL21 DE3 pLysS	DH5α strain
pET30_DNALig1	*DNAlig1* (ORF 13)	pET30b(+) derived expression vector containing 343bp DNA ligase AM (*DNAlig*) gene in frame with the T7 lac promoter gene and C-terminal his tag sequence.	5.56 / 35	*Nde*I: 5' C A ↓ T A T G 3' *Xho*I: 5’ C ↓ T C G A G 3’	Kanamycin	BL21 DE3, BL21 AiBL21 DE3 pLysS	DH5α strain

5.2.3.3 Expression of nucleic acid manipulating enzymes

Recombinant protein expression in one of the most widely preferred hosts such as *E. coli*, or in any other hosts for that matter, generally involves a trial and error approach. This is mainly because every protein is different. Thus, the strategies for the expression and purification of the proteins must be defined for each single case (Hannig and Makrides, 1998; Yildir *et al.*, 1998). However, possible reasons for poor protein expression have been identified and some of these include the toxicity in the host cell, over expression and inability of proteins to fold properly, which results in the formation of inclusion bodies and insolubility, and the formation of mRNA secondary structure preventing effective translation. Furthermore codon incompatibility can result in codons that are inconsistent with the host strain's available supply of tRNAs and can result in poor expression due to a halt in translation (Makrides, 1996; Yin *et al.*, 2007; Adrio and Demain, 2010). However, several approaches have been developed over the years to enhance heterologous protein expression in *E. coli* including selection of suitable expression host (Gottesman, 1996); addressing problems of plasmid and mRNA instability; gene product toxicity; choice of promoter (Goldstein and Doi, 1995); effect of growth medium composition; and the addition of fusion tags in the protein sequences which has been reported to enhance the yield of protein, increase the solubility, and even promote proper folding of the proteins. (Tract, 1984; Gräslund *et al.*, 2008; Joseph *et al.*, 2015).

An initial attempt to express the 9 nucleic acid manipulating enzymes encoding genes, (*PolB, HNHc, RNALig1, E7, RE, RNALig2, Pol A1, PolA2* and *DNAlig)* under standard growth conditions (LB medium, 37°C, 1mM IPTG over 24 hr growth period), sampled at the following time intervals (T0, T1 =1h, T2=2h, T3=3h, T4 =4h, T5=5h, T6=6h and overnight (ON), resulted in no detectable protein bands of the expected size in soluble intracellular fractions, but revealed large amounts of protein aggregated into inclusion bodies (Appendices section D3).

Following initial unsuccessful attempts to express the identified enzymes under the standard condition described above, a number of attempts were undertaken to enhance the solubility of proteins in the cytoplasm of *E. coli* host cells. This included changes in growth temperatures (37 to 18^{o}C); lowering inducer (IPTG) concentration from 1 to 0.1mM; testing various growth media (i.e. LB, 2YT and EnPresso® B), testing different *E. coli* host strains

(such as *E. coli* BL21 DE3, BL21 DE3 pLysS, and BL21 AI cells). A summary showing the different conditions employed to optimise recombinant production of the 9 gene products is outlined in Table D3 in the Appendices. Successful soluble expression was observed only for *Pol A1* and *DNAlig* genes under the optimised conditions described below.

5.2.3.4 pMAL expression strategy

An attempt was also made to express *PolA1* and *DNAlig* encoding protein as a fusion protein using pMAL-C5X expression system (NEB), which contains a multiple cloning site that allows for the translational fusion of the *E. coli* maltose-binding protein (MBP) with the cloned target protein at the N-terminus. The MBP-tag is encoded by the *malE* gene. Transcription of the recombinant gene fused with the MBP-tag is controlled by the induction of the *tac* promoter (P_{tac}). Basal expression is minimised by the binding of the lac repressor to the *lac* operator immediately downstream of P_{tac}. The lac repressor is encoded by the *lac*I gene, (Lauritzen *et al.*, 1991; Kapust and Waugh, 1999). The *PolA1* (2217bp) and *DNAlig* (1032bp) genes were respectively excised from pET28_PolA1 and pET30_DNALig1 expression vectors using *Nde*I and *Xho*I restriction enzymes and sub-cloned into a pMAL-C5X (Genscript) linearized with the same enzyme to yield pMAL-C5X-*PolA1* and pMAL-C5X_*DNAlig* expression plasmids. This cloning strategy enabled both the *PolA1 and DNAlig* genes to be fused to *malE* gene which encodes for a 42.5kDa cytoplasmic MBP to yield an MBP-Pol A1 protein fusion product of 127kDa and MBP-DNAlig fusion product of 78 kDa. In both the constructs, a linker sequence (SNNNNNNNNNN) and a TEV cleavage site (ENLYFQG) (where cleavage occurs after Q, resulting in G becoming the new N-terminal) were used as tag for the two genes. The layout of the two expression cassettes is shown in Figure 5.6.

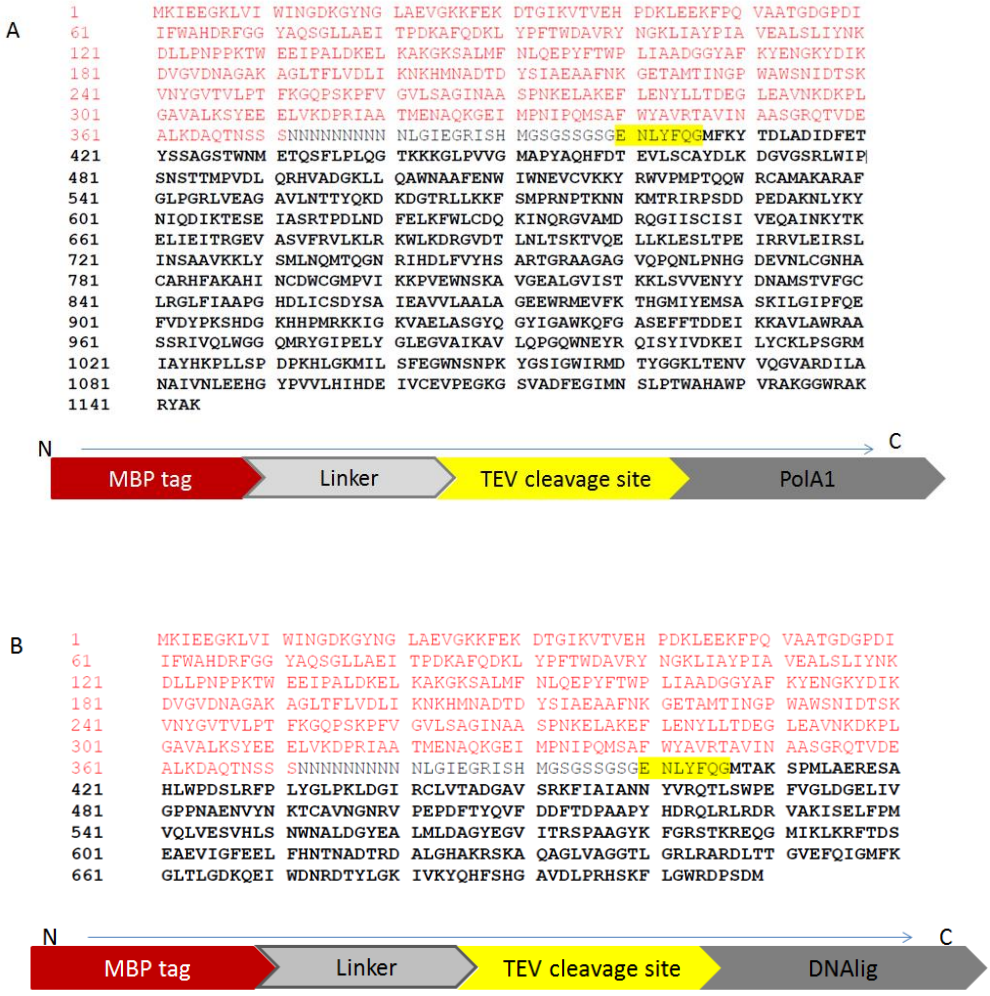


Figure 5.6: Expression cassettes representing sequences of the MBP-Tag-constructs sub-cloned into *E. coli* expression vector pMAL-c5X. A. MBP-Tag-*PolA1* and B. MBP-Tag-*DNAlig*. MBP-tag is represented in red, linker sequence in grey, the Tobacco Etch Virus (TEV cleavage site) highlighted in yellow and the gene sequences in grey.

The total cell protein, soluble and insoluble fractions of the *E. coli* BL21 (DE3) cells harbouring pMAL-C5X_*PolA1 and* pMAL-C5X_*DNAlig* expression plasmids were analysed by SDS-PAGE (Figure 5.7 A and B). SDS-PAGE analysis revealed extra protein band at 127kDa corresponding to the predicted molecular weight of MBP (43kDa) fused to Pol A1 (83kDa) in the induced total, soluble and insoluble protein fractions of *E. coli* BL21 (DE3) cells harbouring pMAL-C5X_*Pol1A1* (Figure 5.7 A). Likewise, an additional protein band at 78kDa corresponding to the predicted molecular weight of MBP (43kDa) fused to DNAlig (35kDa) was observed in the induced total, soluble and insoluble protein fractions of *E. coli* BL21 (DE3) cells harbouring pMAL-C5X_*DNAlig* (Figure 5.7 B).

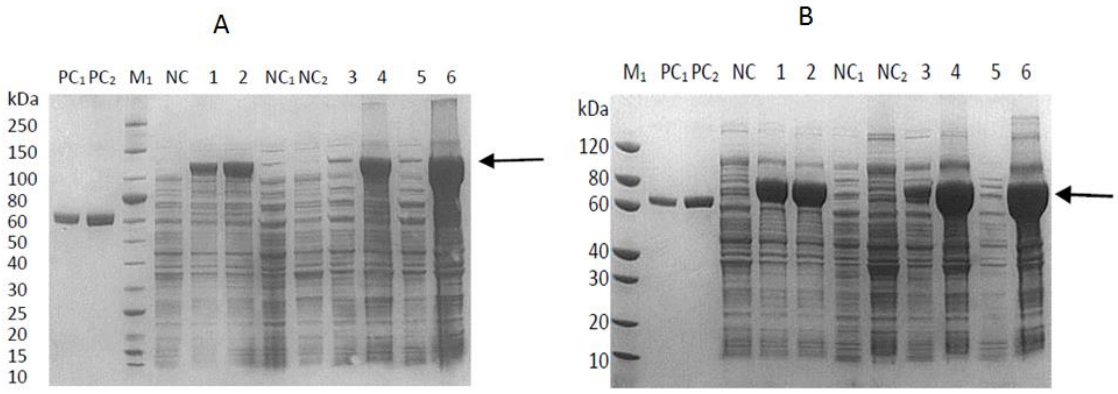


Figure 5.7: **A: SDS-PAGE of A. pMAL-C5X-*PolA1* and B. pMAL-C5X-*DNAlig*. Lane M1: Protein marker, Lane PC_1: BSA (1 µg), Lane PC_2: BSA (2µg), Lane NC: uninduced cell lysate, Lane 1: induced cell lysate for 16h at 15°C, Lane 2: induced cell lysate for 4h at 37°C, Lane NC_1: non-induced supernatant of cell lysate, Lane NC_2: non-induced pellet of cell lysate, Lane 3: induced supernatant of cell lysate for 16h at 15°C, Lane 4: induced pellet of cell lysate for 16h at 15°C, Lane 5: induced supernatant of cell lysate for 4h at 37°C, Lane 6: induced pellet of cell lysate for 4h at 37°C.**

A semi-quantitative analysis of the expression profile based on BSA as a standard using the equation shown in the appendices section D revealed that MBP fused Pol A1 and DNAlig were being expressed at concentration levels of approximately 70 and 120 mg/L respectively.

However, only an estimated 3% and 10% of the MBP-PolA1 and MBP-DNAlig were expressed as soluble fractions, with the rest being expressed as inclusion bodies (Fig 5.7A and B, Lane 3 and 5). Despite numerous attempts to express the two proteins at different temperatures, there was no improvement in the amount of soluble expressed proteins (data not shown). Due to the size of the MBP-tag, which often interferes with protein activity, some studies have recommended its removal prior any further characterisation (Norgard *et al.*, 1978). Furthermore, attempts to purify both MBP-Pol A1 and MBP-DNAlig fused proteins using one step-maltose affinity column and subsequent removal of MBP by cleaving it off with endoproteases at a TEV cleavage site (ENLYFQG) failed, resulting in very negligible amounts of free Pol A1 and DNAlig proteins recoverable for further characterisation (data not shown). The cleavage of the Tag also resulted in the cleaved protein which was insoluble.

5.2.3.5 The *EnBase* ® cultivation technology and purification of DNA ligase

The only successful expression achieved with the pET vector expression system was that of pET30/DNAlig under the *EnBase* ® cultivation technology. The principle of the *EnBase*® cultivation technology is that an enzyme *(EnBase* ®) is used to gradually release glucose from a solubilised polysaccharide substrate into the culture medium as the primary carbon source. The amount and activity of the enzyme directly regulates the rate of release of glucose over a period of time in the culture medium (Krause *et al.*, 2016). The principle of the *EnBase* ® cultivation technology for recombinant protein production optimisation involves; delivering of glucose constantly until the process reaches the end of the protein production phase, an addition of a balanced mixture of inorganic and organic ammonia compounds for self-sustainable pH when there is a lack of glucose and by adjusting the amount of enzyme added, it increases the probability to adapt the system to various aeration conditions (Krause *et al.*, 2016).

Using the *EnBase*® cultivation medium (18^{o}C, incubation temperature, 0.1mM IPTG concentration), a recombinant DNAlig was produced in a biologically soluble form in the cytoplasmic fraction of E. *coli* Bl21 cells. Since the pET30/DNAlig expression construct was designed to allow for recombinant DNAlig to be fused with the C- terminal 6× histidine tag, the soluble recombinant DNAlig was purified in a single step immobilized metal affinity chromatography (IMAC) procedure. Analysing the purified sample revealed a purified protein band at 36kDa corresponding to the predicted molecular mass of 36.895 Da and 344 amino acid length (Figure 5.8A).The concentration of purified DNAlig was estimated to be approximately 75mg of protein from a 1L culture or 75µg/mL (Figure 5.8B).

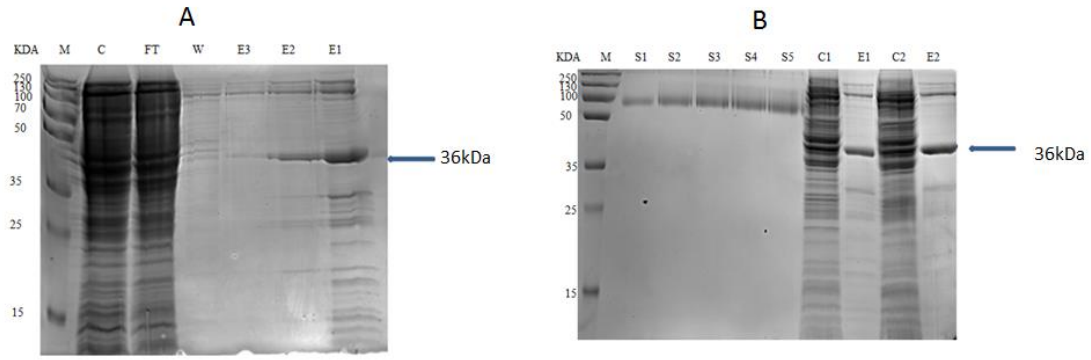


Figure 5.8: **SDS-PAGE gel of the A: crude and purified DNA Ligase and B: BSA gel. Lane M= Marker, Lane C= crude sample, Lane FT = Flow though sample, Lane W= washed sample and**

Lanes E1, E2 and E3 = elution sample, with the expected protein band corresponding to sizes 36kDa (DNAlig) indicated by an arrow. Standard concentration used was 2, 1.5, 1, 0.75, and 0.5µg/µL. 50 µL of the standard and the samples were loaded on the gel.

5.2.3.6 Homology searches and primary structure analysis of DNAlig ORF

The highest sequence identity scores for DNALig in the Genbank database, from a global amino acid alignment were with hypothetical protein from *Elusimicrobia bacterium (30%),* putative DNA ligase from *Actinobacteria bacterium* (29%); hypothetical protein from *Bacillus andreraoultii* (27%) and putative ATP-dependent DNA ligase *Salinibacterium sp.* (27%).

To gain further understanding regarding features that could be both functionally and structurally important in the DNAlig primary structure, a multiple sequence alignment was constructed combined with motif and domain search analysis (Figure 5.9). Despite few areas of sequence homology, the primary structure of DNA ligases is generally characterised of six motifs, i.e.; motif I, III, IIIa, IV, V and VI (Shuman and Schwer, 1995).These motifs were represented by the following sequences in the alignment (where *-represent any amino acid): Motif 1: EYKYDGER (34 -39); Motif III: FILDGEXV (74-81); Motif IIIa: CLFAFDILYL (92-98); Motif IV: XG**EGLXV (216-219) ; Motif V;:WLKXKXDYL (322-328) and motif VI: PRFLRIREDK. The EYKYDGER motif is known to contain the active site lysine nucleophile which, together with the other four motifs, plays an important role in the first step of the DNA ligase reaction wherein the AMP becomes covalently bonded during the nucleotidyl transfer (Martin and MacNeill, 2002). Based on the multiple sequence alignment, it can be deduced that Lys34 constitute the active site nucleophile lysine, whilst the following sequences KFDGNR (39-42), VLDGEL (77-82), LNASTPL (92-99), GCEGFM (192-199), KFKWLST (216-222) and WRLDKTFT (323-329) constitute motif I, III, IIIa, IV, V and VI respectively for ORF13 DNA ligase identified in this study.

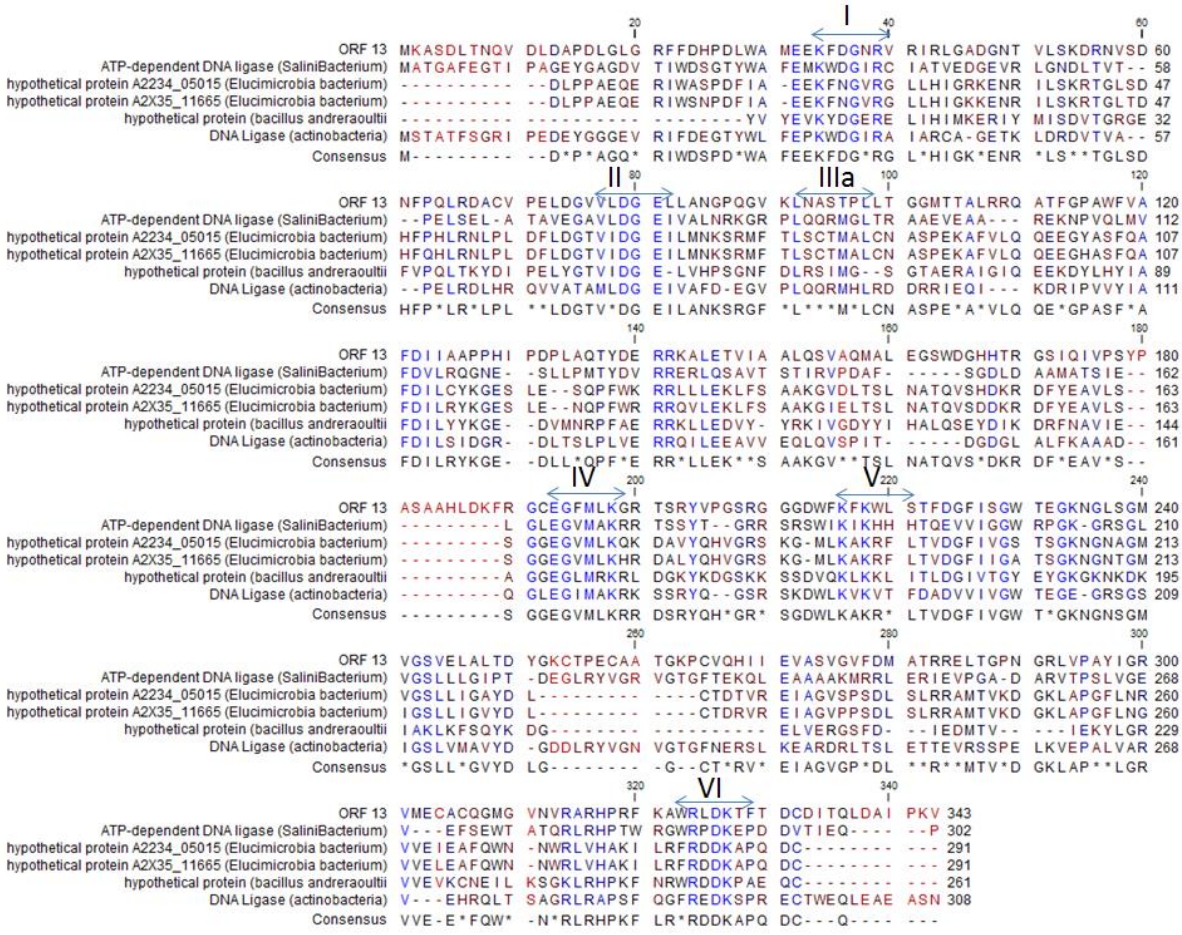


Figure 5.9: **Multiple sequence alignment of the expressed DNAlig protein to its closest hit hypothetical protein A2234_05015 [*Elusimicrobia bacterium* RIFOXYA2_FULL_58_8] crobia bacterium, DNA ligase [*Actinobacteria bacterium*], hypothetical protein A2X35_11665 [*Elusimicrobia bacterium* GWA2_61_42], hypothetical protein [*Bacillus andreraoultii*] and ATP-dependent DNA ligase [*Salinibacterium sp*. S1194]. The sequences that are conserved define a covalent nucleotidyl transferases superfamily. Six motifs I, III, IIIa, IV, V, VI, conserved in ATP-dependent DNA ligases were designated. The conserved motives are highlighted with blue and labelled as I, II, IIIa, IV, V and VI.**

5.2.3.7 Ligation assays

Generally, the determination of the ligase activity is based on a combination of radioactive labelling or fluorescent staining with gel electrophoresis. A registered laboratory for the handling of radioisotopes is also required for the former approach (Marchetti *et al.*, 2006; Zakabunin *et al.*, 2011; Taylor, 2014). More advanced techniques to provide quantifiable DNA ligase assay data include molecular beacon-based assays, electrochemistry-based methods and surface Plasmon resonance (Pergolizzi *et al.*, 2016). Therefore, DNAlig activity

was qualitatively determined using agarose gel electrophoresis that does not involve any radioactive labels. The DNA ligase can be quantitatively assessed by functional activity using known amounts of recombinant protein in a ligation reaction with a vector followed by transformation in *E. coli* and counting of the transformants.

5.2.3.8 Co-factor dependent

To confirm the bioinformatics assignment of the DNAlig protein as an ATP-dependent enzyme, the ligation reaction of pre-digested lambda DNA was performed in the absence of added ATP, using PBS without ATP and a commercial buffer (250mM Tris-HCl (pH 7.6), 50mM $MgCl_2$, 5mM ATP, 5mM DTT, 25% (w/v) polyethylene glycol-8000), containing ATP. Approximately 1.5µg of purified DNAlig protein was used to ligate 0.16pmol ends of lambda DNA digested with *Pst*I restriction enzyme. The positive control, ATP-dependent T4 DNA ligase (Invitrogen), used was able to ligate the DNA in the presence of ATP and unable to ligate the DNA in the absence of ATP as a cofactor. Similarly, it was found that the recombinant DNAlig protein could not perform the ligation reaction efficiently in the absence of ATP as a cofactor, and was able to ligate the DNA fragments in the presence of ATP as a cofactor, represented by the disappearance of the majority of low molecular weight bands and appearance of higher molecular weight band smears (Figure 5.10). Therefore, for further ligation reactions, a decision was made to use a commercial buffer with constituents that are required for ligation reactions of ATP-dependent T4 DNA ligase, which provided optimal additives that are required for the catalytic reaction of the ATP-dependent DNA ligase (Georlette *et al.*, 2000).

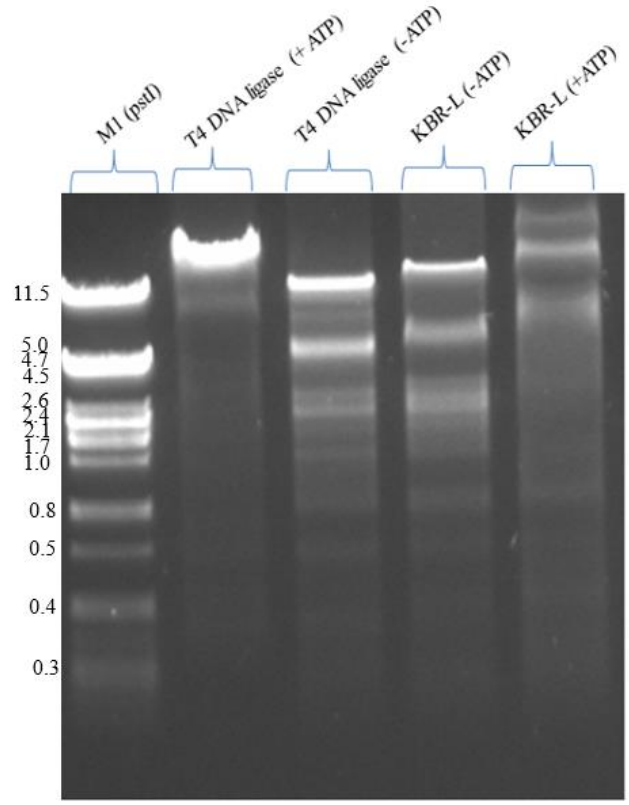


Figure 5.10: **The effect of ATP co-factor on DNAlig protein activity. Lane 1 = lambda DNA digested with *Pst*I as a marker. Lane 2 = ATP-dependent T4 DNA ligase with ATP included in the buffer. Lane 3 = ATP-dependent T4 DNA ligase without ATP included in the buffer. Lane 4 = KBR DNAlig without ATP included in the buffer. Lane 5 = KBR DNAlig with ATP included in the buffer.**

5.2.3.9 Blunt ends ligation assays

The ability of DNAlig to ligate blunt-end DNA molecules was also investigated using lambda DNA. Lambda DNA molecule is a linear double-stranded helix of 48,502bp in length (Sanger *et al.*, 1982). For this purpose, lambda DNA was digested with *Pst*I for sticky end and *EcoR*V for blunt end ligations. The reaction mixture was incubated at 25ºC for 1hr. To define a unit, 3 different amounts of DNAlig (0.5µg, 1.0µg, and 1.5µg) were used. A unit of DNAlig is defined as the amount of enzyme required to ligate 50% of cut lambda DNA in 1hr. As a result, the general increase in molecular weight and the disappearing of the low molecular weight bands with increasing amount of DNAlig suggests that the ligase reaction appears to be working with sticky-ended fragments digested with *Pst*I and blunt-ended fragments digested with *EcoR*V, as indicated by a smear on the agarose gel (Figure 5.11). The appearance of a smear could have resulted from DNAlig remaining very tightly bound

to the ligated lambda DNA after ligation reaction. To prevent gel shift and to check the efficiency of the ligation reaction, in future, it would be advisable to include a 30-minute proteinase K treatment in the presence of SDS (Bauer *et al.*, 2017). Additionally, including SDS in the loading dye will prevent the appearance of a smear and will make it easier to judge the ligation result. Alternatively, ligation of linear DNA with one 5'P group can be done so that the multimers are easier to be observed.

Commercial T4 DNA ligase was used as a positive control and was observed to ligate Lambda DNA, as shown by one single high molecular band as compared to DNAlig protein. Lambda DNA molecule consist of single-stranded complementary 12 nucleotide sticky ends. Restriction digestion is capable of inducing end-to-end DNA assembly which can be mediated by hydrogen bonding and stacking of the bases between complementary base pairs, therefore forming DNA multimers. The nature of the ends (sticky or blunt ends) and the ionic strength of the ligation solution can strongly influence the molecular weight distribution of DNA (Haber and Wirtz, 2000). As seen on the gel, high molecular weight was seen more on the sticky end ligations than the blunt-end ligation, suggesting the formation of undesired very long linear DNA fragments (multimers).

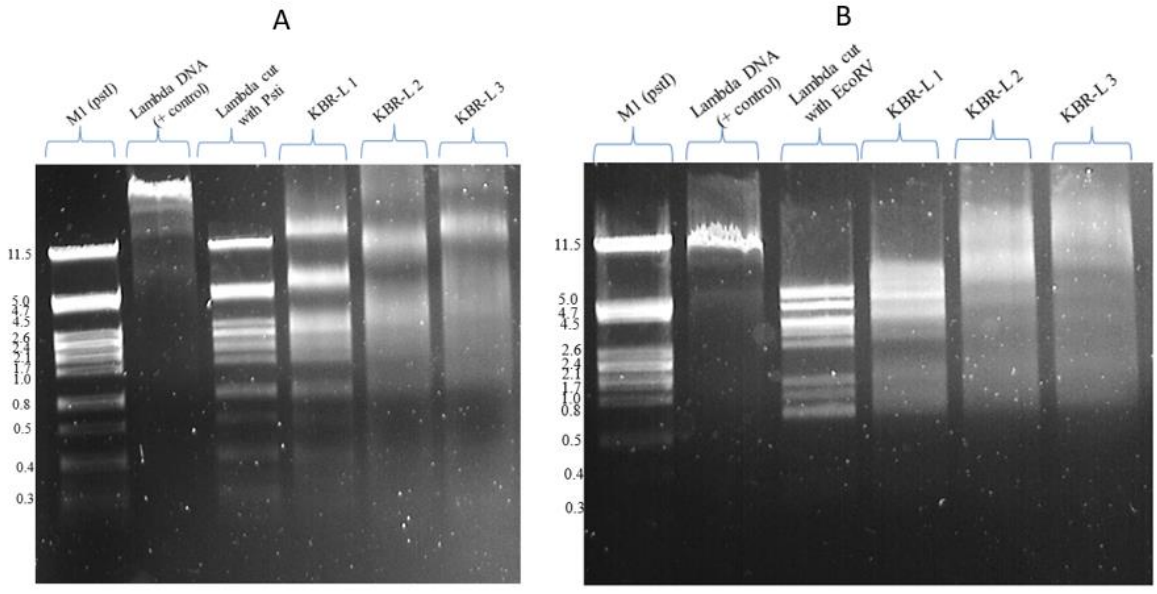


Figure 5.11: **Agarose gel electrophoresis of blunt-ended and sticky-ended DNA ligation by 3 different concentrations of DNAlig. A: lambda DNA digested with *Pst*I and ligated with DNAlig. B: lambda DNA digested with *EcoR*V and ligated with DNAlig. Lane M1= *PstI* marker, Lane 2: Lambda DNA (positive control treated with commercial T4 DNA ligase), Lane 3: Lambda digested with *Pst*I and *EcoR*V, Lane 4, 5 and 6: 0.5µg, 1.0µg and 1.5µg DNAlig**

respectively. Reactions were incubated at room temperature overnight. Lane M, Lambda *Pst*I DNA marker.

The DNAlig was also assayed for its ability to ligate the 3' blunt ends and the 5' sticky ends using 0.16pmols ends of pUC57 cut with *Sma*I and *Nhe*I, respectively. The results showed that DNAlig is capable of ligating both 3' blunt ends and the 5' sticky ends. Ligation of the 5' sticky ends resulted in all the 3 forms of the plasmid and the 3' blunt ends ligation only produced one band and a smear. Commercial T4 DNA ligase effectively ligated both 3' blunt ends and the 5' sticky ends and resulted in 2 forms of plasmids for blunt ends and all forms of plasmids for sticky (Figure 5.12). The smear could have resulted from the ligase bound to the substrate DNA or contamination of the ligation reaction with nuclease. As mentioned above, in future, treatment of ligation reaction with proteinase K and SDS before running the gel will dissociate the enzyme from the substrate DNA. Furthermore, nuclease contamination can be solved by changing running buffer and cleaning up the DNA before running a fresh agarose gel (Bauer *et al.*, 2017).

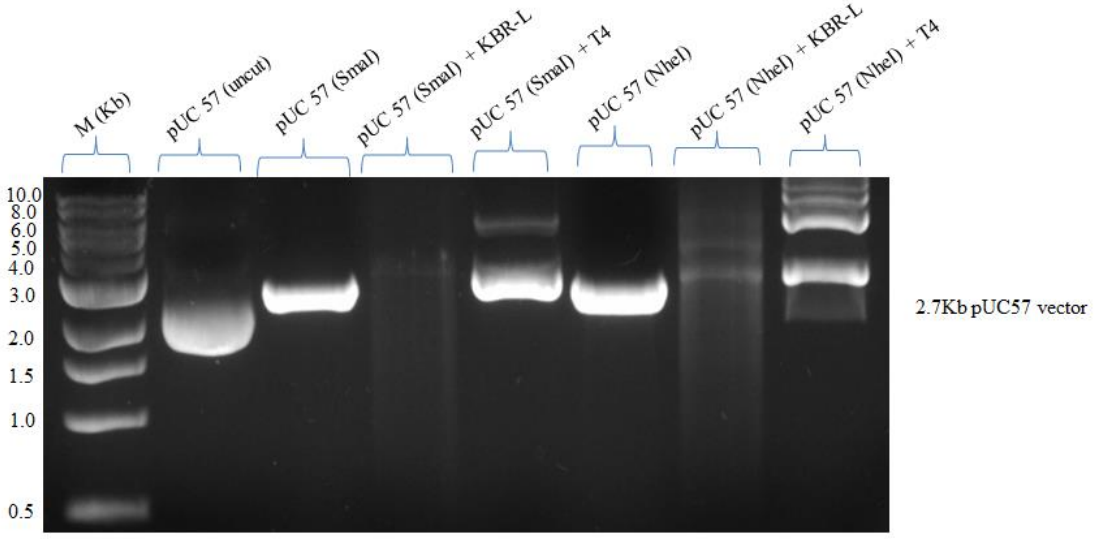


Figure 5.12: **Ligase assays for the ability to ligate the 3' blunt ends and the 5' sticky ends, lanes 1 = 1KB marker, Lane 2 = pUC57 uncut, Lane 3 = pUC57 cut with *Sma*I for 3' blunt ends, Lane 4 = Ligations of 3' blunt ends with KBR DNAlig, Lane 5 = ligation of 3' blunt ends with commercial T4 DNA ligase. Lane 6 = pUC57 cut with *Nhe*I for 5' sticky ends, Lane 7 = ligation of 5' sticky ends with KBR DNAlig and lane 8 = ligation of 5' sticky ends with commercial T4 DNA ligase.**

Results from the transformations of competent DH5α *E. coli* cells with the pUC57 digested with *Sma*I and *Nhe*I and ligated with DNAlig and T4 DNA ligase are outlined in Table 5.4. Concentrations of DNA were adjusted to 1µg for transformations, to rule out any effects due to the concentration of the DNA. Comparing the transformation efficiency of the pUC57 ligated with DNAlig and positive control, the results showed that pUC57 digested with *Sma*I and ligated with DNAlig transformation efficiency was = 920000CFU and pUC57 digested with *Nhe*I and ligated with DNAlig transformation efficiency was = 1600000CFU. However, transformations with positive control T4 DNA ligase (digested and ligated with T4 DNA ligase) resulted in large amount of colonies, which reveals that the pUC57 ligation reaction with DNAlig was relatively less efficient. pUC57 circular vector (undigested and unligated) as a positive control, showed a large amount of colonies from plating 100µL of the transformation reaction, indicating the successful transformation of the cells. No colonies were observed in the negative control (transformation with H_2O). The growth of transformants with pUC57 digested (unligated) and treated with phosphatase was likely due to incomplete digestion of the plasmids, or rather incomplete dephosphorylation. However, the colonies (70CFU for *sma*I and 80CFU for *Nhe*I) (pUC57 digested, dephosphorylated and unligated) were much fewer in number than the positive control (1600 CFU for *smaI* and 2400 CFU for *Nhe*I) (pUC57 digested, dephosphorylated and ligated) and the DNAlig ligated pUC57 plasmids, indicating that that there was ligation occurring in the efficacy of the DNAlig ligation.

Table 5.4: **Transformation efficiency of *E. coli* DH5α with pUC57 digested with *Sma*I and *Nhe*I and ligated with DNAlig and T4 DNA ligase and the efficiency calculation**

Plasmids	Transformation efficiency
pUC57 (supercoiled+ control)	5000000
pUC57_*Sma*I	70000
pUC57_ *Sma*I + KBR ligase	920000
pUC57_*Sma*I + T4 DNA ligase	1600000
pUC57_*Nhe*I	80000

pUC57_*Nhe*I + KBR ligase	2000000
pUC57_*Nhe*I + T4 DNA ligase	2400000

Despite some growth on the pUC57_*Sma*I dephosphorylated (unligated) control plate of the transformations, a decision was made proceed with plasmid extraction and digestion to confirm identity of transformants. This was due to the consideration that much fewer colonies on the control plates were observed as compared to the pUC57 ligated with DNAlig and T4 DNA ligase plates. The plasmids were quantified by NanoDrop and confirmed by agarose gel electrophoresis. Figure 5.13 shows the digested plasmids isolated from the transformations. This shows that pUC57 was successfully linearized, as the uncut plasmid appears to have migrated more rapidly at the bottom of the gel than the linearized plasmid represented by the top band.

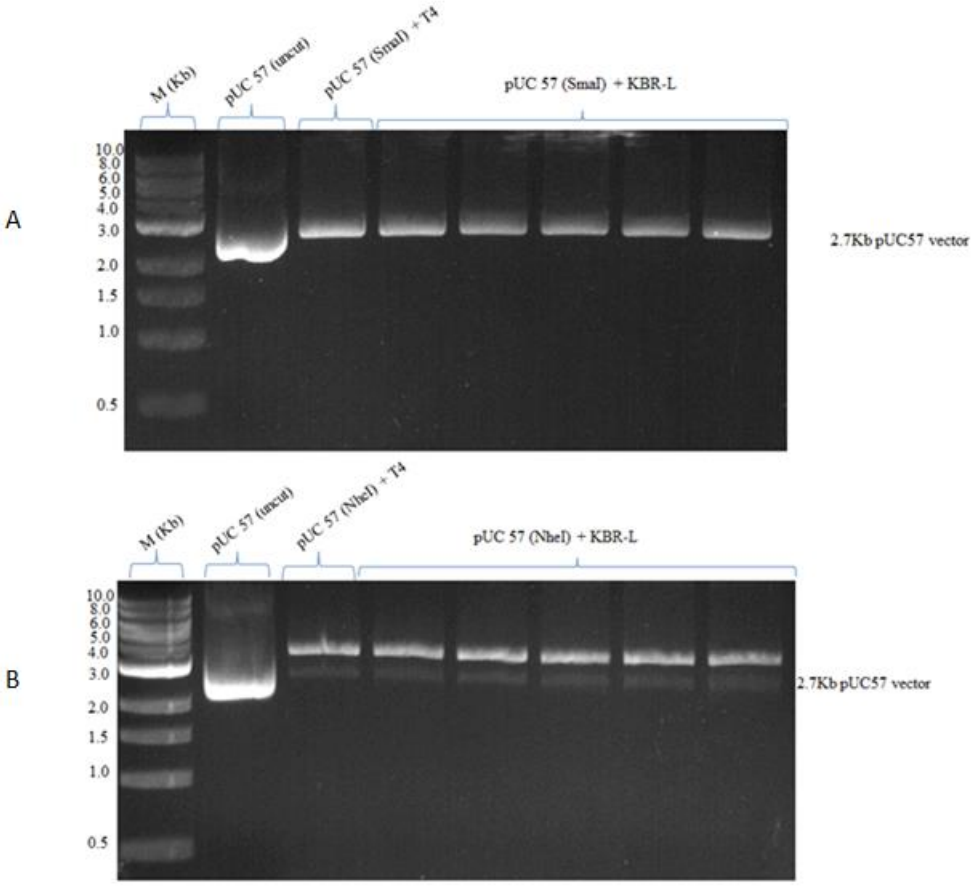


Figure 5.13: **Agarose gel electrophoresis for analysis of digestions of isolated competent DH5α *E. coli* cells transformants of pUC57. Lane 1 = 1kb marker, Lane 2 = undigested pUC57, lane 3 = positive control transformants, Lane 4-8 = Plasmids digested with A: *Sma*I and B: *Nhe*I and ligated with KBR DNAlig protein transformants.**

5.2.3.10 Comparing DNAlig with commercial ligases

Commercial ligases from Invitrogen, Thermo Scientific, and Roche were compared with the recombinant DNAlig expressed in this study. One unit of each commercial enzyme was used to perform ligations as described by the manufactures protocols. For DNAlig, 1µg was added to the ligation reaction. Lambda DNA, 0.16pmols, cut with *EcoR*V and *Sma*I was used to assay the ligases for the ligation of the blunt ends and *EcoR*I *and BamH*I for the sticky ends. The enzyme was assayed at 25ºC overnight. Blunt-end ligations performed with Thermo Scientific and Roche commercial ligase resulted in bands with highest molecular weight (Figure 5.14) indicating that the enzymes are more effective. Since the activity was monitored by joining of lambda fragments, KBR DNAlig was less effective than commercial ligases from Thermo Scientific and Roche commercial. However, DNAlig ligated better than Invitrogen ligase as shown by a smear in the lambda DNA cut with *EcoR*V and a higher molecular weight band in the lambda DNA cut with *Sma*I. Similarly, in sticky ends ligation of lambda DNA cut with *EcoR*I and *BamH*I, ligations performed with Thermo Scientific and Roche commercial ligase were more effective as compared to DNAlig, which was in turn more effective than Invitrogen ligase (Figure 5.15). Ligations using T4 DNA ligase from Thermo Scientific and Roche produced multimers of DNA as shown by high molecular weight band on the agarose gel. This could have resulted from buffer constituents (e.g. higher concentration of Mg^{2+} in the buffer induce multimer formation), as buffers specific for each commercial ligase was used (Haber and Wirtz, 2000). As mentioned earlier, the smear could have resulted from a nuclease activity. However, a transformation assay with the ligation reaction can be able to assess how much DNA ligated and successfully transformed *E. coli* and rule out nuclease contamination.

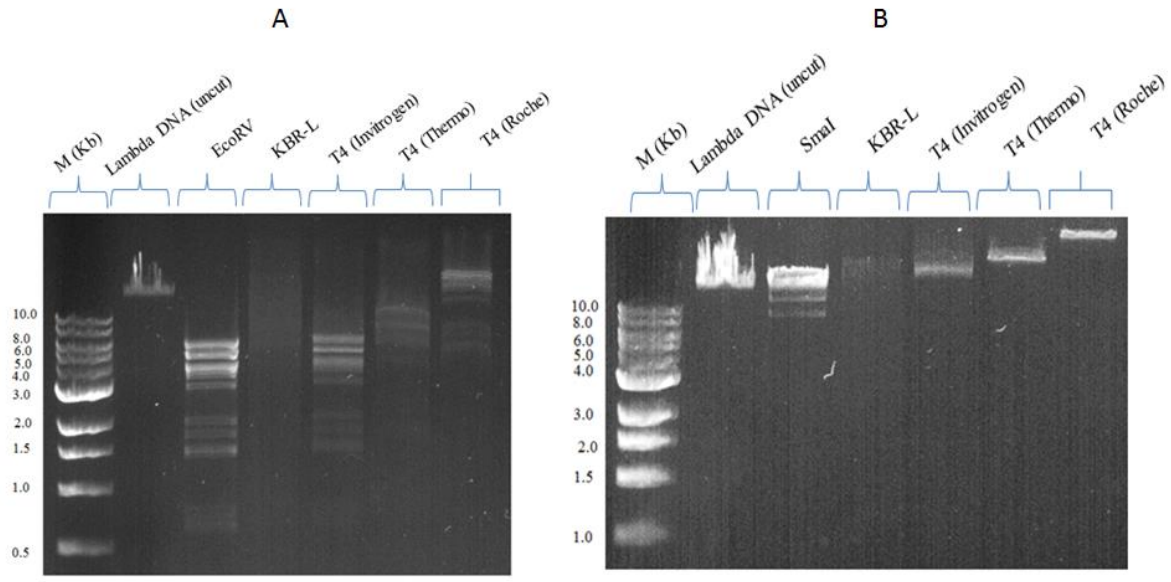


Figure 5.14: **Agarose gels analysis for DNA ligase assays done at 25ºC for blunt-ended lambda DNA cut with A: *EcoR*V and B: *Sma*I. Lane M = 1kb marker, Lane 2 = uncut lambda DNA, Lane 3= lambda DNA cut with *EcoR*V or *Sma*I, Lane 4 = lambda DNA ligated with KBR DNAlig, Lanes 5, 6 and 7 = lambda DNA ligated with T4 ligase from Invitrogen, Thermo Scientific and Roche, respectively.**

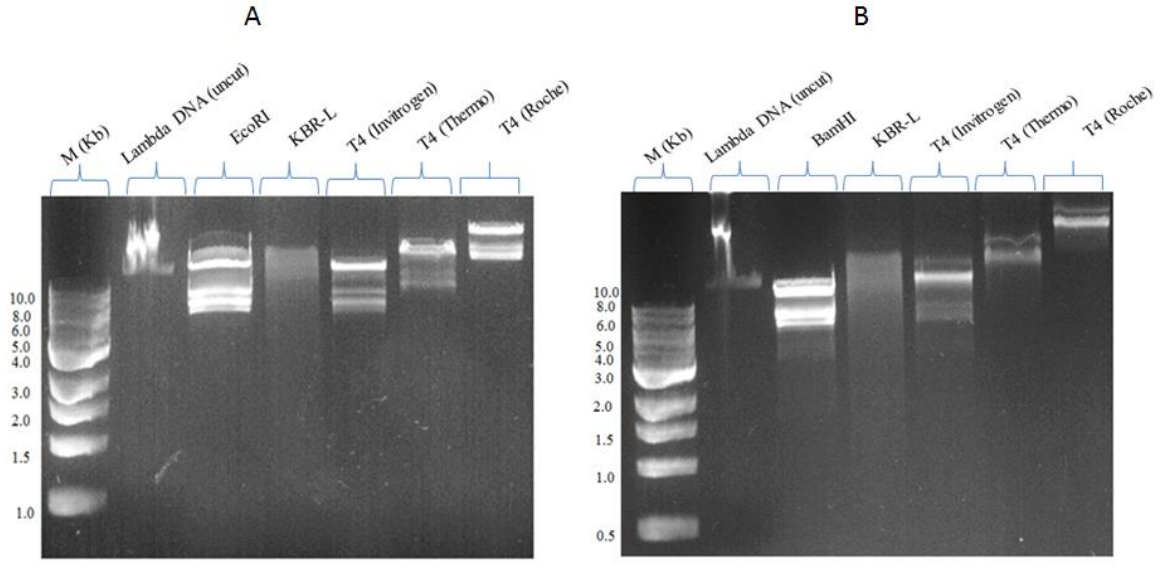


Figure 5.15: **Agarose gel analysis for DNA ligase assays performed at 25ºC for sticky-ended lambda DNA cut with A: *Eco*RI and B: *BamH*I. Lane M = 1kb marker, Lane 2 = uncut lambda DNA, Lane 3= lambda DNA cut with *Eco*RI or *BamH*I, Lane 4 = lambda DNA ligated with KBR DNAlig, Lanes 5, 6 and 7 = lambda DNA ligated with T4 ligase from Invitrogen, Thermo Scientific and Roche, respectively.**

5.2.3.11 Vector and inserts ligation assays

In order to assess the ability of recombinant DNAlig to join intermolecular ends, a 2.7kb pUC57 vector was digested with *Age*I and *Xho*I to release a 1.5kb insert and a 5kb pET28 vector was digested with *Nde*I and *Xho*I to release a 2kb insert in standard conditions at 37°C overnight (data not shown). The resulting fragments were excised from a 1% agarose gel, and extracted using a Geneget gel extraction kit according to the manufacturer's directions. The concentration of the vectors and inserts were calculated by NanoDrop (2.13 pmols ends of the pUC57 vector, 7.37pmols ends of the insert) (2.41pmols ends of the ng/µL pET28 vector and 5.66pmols ends of the insert). A ligation was set up with an insert to vector ratio of 3:1 (50ng vector and 150ng insert amount used) using DNAlig and T4 DNA Ligase (Invitrogen) as a positive control and the reaction was incubated at 4°C overnight (Figure 5.17). As observed in Figure 5.9, ligation is visible for pUC57 in the DNAlig and the control T4 DNA ligase lanes as indicated by the 3 forms of plasmids and the disappearing of the insert band. Ligation was also visible for pET28 in the DNAlig lane indicated by a shift of the band and the disappearing of the insert band. Three forms of plasmids were observed in the positive control lane of pET28 with T4 DNA ligase.

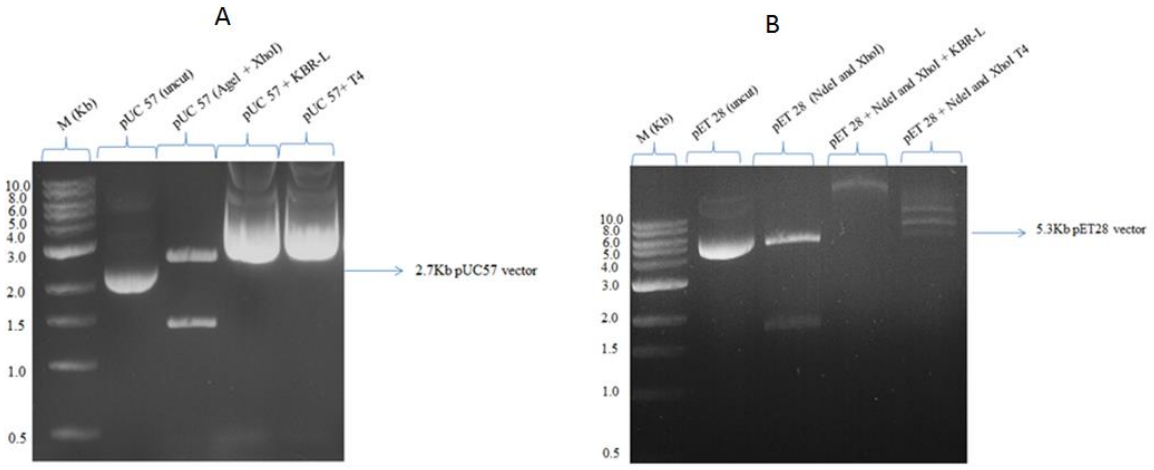


Figure 5.16: **Agarose gel analysis of DNA ligase activity assay indicating the ligation of fragments into a pUC57 vector and pET28 vectors. A: PUC57 vectors digested with *Age*I and *Xho*I and ligated with DNAlig and T4 DNA ligase. B: pET28 vector digested with *Nde*I and *Xho*I and ligated with DNAlig and T4 DNA ligase., Lane 1 = 1KB ladder, Lane 2 = vector uncut, Lane 3 = vector with an insert cut, Lane 4 = vector and an insert ligated with KBR DNAlig and lane 5 = vector and an insert ligated with commercial T4 DNA ligase.**

Competent DH5α *E. coli* cells were used for transformation with the ligation reaction and plated on Luria Broth plates with ampicillin at a concentration of 100µg/mL for pUC57 selection and kanamycin 50µg/mL for pET selection. As positive controls, uncut pUC57 and pET28 were transformed into competent DH5α *E. coli* cells. Reactions with inserts only, vector treated with phosphatase enzyme, and dH_2O instead of ligase enzyme were the negative controls. There was a large amount of growth on positive controls and no colonies observe on negative controls, indicating that the ligation reactions and subsequent transformation procedures were successful. A few background colonies were observed on vectors treated with phosphatase enzymes, but in all instances these were less than those obtained from ligation reactions (Table 5.5)

Table 5.5: **Transformation efficiency of competent DH5α *E. coli* cells with pUC57 and pET28 vectors and inserts**

Plasmids	Total size (kb)	Amount of DNA used (µg)	Transformation efficiency
pUC57 (supercoil+ control)	2.7	~0.1	5500000
pUC57_Insert (KBR ligase)	3.7	~0.1	40000
pUC57_Insert (T4 DNA ligase)	3.7	~0.1	127000
pUC57 dephosphorylated	2.7	~0.1	3000
Insert (pUC57)	1.5	~0.1	0
pET28 (supercoiled+ control)	5.3	~0.1	950000
pET28_Insert (KBR ligase)	7.3	~0.1	10000
pET28_Insert (T4 DNA ligase)	7.3	~0.1	46000
pET28 dephosphorylated	5.3	~0.1	4000
Insert (pET28)	2.0	~0.1	0

Plasmids	Total size (kb)	Amount of DNA used (µg)	Transformation efficiency
H20	N/A	N/A	0

When the products of the pUC57 + 1 kb insert and pET28 + 2kb insert ligation reaction were transformed into competent DH5α *E. coli* cells, approximately 40000CFU (pUC57) and 10000CFU (pET28) were obtained, representing candidates for transformed DH5α *E. coli* cells containing pUC57 + 1 kb insert and pET28 = 2kb insert. Several colonies were screened to identify a transformants with the correct plasmid size. Plasmid extraction was done using Geneget Plasmid Miniprep Kit and the plasmids were quantified using NanoDrop. The plasmids were digested using *Age*I and *Xho*I for pUC57 and *Nde*I and *Xho*I for pET28. Figure 5.18 shows representative picture of the digested plasmids after transformation. Growth on the vector only treated with phosphatase enzyme negative control plate of the transformations was lower than the number of colonies on the positive control and the pUC57 and pET28 with inserts transformation plates. The majority of the colonies on the transformation plates contained plasmids with inserts (see Figure 5.18).

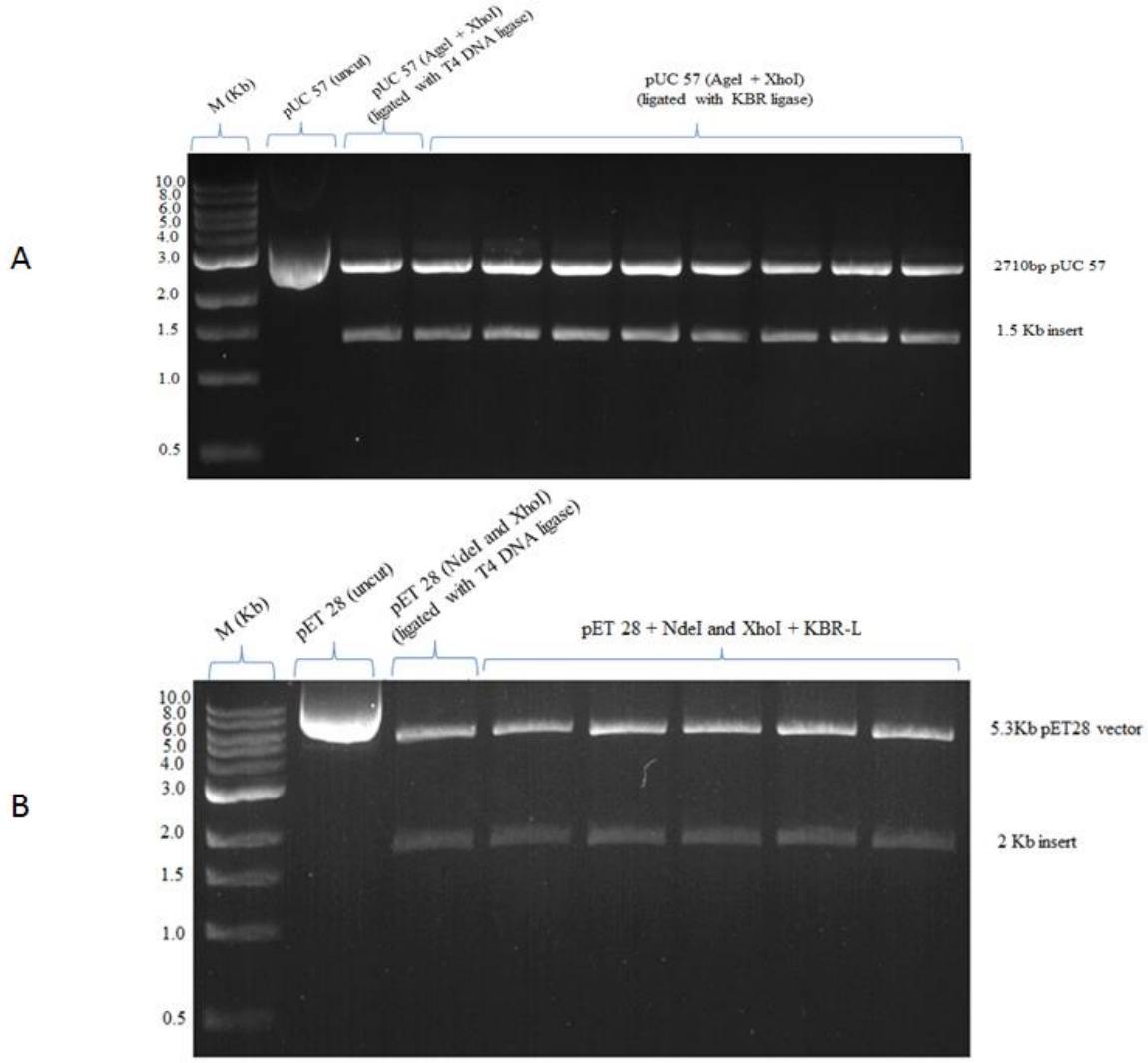


Figure 5.17: **Agarose gel electrophoresis analysis of digestions of isolated competent DH5α *E. coli* cells transformants of A: pUC57 with 1 kb insert and B: pET28 with 2kb insert. Lane 1 = 1kb marker, Lane 2 = undigested plasmids, lane 3 = positive control, Lane 4-11(pUC57) and 4-8 (pET28) = vector and insert transformants ligated with KBR DNAlig.**

5.3 Conclusions

In conclusion, 15 ORFs encoding nucleic acid manipulating enzymes have been identified using a sequence-based screening approach. The identity of their sequences to known homologues in the database ranged from 30 -97%. However, only those with less than 60% identity were subjected to expression studies in order to further explore the proteins that were likely to be novel. Functional screening using DNA polymerase complementation assay was not effective to capture novel gene sequences as the recovered ligases showed high sequence similarity to enzymes, with between 98-99% identities. Of the 9 genes selected from

sequence-based screening for recombinant protein expression, only 2 were successfully expressed in the pMAL expression system. However, the expression levels in the soluble fraction were very low and this prevented the efficient downstream purification and removal of the MBP-tag. An alternative expression strategy, *EnBase* ®, resulted in soluble DNAlig protein which was subsequently purified and qualitatively assayed for activity. Multiple sequence alignment suggests that the DNAlig was an ATP- dependent ligase. This was confirmed using a co-factor dependent DNA ligase assay; which also demonstrated that this DNAlig was able to join a vector and an insert. However, further optimisation of the ligation conditions or improved purity of the recombinant protein may be required to increase the efficiency of the DNAlig reaction since the activity is notably less than commercially available DNA ligases.

GENERAL CONCLUSION

The main goal of this project was to use modern metaviromics techniques to characterise the viral diversity of soil samples from the Cape Floral Region (specifically the Kogelberg Biosphere Reserve *fynbos* soil) and to explore these samples for novel nucleic acid manipulating enzymes. This was achieved by preparing a metagenome and metavirome library, generating a snapshot of the bacterial and viral communities, identifying and expressing novel nucleic acid manipulating enzymes derived from metavirome sequence data. This project employed a number of molecular biology techniques including the analysis of bacterial diversity using 16S rRNA gene marker and the analysis of viral diversity using sequence analysis of metavirome libraries. Functional screening of a metavirome library was also carried out using a complementation in order to identify novel DNA polymerase enzymes. In addition, this work identified several novel DNA modifying gene candidates through the analysis of metavirome nucleic acid sequences.

This study demonstrated that 16S rRNA genes can be used to identify known and unknown bacteria from Kogelberg Biosphere Reserve *fynbos* soil microbial communities. The results of this investigation are concordant with previous observations of the *fynbos* soil which reported the abundance of *Actinobacteria*, *Proteobacteria*, *Acidobacteria*, *Planctomycetes* and *Bacteroidetes* (Stafford *et al.*, 2005; Slabbert *et al.*, 2014; Miyambo *et al.*, 2016; Postma *et al.*, 2016). However, the current understanding of bacterial diversity in *fynbos* soil was based on the association of the microbial communities with *fynbos* plants. The advantages of using 16S rRNA gene techniques coupled with high-throughput NGS and bioinformatics analysis was highlighted in this study. These advantages include low cost and easy accessibility of computational databases for processing data that facilitates the identification of putative novel enzymes. Our results highlight that the sequences identified in this study are likely to represent a portion of the total diversity of bacteria in this *fynbos* soil, since extrapolation of the curves indicates that only 78.2% was sampled with the total diversity estimated to be 2328, which is consistent with other studies of different soil environments which used comparable sequencing technology and sequencing depth (Naveed *et al.*, 2016; Siles and Margesin, 2016; Terrat *et al.*, 2017).

As outlined in this study, metaviromics enabled comprehensive analysis of viral communities in the Kogelberg Biosphere Reserve *fynbos* soil sample and provided

unprecedented access to genetic diversity and genomic information of a variety of viral communities from this environment. The rarefaction analysis revealed that there is under-sampling of the *fynbos* soil bacterial and metaviromes. The rarefaction also shows an incomplete coverage of the bacterial and viral diversity in the Kogelberg Biosphere reserve soil samples, as both rarefaction graph failed to reach a rarefaction plateau. Total diversity was estimated to be 5066 for metaviromes and 2328 for bacterial diversity. Therefore, more species were detected in the metavirome than in the 16S rRNA amplicon sequencing. Majority of viruses in the Kogelberg Biosphere reserve soil metaviromes were bacteriophages which infect host bacterial cells in order to replicate. The viral abundance and activity is expected to have a direct impact on bacteria in the soil by altering the soil bacteria composition and turnover and thereby altering structure and nutrient status and contributing to soil health (Williamson *et al.*, 2005; Auguet *et al.*, 2009).

A comparison of the Kogelberg Biosphere Reserve soil metavirome with that from diverse environments generated a large amount of information on the taxonomic, phylogenetic, and functional nature of genes in a relatively facile and accessible manner. We have provided a comprehensive view of fully sequenced metaviromes. For *Caudovirales* as the major viral group detected in all metavirome samples, a higher number of Kogelberg Biosphere Reserve *fynbos* soil *Caudovirales* taxons were highlighted. These taxons indicate that *Caudovirales* represents related lineages that are distributed worldwide. The similarity of whole-community genetic pool and the clustering of specific clades strongly suggest that viral communities of Kogelberg Biosphere Reserve *fynbos* soil environments are highly stable. Thus, Kogelberg Biosphere reserve *fynbos* soil viral communities seem to form a genetically consistent community that can possibly harbour novel genes coding for nucleic acid manipulating enzymes specifically adapted to these environment. Nonetheless, future development of new technologies, improving computational methods that are robust, enhancing screening techniques, advancing the *de novo* design and selection of novel genes and sequencing cost reductions will further expand the scope of discovering novel genes using metaviromics techniques.

Coupling sequence-based screening and function-based screening provided unpreceded information about the microbial ecology, genetic diversity and detection of novel gene products from the metaviromes library sampled from the Kogelberg Biosphere Reserve *fynbos* soils that represent an under-explored habitat. However, greater emphasis will need

to be placed on new screening methodologies (both functional and sequence-based) for the successful identification of gene products. The functional annotation of genes currently lags behind the generation and deposition of metavirome sequences through high-throughput sequencing.

This work has identified 9 nucleic acid manipulating enzymes using sequence-based metavirome screening which may be novel, as indicated by low sequence identity to sequences available in the databases. These putative genes were codon-optimised for expression in an *E. coli* host. However, most of the recombinant proteins were insoluble, with only two of the recombinants expressing soluble protein using the MBP fusion (pMAL expression system). Although several conditions were tested to improve the expression, there are many possible optimisations, including changing host strains and culture conditions, optimising codon-usage and the addition of various fusion tags. The future development of novel hosts and expression systems may help to overcome these limitations. Through these studies, several novel putative DNA modifying enzymes were identified, but only one novel ATP-dependent DNA ligase was successfully expressed as an active recombinant protein in *E. coli*. Future work should focus on the quantitative characterisation of the DNA ligase activity and optimisation of the conditions to improve its effectiveness in ligating DNA molecules.

The findings from this research will contribute to our knowledge about the presence and diversity of nucleic acid manipulating enzymes in the metaviromes of soils from the Kogelberg Biosphere Reserve. The current value of the global molecular biology market for nucleic acid manipulating enzyme market provides incentives to identify novel genetic, functional and structural nucleic acid manipulating enzymes. The effective and efficient exploitation of these enzymes depends on the development of novel and innovative screening assays to facilitate the discovery of previously unknown genes. This work has clearly demonstrated the effectiveness of combining sequencing and functional screening for the identification of novel genes from the metagenomes contained in environmental samples from *fynbos* soils.

Publications and conference proceedings from this research:

International conference:

Raphela, J., Adriaenssens, E., Tsekoa, T., Rashamuse, K., and Cowan, D. (2015) Metaviromic analysis of soil in Kogelberg Biosphere Reserve. ProkaGENOMICS, European Conference on Prokaryotic and Fungal Genomics. University of Göttingen, Germany

Local conference

J Segobola, E Adriaenssens, T Tsekoa, K Rashamuse, D Cowan. (2018) The metavirome of Kogelberg Biosphere Reserve *fynbos* soil. Microbes: Livelihoods, economy and environment (SASM). Misty Hills Hotel and Conference Center, Muldersdrift, Johannesburg, South Africa.

J Segobola, E Adriaenssens, T Tsekoa, K Rashamuse, D Cowan. (2018) The metavirome of Kogelberg Biosphere Reserve *fynbos* soil: Emerging Researchers Symposium (ERS 2018) "Igniting the Next-Generation for Industrial Development". CSIR. South Africa

Publications

Segobola, J., Adriaenssens, E., Tsekoa, T., Rashamuse, K., and Cowan, D. (2018) Exploring Viral Diversity in a Unique South African Soil Habitat. *Sci. Rep.* **8**: 1–13.

REFERENCES

Ackermann, H.W. (2009) Basic phage electron microscopy. In, *Bacteriophages*. Springer, pp. 113–126.

Ackermann, H.W. (2006) Classification of bacteriophages. *The bacteriophages* **2**: 8–16.

Adriaenssens, E.M. and Cowan, D.A. (2014) Using signature genes as tools to assess environmental viral ecology and diversity. *Appl. Environ. Microbiol.* **80**: 4470–4480.

Adriaenssens, E.M., Van Zyl, L., De Maayer, P., Rubagotti, E., Rybicki, E., Tuffin, M., and Cowan, D.A. (2015) Metagenomic analysis of the viral community in Namib Desert hypoliths. *Environ. Microbiol.* **17**: 480–495.

Adriaenssens, E.M., van Zyl, L.J., Cowan, D.A., and Trindade, M.I. (2016) Metaviromics of Namib Desert Salt Pans: A Novel Lineage of Haloarchaeal Salterproviruses and a Rich Source of ssDNA Viruses. *Viruses* **8**: 14.

Adrio, J.L. and Demain, A.L. (2010) Recombinant organisms for production of industrial products. *Bioeng. Bugs* **1**: 116–131.

Allsopp, N. and Stock, W.D. (1995) Relationships between Seed Reserves, Seedling Growth and Mycorrhizal Responses in 14 Related Shrubs (Rosidae) from a Low-Nutrient Environment. *Funct. Ecol.* **9**: 248.

Alma'abadi, A.D., Gojobori, T., Mineta, K., Alma'abadi, A.D., Gojobori, T., and Mineta, K. (2015) Marine Metagenome as A Resource for Novel Enzymes. *Genomics, Proteomics Bioinforma.* **13**: 290–295.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.* **215**: 403–410.

Andraos, N., Tabor, S., and Richardson, C.C. (2004) The highly processive DNA polymerase of bacteriophage T5: Role of the unique N and C termini. *J. Biol. Chem.* **279**: 50609–50618.

Angly, F.E., Felts, B., Breitbart, M., Salamon, P., Edwards, R.A., Carlson, C., *et al*. (2006) The marine viromes of four oceanic regions. *PLoS Biol.* **4**: e368.

Angly, F.E., Willner, D., Prieto-Davó, A., Edwards, R.A., Schmieder, R., Vega-Thurber, R., *et al*. (2009) The GAAS metagenomic tool and its estimations of viral and microbial average

REFERENCES

genome size in four major biomes. *PLoS Comput. Biol.* **5**: e1000593.

Ansorge, W.J. (2009) Next-generation DNA sequencing techniques. *N. Biotechnol.* **25**: 195–203.

Arndt, D., Xia, J., Liu, Y., Zhou, Y., Guo, A.C., Cruz, J.A., *et al*. (2012) METAGENassist: A comprehensive web server for comparative metagenomics. *Nucleic Acids Res.* **40**: 1–8.

Auguet, J.C., Montanié, H., Hartmann, H.J., Lebaron, P., Casamayor, E.O., Catala, P., and Delmas, D. (2009) Potential Effect of Freshwater Virus on the Structure and Activity of Bacterial Communities in the Marennes-Oléron Bay (France). *Microb. Ecol.* **57**: 295–306.

Aziz, R.K., Bartels, D., Best, A.A., DeJongh, M., Disz, T., Edwards, R.A., *et al*. (2008) The RAST Server: Rapid Annotations using Subsystems Technology. *BMC Genomics* **9**: 75.

Aziz, R.K., Devoid, S., Disz, T., Edwards, R.A., Henry, C.S., Olsen, G.J., *et al*. (2012) SEED Servers: High-Performance Access to the SEED Genomes, Annotations, and Metabolic Models. *PLoS One* **7**: e48053.

Bag, S., Saha, B., Mehta, O., Anbumani, D., Kumar, N., Dayal, M., *et al*. (2016) An Improved Method for High Quality Metagenomics DNA Extraction from Human and Environmental Samples. *Sci. Rep.* **6**: 26775.

Bairoch, A. and Apweiler, R. (2000) The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.* **28**: 45–8.

Baltimore, D. (1971) Expression of animal virus genomes. *Bacteriol. Rev.* **35**: 235–241.

Bauer, R.J., Zhelkovsky, A., Bilotti, K., Crowell, L.E., Evans, T.C., McReynolds, L.A., and Lohman, G.J.S. (2017) Comparative analysis of the end-joining activity of several DNA ligases. *PLoS One* **12**: e0190062.

Bej, A.K., Mahbubani, M.H., Miller, R., DiCesare, J.L., Haff, L., and Atlas, R.M. (1990) Multiplex PCR amplification and immobilized capture probes for detection of bacterial pathogens and indicators in water. *Mol. Cell. Probes* **4**: 353–365.

Béjà, O. (2004) To BAC or not to BAC: Marine ecogenomics. *Curr. Opin. Biotechnol.* **15**:187–190

Béjà, O., Suzuki, M.T., Heidelberg, J.F., Nelson, W.C., Preston, C.M., Hamada, T., *et al*. (2002) Unsuspected diversity among marine aerobic anoxygenic phototrophs. *Nature* **415**:

REFERENCES

630–633.

Bertani, G. (1953) Lysogenic versus lytic cycle of phage multiplication. *Cold Spring Harb. Symp. Quant. Biol.* **18**: 65–70.

Blondal, T., Hjorleifsdottir, S.H., Fridjonsson, O.F., Ævarsson, A., Skirnisdottir, S., Hermannsdottir, A.G., Hreggvidsson, G.O., Smith, A.V., and Kristjansson, J.K. (2003) Discovery and characterization of a thermostable bacteriophage RNA ligase homologous to T4 RNA ligase 1. *Nucleic Acids Res.* **31**: 7247–7254.

Bobrova, O., Kristoffersen, J.B., Oulas, A., and Ivanytsia, V. (2016) Metagenomic 16s rrna investigation of microbial communities in the Black Sea estuaries in South-West of Ukraine. *Acta Biochim. Pol.* **63**.2015-1145

Booysen, D. and Booysen, D. (2011) Characterisation of a DNA ligase from an Antarctic metagenomic library. *Department of Biotechnology, University of the Western Cape Bellville* (Thesis)

Borneman, J., Skroch, P.W., O'Sullivan, K.M., Palus, J.A., Rumjanek, N.G., Jansen, J.L., *et al*. (1996) Molecular microbial diversity of an agricultural soil in Wisconsin. *Appl. Environ. Microbiol.* **62**: 1935–43.

Boutet, E., Lieberherr, D., Tognolli, M., Schneider, M., and Bairoch, A. (2007) UniProtKB/Swiss-Prot. *Methods Mol. Biol.* **406**: 89–112.

Breitbart, M., Felts, B., Kelley, S., Mahaffy, J.M., Nulton, J., Salamon, P., and Rohwer, F. (2004) Diversity and population structure of a near–shore marine–sediment viral community. *Proc. R. Soc. London.Series B Biol. Sci.* **271**: 565–574.

Breitbart, M., Miyake, J.H., and Rohwer, F. (2004) Global distribution of nearly identical phage-encoded DNA sequences. *FEMS Microbiol. Lett.* **236**: 249–56.

Breitbart, M. and Rohwer, F. (2005) Here a virus, there a virus, everywhere the same virus? *Trends Microbiol.* **13**: 278–284.

Breitbart, M., Salamon, P., Andresen, B., Mahaffy, J.M., Segall, A.M., Mead, D., *et al*. (2002) Genomic analysis of uncultured marine viral communities. *Proc. Natl. Acad. Sci. U. S. A.* **99**: 14250–5.

Brown, T.A. (2016) Gene Cloning and DNA Analysis: An Introduction. *John Wiley Sons.*

REFERENCES

Brum, J.R. and Sullivan, M.B. (2015) Rising to the challenge: accelerated pace of discovery transforms marine virology. *Nat. Rev. Microbiol.* **13**: 147–159.

Brussaard, C.P.D. (2004) Optimization of procedures for counting viruses by flow cytometry. *Appl. Environ. Microbiol.* **70**: 1506–13.

Buermans, H.P.J. and den Dunnen, J.T. (2014) Next generation sequencing technology: Advances and applications. *Biochim. Biophys. Acta - Mol. Basis Dis.* **1842**: 1932–1941.

Burgess-Brown, N.A., Sharma, S., Sobott, F., Loenarz, C., Oppermann, U., and Gileadi, O. (2008) Codon optimization can improve expression of human genes in Escherichia coli: A multi-gene study. *Protein Expr. Purif.* **59**: 94–102.

Cancilla, M.R., Powell, I.B., Hillier, A.J., and Davidson, B.E. (1992) Rapid genomic fingerprinting of Lactococcus lactis strains by arbitrarily primed polymerase chain reaction with 32P and fluorescent labels. *Appl. Environ. Microbiol.* **58**: 1772–5.

Caporaso, J.G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F.D., Costello, E.K., *et al.* (2010) QIIME allows analysis of high-throughput community sequencing data. *Nat. Methods* **7**: 335–336.

Caporaso, J.G., Lauber, C.L., Walters, W. A, Berg-Lyons, D., Huntley, J., Fierer, N., *et al.* (2012) Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. *ISME J.* **6**: 1621–1624.

Caravaca, F., Barea, J.., and Roldán, A. (2002) Synergistic influence of an arbuscular mycorrhizal fungus and organic amendment on Pistacia lentiscus L. seedlings afforested in a degraded semiarid soil. *Soil Biol. Biochem.* **34**: 1139–1145.

Carrigg, C., Rice, O., Kavanagh, S., Collins, G., and O'Flaherty, V. (2007) DNA extraction method affects microbial community profiles from soils and sediment. *Appl. Microbiol. Biotechnol.* **77**: 955–964.

Casas, V. and Rohwer, F. (2007) Phage metagenomics. *Methods Enzymol.* **421**: 259–268.

Case, R.J., Boucher, Y., Dahllöf, I., Holmström, C., Doolittle, W.F., and Kjelleberg, S. (2007) Use of 16S rRNA and rpoB genes as molecular markers for microbial ecology studies. *Appl. Environ. Microbiol.* **73**: 278–88.

Centeno, C.M., Legendre, P., Beltrán, Y., Alcántara-Hernández, R.J., Lidström, U.E., Ashby,

REFERENCES

M.N., *et al*. (2012) Microbialite genetic diversity and composition relate to environmental variables. *FEMS Microbiol. Ecol.* **82**: 724–735.

Cheng, J., Romantsov, T., Engel, K., Doxey, A.C., Rose, D.R., Neufeld, J.D., and Charles, T.C. (2017) Functional metagenomics reveals novel β-galactosidases not predictable from gene sequences. *PLoS One* **12**: e0172545.

Choi, K.H. (2012) Viral Polymerases. *Adv Exp Med Biol* **726**: 267–304.

Cole, J.R., Wang, Q., Cardenas, E., Fish, J., Chai, B., Farris, R.J., *et al.* (2009) The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. *Nucleic Acids Res.* **37**: D141–D145.

Cowling, R., Pressey, R., Rouget, M., and Lombard, A. (2003) A conservation plan for a global biodiversity hotspot—the Cape Floristic Region, South Africa. *Biol. Conserv.* **112**: 191–216.

Cowling, R. (1992) The ecology of *fynbos*. Nutrients, fire and diversity. *Town. Oxford Univ.* **41**.

Culley, A.I., Lang, A.S., and Suttle, C.A. (2006) Metagenomic analysis of coastal RNA virus communities. *Science (80-.).* **312**: 1795–1798.

Culligan, E.P., Marchesi, J.R., Hill, C., and Sleator, R.D. (2014) Combined metagenomic and phenomic approaches identify a novel salt tolerance gene from the human gut microbiome. *Front. Microbiol.* **5**: 189.

Culligan, E.P., Sleator, R.D., Marchesi, J.R., and Hill, C. (2014) Metagenomics and novel gene discovery: promise and potential for novel therapeutics. *Virulence* **5**: 399–412.

D'Argenio, V., Casaburi, G., Precone, V., and Salvatore, F. (2014) Comparative metagenomic analysis of human gut microbiome composition using two different bioinformatic pipelines. *Biomed Res. Int.* **2014**:.

DeCastro, M.-E., Rodríguez-Belmonte, E., and González-Siso, M.-I. (2016) Metagenomics of Thermophiles with a Focus on Discovery of Novel Thermozymes. *Front. Microbiol.* **7**: 1521.

Delmont, T.O., Robe, P., Cecillon, S., Clark, I.M., Constancias, F., Simonet, P., *et al*. (2011) Accessing the soil metagenome for studies of microbial diversity. *Appl. Environ.*

REFERENCES

Microbiol. **77**: 1315–24.

Delwart, E.L. (2007) Viral metagenomics. *Rev. Med. Virol.* **17**: 115–131.

Dinsdale, E. a, Edwards, R. a, Hall, D., Angly, F., Breitbart, M., Brulc, J.M., *et al.* (2008) Functional metagenomic profiling of nine biomes. *Nature* **452**: 629–632.

Dinsdale, E.A., Pantos, O., Smriga, S., Edwards, R.A., Angly, F., Wegley, L., *et al.* (2008) Microbial ecology of four coral atolls in the Northern Line Islands. *PLoS One* **3**:.

Doherty, A.J. and Suh, S.W. (2000) Structural and mechanistic conservation in DNA ligases. *Nucleic Acids Res.* **28**: 4051–4058.

Doolittle, W.F. (1988) Bacterial evolution. *Can. J. Microbiol.* **34**: 547–551.

Dorigo, U., Jacquet, S., and Humbert, J.-F. (2004) Cyanophage diversity, inferred from g20 gene analyses, in the largest natural lake in France, Lake Bourget. *Appl. Environ. Microbiol.* **70**: 1017–22.

Duhaime, M.B., Deng, L., Poulos, B.T., Sullivan, M.B., Crowgey, E., Polson, S.W., *et al.* (2012) Towards quantitative metagenomics of wild viruses and other ultra-low concentration DNA samples: a rigorous assessment and optimization of the linker amplification method. *Environ. Microbiol.* **14**: 2526–2537.

Dziewit, L., Oscik, K., Bartosik, D., and Radlinska, M. (2014) Molecular characterization of a novel temperate sinorhizobium bacteriophage, ΦLM21, encoding DNA methyltransferase with CcrM-like specificity. *J. Virol.* **88**: 13111–24.

Dzikiti, S., Bugan, R., and Israel, S. (2014) Measurement and modelling of evapotranspiration in three *fynbos* vegetation types.*Water SA.* **40**: 189–198.

Easton, S. (2009) Functional and Metagenomic Analysis of the Human Tongue Dorsum using Phage Display. *University College London, Department of Structural and Molecular Biology*. Thesis

Edwards, R.A. and Rohwer, F. (2005) Viral metagenomics. *Nat. Rev. Microbiol.* **3**: 504–510.

Errol, C. (2003) DNA damage and repair. *Nature* **421**: 436–440.

Ewing, B., Ewing, B., Hillier, L., Hillier, L., Wendl, M.C., Wendl, M.C., *et al.* (2005) Base-Calling of Automated Sequencer Traces Using. *Genome Res.* 175–185.

REFERENCES

Fahnert, B., Lilie, H., and Neubauer, P. (2004) Inclusion Bodies: Formation and Utilisation. *Adv Biochem Engin/Biotechnol* **89**: 93–142.

Fancello, L., Trape, S., Robert, C., Boyer, M., Popgeorgiev, N., Raoult, D., and Desnues, C. (2013) Viruses in the desert: a metagenomic survey of viral communities in four perennial ponds of the Mauritanian Sahara. *ISME J.* **7**: 359–69.

Fantle, M.S., Dittel, A.I., Schwalm, S.M., Epifanio, C.E., Fogel, M.L., Etherington, L.L., *et al.* (2003) Sequence-Based Metagenomic Analysis. *J. Exp. Mar. Bio. Ecol.* **72**: 179–198.

Farrelly, V., Rainey, F.A., and Stackebrandt, E. (1995) Effect of genome size and rrn gene copy number on PCR amplification of 16S rRNA genes from a mixture of bacterial species. *Appl. Environ. Microbiol.* **61**: 2798–801.

Felczykowska, A., Krajewska, A., Zielińska, S., Łoś, J.M., Bloch, S.K., and Nejman-Faleńczyk, B. (2015) The most widespread problems in the function-based microbial metagenomics. *Acta Biochim. Pol.* **62**: 161–166.

Fernández-Álvaro, E., Kourist, R., Winter, J., Böttcher, D., Liebeton, K., Naumer, C., *et al.* (2010) Enantioselective kinetic resolution of phenylalkyl carboxylic acids using metagenome-derived esterases. *Microb. Biotechnol.* **3**: 59–64.

Ferrer, M., Beloqui, A., Timmis, K.N., and Golyshin, P.N. (2009) Metagenomics for mining new genetic resources of microbial communities. *J. Mol. Microbiol. Biotechnol.* **16**: 109–23.

Ferrer, M., Golyshina, O. V., Chernikova, T.N., Khachane, A.N., Dos Santos, V.A.P.M., Yakimov, M.M., *et al*. (2005) Microbial enzymes mined from the Urania deep-sea hypersaline anoxic basin. *Chem. Biol.* **12**: 895–904.

Ferrer, M., Martínez-martínez, M., Bargiela, R., Streit, W.R., Golyshina, O. V., and Golyshin, P.N. (2015) Minireview Estimating the success of enzyme bioprospecting through metagenomics : current status and future trends. **9**: 22–34.

Fierer, N., Breitbart, M., Nulton, J., Salamon, P., Lozupone, C., Jones, R., *et al*. (2007) Metagenomic and small-subunit rRNA analyses reveal the genetic diversity of bacteria, archaea, fungi, and viruses in soil. *Appl. Environ. Microbiol.* **73**: 7059–7066.

Fox, K.R. (1988) DNAase I footprinting of restriction enzymes. *Biochem. Biophys. Res. Commun.* **155**: 779–785.

REFERENCES

Frostegård, A., Courtois, S., Ramisse, V., Clerc, S., Bernillon, D., Le Gall, F., *et al*. (1999) Quantification of bias related to the extraction of DNA directly from soils. *Appl. Environ. Microbiol.* **65**: 5409–20.

Fuhrman, J.A. (1999) Marine viruses and their biogeochemical and ecological effects. *Nature* **399**: 541–548.

Fuhrman, J.A. and Campbell, L. (1998) Marine ecology: microbial microdiversity. *Nature* **393**: 410–411.

Fukuyo, M., Nakano, T., Zhang, Y., Furuta, Y., Ishikawa, K., Watanabe-Matsui, M., *et al*. (2015) Restriction-modification system with methyl-inhibited base excision and abasic-site cleavage activities. *Nucleic Acids Res.* **43**: 2841–52.

Gasol, J.M. and Moran, X.A.G. (2015) Flow Cytometric Determination of Microbial Abundances and Its Use to Obtain Indices of Community Structure and Relative Activity. *Springer Protoc. Handbooks* 1–29.

Georgiou, G. and Segatori, L. (2005) Preparative expression of secreted proteins in bacteria: Status report and future prospects. *Curr. Opin. Biotechnol.* **16**: 538–545.

Georlette, D., Jónsson, Z.O., Van Petegem, F., Chessa, J.P., Van Beeumen, J., Hübscher, U., and Gerday, C. (2000) A DNA ligase from the psychrophile Pseudoalteromonas haloplanktis gives insights into the adaptation of proteins to low temperatures. *Eur. J. Biochem.* **267**: 3502–3512.

Ghosh, T.S., Mohammed, M.H., Komanduri, D., and Mande, S.S. (2011) ProViDE: A software tool for accurate estimation of viral diversity in metagenomic samples. *Bioinformation* **6**: 91–94.

Gillespie, D. E., Rondon, M. R., Williamson, L. L., & Handelsman, J. (2005) Metagenomic libraries from uncultured microorganisms. In, *Molecular microbial ecology, Routledge. Taylor and Francis Group, Florence*., pp. 261–280.

Giordano, F., Aigrain, L., Quail, M.A., Coupland, P., Bonfield, J.K., Davies, R.M., *et al*. (2017) De novo yeast genome assemblies from MinION, PacBio and MiSeq platforms. *Sci. Rep.* **7**: 3935.

Goff, S.P. (2004) Genetic reprogramming by retroviruses: Enhanced suppression of translational termination. *Cell Cycle* **3**: 123–125.

REFERENCES

Goldblatt, P. (1997) Floristic diversity in the Cape Flora of South Africa. *Biodivers. Conserv.* **6**: 359–377.

Goldstein, M.A. and Doi, R.H. (1995) Biotechnology annual review. In, *Biotechnology annual review.*, pp. 105–128.

Gonçalves, A.C.S., Dos Santos, A.C.F., Dos Santos, T.F., Pessoa, T.B.A., Dias, J.C.T., and Rezende, R.P. (2015) High yield of functional metagenomic library from mangroves constructed in fosmid vector. *Genet. Mol. Res.* **14**: 11841–11847.

Gottesman, S. (1996) Proteases and their targets in escherichia coli. *Annu. Rev. Genet.* **30**: 465–506.

Gräslund, S., Nordlund, P., Weigelt, J., Hallberg, B.M., Bray, J., Gileadi, O., *et al*. (2008) Protein production and purification. *Nat. Methods* **5**: 135–146.

Griffiths, R.I., Thomson, B.C., James, P., Bell, T., Bailey, M., and Whiteley, A.S. (2011) The bacterial biogeography of British soils. *Environ. Microbiol.* **13**: 1642–1654.

Grisshammer, R. and Tate, C.G. (1995) Overexpression of integral membrane proteins for structural studies. *Q. Rev. Biophys.* **28**: 315–422.

Guazzaroni, M.-E., Silva-Rocha, R., and Ward, R.J. (2015) Synthetic biology approaches to improve biocatalyst identification in metagenomic library screening. *Microb. Biotechnol.* **8**: 52–64.

Gustafsson, C., Govindarajan, S., and Minshull, J. (2004) Codon bias and heterologous protein expression. *Trends Biotechnol.* **22**: 346–353.

Haber, C. and Wirtz, D. (2000) Shear-Induced Assembly of λ-Phage DNA. *Biophys. J.* **79**: 1530–1536.

Hall, R.J., Wang, J., Todd, A.K., Bissielo, A.B., Yen, S., Strydom, H., *et al*. (2014) Evaluation of rapid and simple techniques for the enrichment of viruses prior to metagenomic virus discovery. *J. Virol. Methods* **195**: 194–204.

Hamilton, S.C., Joseph, W., Davis, M.C., Biotech, A.P., Farchaus, J.W., and Davis, M.C. (2001) DNA polymerases as engines for biotechnology. *Biotechniques* **31**: 370–383.

Handelsman, J. (2004) Metagenomics: application of genomics to uncultured microorganisms. *Microbiol. Mol. Biol. Rev.* **68**: 669–685.

REFERENCES

Hannig, G. and Makrides, S.C. (1998) Strategies for optimizing heterologous protein expression in Escherichia coli. *Trends Biotechnol.* **16**: 54–60.

Haq, I.U., Chaudhry, W.N., Akhtar, M.N., Andleeb, S., and Qadri, I. (2012) Bacteriophages and their implications on future biotechnology: a review. *Virol J* **9**:.

Harmsen, H.J.M., Raangs, G.C., Franks, A.H., Wildeboer-Veloo, A.C.M., and Welling, G.W. (2002) The Effect of the Prebiotic Inulin and the Probiotic Bifidobacterium longum on the Fecal Microflora of Healthy Volunteers Measured by FISH and DGGE. *Microb. Ecol. Health Dis.* **14**: 212–220.

Hastings, R.C., Butler, C., Singleton, I., Saunders, J.R., and McCarthy, A.J. (2000) Analysis of ammonia-oxidizing bacteria populations in acid forest soil during conditions of moisture limitation. *Lett. Appl. Microbiol.* **30**: 14–18.

Haynes, R.J. and Swift, R.S. (1989) Effect of rewetting air-dried soils on pH and accumulation of mineral nitrogen. *J. Soil Sci.* **40**: 341–347.

Head, I.M., Saunders, J.R., and Pickup, R.W. (1998) Microbial Evolution, Diversity, and Ecology: A Decade of Ribosomal RNA Analysis of Uncultivated Microorganisms. *Microb. Ecol.* **35**: 1–21.

Hendrix, R.W., Smith, M.C., Burns, R.N., Ford, M.E., and Hatfull, G.F. (1999) Evolutionary relationships among diverse bacteriophages and prophages: all the world's a phage. *Proc. Natl. Acad. Sci. U. S. A.* **96**: 2192–7.

Holovachov, O., Haenel, Q., Bourlat, S.J., and Jondelius, U. (2017) Taxonomy assignment approach determines the efficiency of identification of OTUs in marine nematodes. *R. Soc. Open Sci.* **4**:.

Hugenholtz, P., Staley, J., Konopka, A., Galvez, A., Maqueda, M., Martinez-Bueno, M., *et al.* (2002) Exploring prokaryotic diversity in the genomic era. *Genome Biol.* **3**: 3.1.

Hulo, C., De Castro, E., Masson, P., Bougueleret, L., Bairoch, A., Xenarios, I., and Le Mercier, P. (2011) ViralZone: A knowledge resource to understand virus diversity. *Nucleic Acids Res.* **39**: 576–582.

Hurwitz, B.L., Deng, L., Poulos, B.T., and Sullivan, M.B. (2013) Evaluation of methods to concentrate and purify ocean virus communities through comparative, replicated metagenomics. *Environ. Microbiol.* **15**: 1428–1440.

REFERENCES

Huse, S.M., Mark Welch, D.B., Voorhis, A., Shipunova, A., Morrison, H.G., Eren, A.M., and Sogin, M.L. (2014) VAMPS: a website for visualization and analysis of microbial population structures. *BMC Bioinformatics* **15**: 41.

Huson, D.H., Auch, A.F., Qi, J., and Schuster, S.C. (2007) MEGAN analysis of metagenomic data. *Genome Res.* **17**: 377–86.

Joseph, B.C., Pichaimuthu, S., and Srimeenakshi, S. (2015) An Overview of the Parameters for Recombinant Protein Expression in Escherichia coli. *J. Cell Sci. Ther.* **06**:.

Ju, F. and Zhang, T. (2015) 16S rRNA gene high-throughput sequencing data mining of microbial diversity and interactions. *Appl. Microbiol. Biotechnol.* **99**: 4119–4129.

Kaiser, K., Wemheuer, B., Korolkow, V., Wemheuer, F., Nacke, H., Schöning, I., *et al*. (2016) Driving forces of soil bacterial community structure, diversity, and function in temperate grasslands and forests. *Sci. Rep.* **6**: 33696.

Kanehisa, M., Goto, S., Sato, Y., Furumichi, M., and Tanabe, M. (2012) KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res.* **40**: D109–D114.

Kapust, R.B. and Waugh, D.S. (1999) Escherichia coli maltose-binding protein is uncommonly effective at promoting the solubility of polypeptides to which it is fused. *The Protein Society.* **8***:1668–1674.*

Keegan, K.P., Glass, E.M., and Meyer, F. (2016) MG-RAST, a Metagenomics Service for Analysis of Microbial Community Structure and Function. *Methods Mol. Biol.* **1399**. 207–233.

Keeley, J.E. (2013) A Comparative Overview of Fire and Flora in Mediterranean Climate Ecosystems. *Israel Journal of Ecology & Evolution***, 58:***2-3, 123-135*

Kennedy, J., Marchesi, J.R., and Dobson, A.D.W. (2008) Marine metagenomics: strategies for the discovery of novel enzymes with biotechnological applications from marine environments. *Microb Cell Fact* **7**:2

Khalili, M., Soleyman, M.R., Baazm, M., and Beyer, C. (2015) High-level expression and purification of soluble bioactive recombinant human heparin-binding epidermal growth factor in Escherichia coli. *Cell Biol. Int.* **39**: 858–864.

Kim, J.C., Lee, S.W., Won, K., Lim, H.K., Kim, J.C., Choi, G.J., and Cho, K.Y. (2004)

REFERENCES

Screening for novel lipolytic enzymes from uncultured soil microorganisms. *Appl. Microbiol. Biotechnol.* **65**: 720–726.

Kim, K.H. and Bae, J.W. (2011) Amplification methods bias metagenomic libraries of uncultured single-stranded and double-stranded DNA viruses. *Appl. Environ. Microbiol.* **77**: 7663–8.

Kim, M.S., Park, E.J., Roh, S.W., and Bae, J.W. (2011) Diversity and Abundance of Single-Stranded DNA Viruses in Human Feces. *Appl. Environ. Microbiol.* **77**: 8062–8070.

Knietsch, A., Bowien, S., Whited, G., Gottschalk, G., and Daniel, R. (2003) Identification and characterization of coenzyme B12-dependent glycerol dehydratase- and diol dehydratase-encoding genes from metagenomic DNA libraries derived from enrichment cultures. *Appl. Environ. Microbiol.* **69**: 3048–60.

Kodzius, R. and Gojobori, T. (2015) Marine metagenomics as a source for bioprospecting. *Mar. Genomics* **24**: 21–30.

Koonin, E. V. and Dolja, V. V. (2018) Metaviromics: a tectonic shift in understanding virus evolution. *Virus Res.* **246**: A1–A3.

Koonin, E. V, Senkevich, T.G., and Dolja, V. V (2006) The ancient Virus World and evolution of cells. *Biol. Direct* **1**: 29.

Krause, M., Neubauer, A., and Neubauer, P. (2016) The fed-batch principle for the molecular biology lab: controlled nutrient diets in ready-made media improve production of recombinant proteins in Escherichia coli. *Microb. Cell Fact.* **15**: 110.

Lako, J. (2005) Analysis of ammonia-oxidizing bacteria associated with the roots of Proteaceae plant species in soils of *fynbos* ecosystem. *Department of Biotechnology, University of the Western Cape.**2005***

Lam, E.T., Hastie, A., Lin, C., Ehrlich, D., Das, S.K., Austin, M.D., *et al*. (2012) Genome mapping on nanochannel arrays for structural variation analysis and sequence assembly. *Nat. Biotechnol.* **30**: 771–6.

Lam, K.N., Cheng, J., Engel, K., Neufeld, J.D., and Charles, T.C. (2015) Current and future resources for functional metagenomics. *Front. Microbiol.* **6**: 1196.

LaMontagne, M.G., Michel, F.C., Holden, P.A., and Reddy, C.A. (2002) Evaluation of

REFERENCES

extraction and purification methods for obtaining PCR-amplifiable DNA from compost for microbial community analysis. *J. Microbiol. Methods* **49**: 255–264.

Lauritzen, C., Tüchsen, E., Hansen, P.E., and Skovgaard, O. (1991) BPTI and N-terminal extended analogues generated by factor Xa cleavage and cathepsin C trimming of a fusion protein expressed in Escherichia coli. *Protein Expr. Purif.* **2**: 372–378.

Lee, D.E., Lee, J., Kim, Y.M., Myeong, J.I., and Kim, K.H. (2016) Uncultured bacterial diversity in a seawater recirculating aquaculture system revealed by 16S rRNA gene amplicon sequencing. *J. Microbiol.* **54**: 296–304.

Leemhuis, H., Kelly, R.M., and Dijkhuizen, L. (2009) Directed evolution of enzymes: Library screening strategies. *IUBMB Life* **61**: 222–228.

Li, S., Yang, X., Yang, S., Zhu, M., and Wang, X. (2012) Technology prospecting on enzymes: Application, marketing and engineering. *Comput. Struct. Biotechnol. J.* **2**:.

Littlechild, J. (2016) Novel enzymes from metagenomics. *N. Biotechnol.* **33**: S61–S62.

Liu, L., Li, Y., Li, S., Hu, N., He, Y., Pong, R., *et al*. (2012) Comparison of next-generation sequencing systems. *Biomed Res. Int.* **2012**:.

Loenen, W. a M., Dryden, D.T.F., Raleigh, E. a., Wilson, G.G., and Murrayy, N.E. (2014) Highlights of the DNA cutters: A short history of the restriction enzymes. *Nucleic Acids Res.* **42**: 3–19.

Lopes, A., Amarir-Bouhram, J., Faure, G., Petit, M.-A., and Guerois, R. (2010) Detection of novel recombinases in bacteriophage genomes unveils Rad52, Rad51 and Gp2.5 remote homologs. *Nucleic Acids Res.* **38**: 3952–3962.

Lorenz, P., Liebeton, K., Niehaus, F., and Eck, J. (2002) Screening for novel enzymes for biocatalytic processes: accessing the metagenome as a resource of novel functional sequence space. *Curr. Opin. Biotechnol.* **13**: 572–577.

Louten, J. (2016) Essential Human Virology. *Elsevier*. p 19-29

Lu, J., Santo Domingo, J., and Shanks, O.C. (2007) Identification of chicken-specific fecal microbial sequences using a metagenomic approach. *Water Res.* **41**: 3561–3574.

Madhavan, A. and Sindhu, R. (2017) Metagenome Analysis : a Powerful Tool for Enzyme Bioprospecting. *Appl Biochem Biotechnol.* **183**. 636–651.

REFERENCES

Makhalanyane, T.P., Valverde, A., Lacap, D.C., Pointing, S.B., Tuffin, M.I., and Cowan, D.A. (2013) Evidence of species recruitment and development of hot desert hypolithic communities. *Environ. Microbiol. Rep.* **5**: 219–224.

Makrides, S.C. (1996) Strategies for achieving high-level expression of genes in Escherichia coli. *Microbiol. Rev.* **60**: 512–38.

Mann, N.H. (2005) The third age of phage. *PLoS Biol.* **3**: e182.

Marchetti, C., Walker, S.A., Odreman, F., Vindigni, A., Doherty, A.J., and Jeggo, P. (2006) Identification of a novel motif in DNA ligases exemplified by DNA ligase IV. *DNA Repair (Amst).* **5**: 788–798.

Margulies, M., Egholm, M., Altman, W.E., Attiya, S., Bader, J.S., Bemben, L.A., *et al*. (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**: 376.

Marine, R., McCarren, C., Vorrasane, V., Nasko, D., Crowgey, E., Polson, S.W., and Wommack, K. (2014) Caught in the middle with multiple displacement amplification: the myth of pooling for avoiding multiple displacement amplification bias in a metagenome. *Microbiome* **2**: 3.

Martin-Laurent, F., Philippot, L., Hallet, S., Chaussod, R., Germon, J.C., Soulas, G., and Catroux, G. (2001) DNA extraction from soils: old bias for new microbial diversity analysis methods. *Appl. Environ. Microbiol.* **67**: 2354–9.

Martin, I. V and MacNeill, S.A. (2002) ATP-dependent DNA ligases. *Genome Biol.* **3**: 3.5.

Martínez-Martínez, M., Bargiela, R., and Ferrer, M. (2016) Metagenomics and the Search for Industrial Enzymes. *Elsevier*. 167–184.

Mathur, E.. (1996) Purified thermostable Pyrococcus furiosus DNA polymerase that migrates on a non-denaturing polyacrylamide gel faster than phosphorylase B and Taq polymerase and more slowly. *U.S. Pat.* **5**:545,552.

McDonald, D., Price, M.N., Goodrich, J., Nawrocki, E.P., DeSantis, T.Z., Probst, A., *et al*. (2012) An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *ISME J.* **6**: 610–618.

McDonald, D.J., Juritz, J.M., Cowling, R.M., and Knottenbelt, W.J. (1995) Modelling the biological aspects of local endemism in South African *fynbos*. *Plant Syst. Evol.* **195**: 137–

147.

Merseguel, K.B., Nishikaku, A.S., Rodrigues, A.M., Padovan, A.C., e Ferreira, R.C., de Azevedo Melo, A.S., *et al*. (2015) Genetic diversity of medically important and emerging Candida species causing invasive infection. *BMC Infect. Dis.* **15**: 57.

Meyer, F., Paarmann, D., D'Souza, M., Olson, R., Glass, E.M., Kubal, M., *et al*. (2008) The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics* **9**: 386.

Mierendorf, R.C., Morris, B.B., Hammer, B., and Novy, R.E. (1998) Expression and Purification of Recombinant Proteins Using the pET System. *Humana Press, New Jersey*. 257–292.

Miyambo, T., Makhalanyane, T.P., Cowan, D.A., and Valverde, A. (2016) Plants of the *fynbos* biome harbour host species-specific bacterial communities. *FEMS Microbiol. Lett.* **363**:15

Miyashita, N.T., Iwanaga, H., Charles, S., Diway, B., Sabang, J., and Chong, L. (2013) Soil bacterial community structure in five tropical forests in Malaysia and one temperate forest in Japan revealed by pyrosequencing analyses of 16S rRNA gene sequence variation. *Genes Genet. Syst.* **88**: 93–103.

Mizuno, C.M., Ghai, R., Saghaï, A., López-García, P., and Rodriguez-Valera, F. (2016) Genomes of Abundant and Widespread Viruses from the Deep Ocean. *MBio* **7**: 805-16.

Moffatt, B.A. and Studier, F.W. (1987) T7 lysozyme inhibits transcription by T7 RNA polymerase. *Cell.* **49**: 221–227.

Mokili, J.L., Rohwer, F., and Dutilh, B.E. (2012) Metagenomics and future perspectives in virus discovery. *Curr. Opin. Virol.* **2**: 63–77.

Moroenyane, I., Chimphango, S.B.M., Wang, J., Kim, H.-K., and Adams, J.M. (2016) Deterministic assembly processes govern bacterial community structure in the *fynbos*, South Africa. *Microb. Ecol.* **72**: 313–323.

Moser, M.J., DiFrancesco, R.A., Gowda, K., Klingele, A.J., Sugar, D.R., Stocki, S., *et al*. (2012) Thermostable DNA polymerase from a viral metagenome is a potent rt-PCR enzyme. *PLoS One* **7**: e38371.

Mucina, L. and Rutherford, M.C. (2006) The Vegetation of South Africa, Lesotho and

REFERENCES

Swaziland Vegetation Map of the Mitchell Plateau Region View project Great Escarpment Biodiversity Research Programme View project. *Plant and Soil.* 1-2; p1-23

Muller, J., Szklarczyk, D., Julien, P., Letunic, I., Roth, A., Kuhn, M., *et al*. (2010) eggNOG v2.0: extending the evolutionary genealogy of genes with enhanced non-supervised orthologous groups, species and functional annotations. *Nucleic Acids Res.* **38**: D190-5.

Murray, N.E., Bruce, S.A., and Murray, K. (1979) Molecular cloning of the DNA ligase gene from bacteriophage T4. *J. Mol. Biol.* **132**: 493–505.

Myers, N., Mittermeier, R.A., Mittermeier, C.G., da Fonseca, G.A.B., and Kent, J. (2000) Biodiversity hotspots for conservation priorities. *Nature.* **403**: 853–858.

Nagano, K., Wachi, M., Takada, A., Takaku, F., Hirasawa, T., and Nagai, K. (1999) fcsA29 mutation is an allele of polA gene of Escherichia coli. *Biosci. Biotechnol. Biochem.* **63**: 427–429.

Naveed, M., Herath, L., Moldrup, P., Arthur, E., Nicolaisen, M., Norgaard, T., *et al*. (2016) Spatial variability of microbial richness and diversity and relationships with soil organic carbon, texture and structure across an agricultural field. *Appl. Soil Ecol.* **103**: 44–55.

Niehaus, F., Bertoldo, C., Kähler, M., and Antranikian, G. (1999) Extremophiles as a source of novel enzymes for industrial application. *Appl. Microbiol. Biotechnol.* **51**: 711–729.

Norgard, M. V., Keem, K., and Monahan, J.J. (1978) Factors affecting the transformation of Escherichia coli strain χ1776 by pBR322 plasmid DNA. *Gene* **3**: 279–292.

Oksanen J, Blanchet FG, Kindt R, L.P. and others (2016) 'vegan' 2.3.4.4— community ecology package. http: //CRAN.R-project.org/package=vegan.

Osborn, M. (2005) Molecular microbial ecology. *Garland Science.*

Overbeek, R., Begley, T., Butler, R.M., Choudhuri, J. V, Chuang, H.-Y., Cohoon, M., *et al*. (2005) The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic Acids Res.* **33**: 5691–702.

Pace, N.R., Stahl, D.A., Lane, D.J., and Olsen, G.J. (1986) The Analysis of Natural Microbial Populations by Ribosomal RNA Sequences. *Springer, Boston, MA*, pp. 1–55.

Park, J.E., Lee, B.-T., Kim, B.-Y., and Son, A. (2018) Bacterial community analysis of stabilized soils in proximity to an exhausted mine. *Environ. Eng. Res.* **23**: 420–429.

REFERENCES

Paul, E.A. (1988) Soil Microbiology, Ecology and Biochemistry. *Elsevier Science*.286

Paul, J.H., Jiang, S.C., and Rose, J.B. (1991) Concentration of viruses and dissolved DNA from aquatic environments by vortex flow filtration. *Appl. Environ. Microbiol.* **57**: 2197–2204.

Payeta, J.P. and Suttle, C.A. (2013) To kill or not to kill: The balance between lytic and lysogenic viral infection is driven by trophic status. *Limnol.Oceanogr* **58**: 465–474.

Penadés, J.R., Chen, J., Quiles-Puchalt, N., Carpena, N., and Novick, R.P. (2015) Bacteriophage-mediated spread of bacterial virulence genes. *Curr. Opin. Microbiol.* **23**: 171–178.

Pergolizzi, G., Wagner, G.K., and Bowater, R.P. (2016) Biochemical and structural characterization of DNA ligases from bacteria and archaea. *Biosci. Rep.* **36**: e00391–e00391.

Petersen, T.N., Brunak, S., von Heijne, G., and Nielsen, H. (2011) SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nat. Methods* **8**: 785–786.

Plummer, E., Twin, J., Bulach, D.M., Garl, S.M., and Tabrizi, S.N. (2015) A Comparison of Three Bioinformatics Pipelines for the Analysis of Preterm Gut Microbiota using 16S rRNA Gene Sequencing Data. *J. Proteomics Bioinform.* **8**:12

Pluthero, F.G. (1993) Rapid purification of high-activity Taq DNA polymerase. *Nucleic Acids Res.* **21**: 4850–4851.

Postma, A., Slabbert, E., Postma, F., and Jacobs, K. (2016) Soil bacterial communities associated with natural and commercial Cyclopia spp. *FEMS Microbiol. Ecol.* **92**:.

Pruitt, K.D., Tatusova, T., and Maglott, D.R. (2005) NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.* **33**: D501-4.

Racine, J.S. (2012) RStudio: A Platform-Independent IDE for R and Sweave. *J. Appl. Econom.* **27**: 167–172.

Rajendhran, J. and Gunasekaran, P. (2008) Strategies for accessing soil metagenome for desired applications. *Biotechnol. Adv.* **26**: 576–90.

Ramond, J.B., Lako, J.D.W., Stafford, W.H.L., Tuffin, M.I., and Cowan, D.A. (2015) Evidence of novel plant-species specific ammonia oxidizing bacterial clades in acidic South African

REFERENCES

Fynbos soils. *J. Basic Microbiol.* **55**: 1040–1047.

Rappé, M.S. and Giovannoni, S.J. (2003) The uncultured microbial majority. *Annu. Rev. Microbiol.* **57**: 369–394.

Rashid, M. and Stingl, U. (2015) Contemporary molecular tools in microbial ecology and their application to advancing biotechnology. *Biotechnol. Adv.* **33**: 1755–1773.

Reyes, G.R. and Kim, J.P. (1991) Sequence-independent, single-primer amplification (SISPA) of complex DNA populations. *Mol. Cell. Probes* **5**: 473–481.

Reysenbach, A. L., Pace, N. R., Robb, F. T., & Place, A.R. (1995) Archaea: a laboratory manual—thermophiles. *Cold Spring Harb. Protoc* **16**: 101–107.

Reysenbach, A. and Pace, N. (1995) Reliable amplification of hyperthermophilic archaeal 16S rRNA genes by the polymerase chain reaction. In, *Archaea: a laboratory manual. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY*. 101–107.

Ricchetti, M. and Buc, H. (1993) *E. coli* DNA polymerase I as a reverse transcriptase. *EMBO J.* **12**: 387–96.

Richards, M.B., Stock, W.D., and Cowling, R.M. (1997) Soil nutrient dynamics and community boundaries in the *fynbos* vegetation of South Africa. *Plant Ecol.* **130**: 143–153.

Riesenfeld, C.S., Schloss, P.D., and Handelsman, J. (2004) Metagenomics: genomic analysis of microbial communities. *Annu.Rev.Genet.* **38**: 525–552.

Rittié, L. and Perbal, B. (2008) Enzymes used in molecular biology: a useful guide. *J. Cell Commun. Signal.* **2**: 25–45.

Rodrigue, S., Materna, A.C., Timberlake, S.C., Blackburn, M.C., Malmstrom, R.R., Alm, E.J., and Chisholm, S.W. (2010) Unlocking short read sequencing for metagenomics. *PLoS One* **5**:7

Rodriguez-Brito, B., Li, L., Wegley, L., Furlan, M., Angly, F., Breitbart, M., *et al*. (2010) Viral and microbial community dynamics in four aquatic environments. *Isme J* **4**: 739–751.

Rohwer, F. and Barott, K. (2013) Viral information. *Biol. Philos.* **28**: 283–297.

Rohwer, F. and Edwards, R. (2002) The Phage Proteomic Tree: a genome-based taxonomy for phage. *J. Bacteriol.* **184**: 4529–4535.

REFERENCES

Rondon, M.R. and Al, E. (2000) Cloning the metagenome: a stragy for accessing the genetic and functional diversity of uncultured microorganisms. *Appl. Environ. Microbiol.* **66**: 2541–2547.

Rosario, K. and Breitbart, M. (2011) Exploring the viral world through metagenomics. *Curr. Opin. Virol.* **1**: 289–297.

Rose, R., Constantinides, B., Tapinos, A., Robertson, D.L., and Prosperi, M. (2016) Challenges in the analysis of viral metagenomes. *Virus Evol.* **2(2)**:

Rosenberg, A.H., Lade, B.N., Dao-shan, C., Lin, S.-W., Dunn, J.J., and Studier, F.W. (1987) Vectors for selective expression of cloned DNAs by T7 RNA polymerase. *Gene* **56**: 125–135.

Rosseel, T., Van Borm, S., Vandenbussche, F., Hoffmann, B., van den Berg, T., Beer, M., and Höper, D. (2013) The Origin of Biased Sequence Depth in Sequence-Independent Nucleic Acid Amplification and Optimization for Efficient Massive Parallel Sequencing. *PLoS One* **8**: e76144.

Rothberg, J.M., Hinz, W., Rearick, T.M., Schultz, J., Mileski, W., Davey, M., *et al*. (2011) An integrated semiconductor device enabling non-optical genome sequencing. *Nature* **475**: 348–352.

Rousk, J., Bååth, E., Brookes, P.C., Lauber, C.L., Lozupone, C., Caporaso, J.G., *et al*. (2010) Soil bacterial and fungal communities across a pH gradient in an arable soil. *ISME J.* **4**: 1340–1351.

Roux, S., Enault, F., Ravet, V., Colombet, J., Bettarel, Y., Auguet, J.C., *et al*. (2015) Analysis of metagenomic data reveals common features of halophilic viral communities across continents. *Environ. Microbiol.* **18**: 889–903.

Roux, S., Enault, F., Robin, A., Ravet, V., Personnic, S., Theil, S., *et al*. (2012) Assessing the Diversity and Specificity of Two Freshwater Viral Communities through Metagenomics. *PLoS One* **7**: e33641.

Roux, S., Hallam, S.J., Woyke, T., and Sullivan, M.B. (2015) Viral dark matter and virus-host interactions resolved from publicly available microbial genomes. *Elife* **4**: e08490.

Roux, S., Hawley, A.K., Torres Beltran, M., Scofield, M., Schwientek, P., Stepanauskas, R., *et al*. (2014) Ecology and evolution of viruses infecting uncultivated SUP05 bacteria as

revealed by single-cell- and meta-genomics. *Elife* **3**: 1–20.

Roux, S., Solonenko, N.E., Dang, V.T., Poulos, B.T., Schwenck, S.M., Goldsmith, D.B., *et al*. (2016) Towards quantitative viromics for both double-stranded and single-stranded DNA viruses. *PeerJ* **4**: e2777.

Roux, S., Tournayre, J., Mahul, A., Debroas, D., and Enault, F. (2014) Metavir 2: new tools for viral metagenome comparison and assembled virome analysis. *BMC Bioinformatics* **15**: 76.

Roychoudhury, R., Jay, E., and Wu, R. (1976) Terminal labeling and addition of homopolymer tracts to duplex DNA fragments by terminal deoxynucleotidyl transferase. *Nucleic Acids Res.* **3**: 863–878.

Sambrook, J. and Russell, D.W. (2001) Molecular cloning. *A Laboratory Manual.Third edition Cold Spring Harbor Laboratory Press Cold Spring Harbor NY*. 1-3.

Sander, J.D. and Joung, J.K. (2014) CRISPR-Cas systems for editing, regulating and targeting genomes. *Nat. Biotechnol.* **32**: 347–355.

Sanger, F., Coulson, A.R., Hong, G.F., Hill, D.F., and Petersen, G.B. (1982) Nucleotide sequence of bacteriophage λ DNA. *J. Mol. Biol.* **162**: 729–773.

Santos, F., Yarza, P., Parro, V., Briones, C., and Antón, J. (2010) The metavirome of a hypersaline environment. *Environ. Microbiol.* **12**: 2965–2976.

Schloss, P.D. and Handelsman, J. (2003) Biotechnological prospects from metagenomics. *Curr. Opin. Biotechnol.* **14**: 303–310.

Schloss, P.D., Westcott, S.L., Ryabin, T., Hall, J.R., Hartmann, M., Hollister, E.B., *et al*. (2009) Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl. Environ. Microbiol.* **75**: 7537–41.

Schmidt, H.F., Sakowski, E.G., Williamson, S.J., Polson, S.W., and Wommack, K.E. (2014) Shotgun metagenomics indicates novel family A DNA polymerases predominate within marine virioplankton. *ISME J.* **8**: 103–114.

Schmidt, T.M., DeLong, E.F., and Pace, N.R. (1991) Analysis of a marine picoplankton community by 16S rRNA gene cloning and sequencing. *J. Bacteriol.* **173**: 4371–4378.

REFERENCES

Schmitz, J.E., Schuch, R., and Fischetti, V.A. (2010) Identifying active phage lysins through functional viral metagenomics. *Appl. Environ. Microbiol.* **76**: 7181–7.

Schoenfeld, T., Liles, M., Wommack, K.E., Polson, S.W., Godiska, R., and Mead, D. (2010) Functional viral metagenomics and the next generation of molecular tools. *Trends Microbiol.* **18**: 20–9.

Schoenfeld, T.W., Hermersmann, N., Moser, M., Renneckar, D., Dhodda, V., and Mead, D. (2011) Functional Viral Metagenomics and the Development of New Enzymes for DNA and RNA Amplification and Sequencing. In, *Handbook of Molecular Microbial Ecology II: Metagenomics in Different Habitats. Springer US, Boston, MA*. 563–577.

Shao, Q., Trinh, J.T., McIntosh, C.S., Christenson, B., Balázsi, G., and Zeng, L. (2016) Lysis-lysogeny coexistence: prophage integration during lytic development. *Microbiologyopen*.

Shoaib, M., Baconnais, S., Mechold, U., Le Cam, E., Lipinski, M., and Ogryzko, V. (2008) Multiple displacement amplification for complex mixtures of DNA fragments. *BMC Genomics* **9**: 415.

Shuman, S. and Schwer, B. (1995) RNA capping enzyme and DNA ligase: a superfamily of covalent nucleotidyl transferases. *Mol. Microbiol.* **17**: 405–420.

Siavash Bashiri, D.V.N.I. (2014) Optimization of Protein Expression in Escherichia Coli. *BioPharm Int.* **28**: 42–44.

Siles, J.A. and Margesin, R. (2016) Abundance and Diversity of Bacterial, Archaeal, and Fungal Communities Along an Altitudinal Gradient in Alpine Forest Soils: What Are the Driving Factors? *Microb. Ecol.* **72**: 207–220.

Simon, C. and Daniel, R. (2011) Metagenomic analyses: past and future trends. *Appl. Environ. Microbiol.* **77**: 1153–1161.

Simon, C., Herath, J., Rockstroh, S., and Daniel, R. (2009) Rapid identification of genes encoding DNA polymerases by function-based screening of metagenomic libraries derived from glacial ice. *Appl. Environ. Microbiol.* **75**: 2964–2968.

Singh, S.M. and Panda, A.K. (2005) Solubilization and refolding of bacterial inclusion body proteins. *J. Biosci. Bioeng.* **99**: 303–310.

Slabbert, E., Jacobs, S.M., and Jacobs, K. (2014) The soil bacterial communities of South

REFERENCES

African *fynbos* riparian ecosystems invaded by Australian Acacia species. *PLoS One* **9**: 1–10.

Slabbert, E., Kongor, R.Y., Esler, K.J., and Jacobs, K. (2010) Microbial diversity and community structure in *fynbos* soil. *Mol. Ecol.* **19**: 1031–1041.

Smits, S.A. and Ouverney, C.C. (2010) jsPhyloSVG: A Javascript Library for Visualizing Interactive and Vector-Based Phylogenetic Trees on the Web. *PLoS One* **5**: e12267.

Snyder, J.C., Bolduc, B., and Young, M.J. (2015) 40 Years of archaeal virology: Expanding viral diversity. *Virology* **479**–**480**: 369–78.

Sohoni, S.V., Nelapati, D., Sathe, S., Javadekar-Subhedar, V., Gaikaiwari, R.P., and Wangikar, P.P. (2015) Optimization of high cell density fermentation process for recombinant nitrilase production in *E. coli*. *Bioresour. Technol.* **188**: 202–208.

Spot, M., Collections, A., and Headquarters, H. (2012) Examples of Restriction Enzymes. *Health.* 3–5.

Spriggs, A.C., Stock, W.D., and Dakora, F.D. (2003) Influence of mycorrhizal associations on foliar δ15N values of legume and non-legume shrubs and trees in the *fynbos* of South Africa: Implications for estimating N2 fixation using the 15N natural abundance method. *Plant Soil* **255**: 495–502.

Srinivasiah, S., Lovett, J., Polson, S., Bhavsar, J., Ghosh, D., Roy, K., *et al*. (2013) Direct assessment of viral diversity in soils by random PCR amplification of polymorphic DNA. *Appl. Environ. Microbiol.* **79**: 5450–7.

Stafford, W.H.L., Baker, G.C., Brown, S.A., Burton, S.G., and Cowan, D.A. (2005) Bacterial diversity in the rhizosphere of Proteaceae species. *Environ. Microbiol.* **7**: 1755–1768.

Steward, G.F. (2001) Fingerprinting viral assemblages by pulsed field gel electrophoresis (PFGE). *Methods Microbiol.* **30**: 85–104.

Strazzulli, A., Fusco, S., Cobucci-Ponzano, B., Moracci, M., and Contursi, P. (2017) Metagenomics of microbial and viral life in terrestrial geothermal environments. *Rev. Environ. Sci. Bio/Technology* **16**: 425–454.

Struhl, K. (2009) Enzymatic manipulation of DNA and RNA. *Current Protocols in Molecular Biology.86*

REFERENCES

Studier, F.W. and Moffatt, B.A. (1986) Use of bacteriophage T7 RNA polymerase to direct selective high-level expression of cloned genes. *J. Mol. Biol.* **189**: 113–130.

Suzuki, R. and Shimodaira, H. (2006) Pvclust: An R package for assessing the uncertainty in hierarchical clustering. *Bioinformatics* **22**: 1540–1542.

Tabor, S. and Richardson, C.C. (1985) A Bacteriophage T7 RNA Polymerase/Promoter System for Controlled Exclusive Expression of Specific Genes A bacteriophage T7 RNA polymer: controlled exclusive expression of (T7 DNA polymerase/T7 gene 5 protein/proteolysis/f,-lactamase/r. *Source Proc. Natl. Acad. Sci. United States Am. Biochem.* **82**: 1074–1078.

Tamura, K., Stecher, G., Peterson, D., Filipski, A., and Kumar, S. (2013) MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. *Mol. Biol. Evol.* **30**: 2725–9.

Tan, S.C. and Yiap, B.C. (2009) DNA, RNA, and protein extraction: the past and the present. *J. Biomed. Biotechnol.* **2009**: 574398.

Tatusov, R.L., Fedorova, N.D., Jackson, J.D., Jacobs, A.R., Kiryutin, B., Koonin, E. V, *et al.* (2003) The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* **4**: 41.

Taylor, M.R. (2014) The Role of Divalent Metal Ions in Enzymatic DNA Ligation. *Biol. Chem.***2014**

Teeling, H. and Glockner, F.O. (2012) Current opportunities and challenges in microbial metagenome analysis--a bioinformatic perspective. *Brief. Bioinform.* **13**: 728–742.

Terpe, K. (2006) Overview of bacterial expression systems for heterologous protein production: From molecular and biochemical fundamentals to commercial systems. *Appl. Microbiol. Biotechnol.* **72**: 211–222.

Terrat, S., Horrigue, W., Dequietd, S., Saby, N.P.A., Lelièvre, M., Nowak, V., *et al.* (2017) Mapping and predictive variations of soil bacterial richness across France. *PLoS One* **12**: e0186766.

Terrón-González, L., Medina, C., Limón-Mortés, M.C., and Santero, E. (2013) Heterologous viral expression systems in fosmid vectors increase the functional analysis potential of metagenomic libraries. *Sci. Rep.* **3**: 1107.

REFERENCES

Thomas, T., Gilbert, J., and Meyer, F. (2012) Metagenomics - a guide from sampling to data analysis. *Microb. Inform. Exp.* **2**: 3.

Thurber, R. V, Haynes, M., Breitbart, M., Wegley, L., and Rohwer, F. (2009) Laboratory procedures to generate viral metagenomes. *Nat. Protoc.* **4**: 470–83.

Torsvik, V., Goksøyr, J., and Daae, F.L. (1990) High diversity in DNA of soil bacteria. *Appl. Environ. Microbiol.* **56**: 782–7.

Armstrong, J., Brown, R.S., and Tsugita, A. Primary structure and genetic organization of phage T4 DNA ligase (1984). *Nucleic Acids Research*. **12**: 6397–6414.

Trill, J.J., Kirkpatrick, R., Shatzman, A.R., and Marcy, A. (2001) Eukaryotic expression. *Molecular Biology Problem Solver: A Laboratory Guide.* **7:**491-542

Tripathi, M., Narain, D., Vikram, S., Singh, V.S., and Kumar, S. (2018) Metagenomic Approach towards Bioprospection of Novel Biomolecule(s) and Environmental Bioremediation. *ť Contrib. Equal. Annu. Res. Rev. Biol. Egypt. Pet. Res. Institute, Egypt* **22**: 1–12.

Uchiyama, T. and Miyazaki, K. (2009) Functional metagenomics for enzyme discovery: challenges to efficient screening. *Curr. Opin. Biotechnol.* **20**: 616–22.

Vallejo, L.F. and Rinas, U. (2004) Strategies for the recovery of active proteins through refolding of bacterial inclusion body proteins. *Microb. Cell Fact.* **3**: 1–12.

Venter, J.C., Remington, K., Heidelberg, J.F., Halpern, A.L., Rusch, D., Eisen, J.A., *et al.* (2004) Environmental Genome Shotgun Sequencing of the Sargasso Sea. *Science.* **304**:5667

Vergin, K.L., Beszteri, B., Monier, A., Cameron Thrash, J., Temperton, B., Treusch, A.H., *et al.* (2013) High-resolution SAR11 ecotype dynamics at the Bermuda Atlantic Time-series Study site by phylogenetic placement of pyrosequences. *ISME J.* **7**: 1322–1332.

Vester, J.K., Glaring, M.A., and Stougaard, P. (2014) Discovery of novel enzymes with industrial potential from a cold and alkaline environment by a combination of functional metagenomics and culturing. *Microbial Cell Factories.* 1–14.

Vierheilig, J., Savio, D., Ley, R.E., Mach, R.L., Farnleitner, A.H., and Reischer, G.H. (2015) Potential applications of next generation DNA sequencing of 16S rRNA gene amplicons

REFERENCES

in microbial water quality monitoring. *Water Sci. Technol.* **72**: 1962–1972.

Wang, M.Y. (2016) The bacterial communities of sand-like surface soils of the San Rafael Swell (Utah , USA) and the Desert of Maine (USA). *Agricultural sciences. Université Paris-Saclay*

Watkins, S.C., Hatzopoulos, T., and Putonti, C. (2016) The Use of Informativity in the Development of Robust Metaviromics-based Examinations. *bioRxiv* 1–17.

Watkins, S.C., Kuehnle, N., Ruggeri, C.A., Malki, K., Bruder, K., Elayyan, J., *et al*. (2016) Assessment of a metaviromic dataset generated from nearshore Lake Michigan. *Mar. Freshw. Res.* **67**: 1700.

Weinbauer, M.G. (2004) Ecology of prokaryotic viruses. *FEMS Microbiol. Rev.* **28**: 127–181.

Wilhelm, S.W. and Matteson, A.R. (2008) Freshwater and marine virioplankton: a brief overview of commonalities and differences. *Freshw. Biol.* **53**: 1076–1089.

Wilkinson, A., Day, J., and Bowater, R. (2001) Bacterial DNA ligases. *Mol. Microbiol.* **40**: 1241–1248.

Williamson, K.E., Radosevich, M., Smith, D.W., and Wommack, K.E. (2007) Incidence of lysogeny within temperate and extreme soil environments. *Environ. Microbiol.* **9**: 2563–2574.

Williamson, K.E., Radosevich, M., and Wommack, K.E. (2005) Abundance and diversity of viruses in six Delaware soils. *Appl. Environ. Microbiol.* **71**: 3119–3125.

Willner, D., Thurber, R.V., and Rohwer, F. (2009) Metagenomic signatures of 86 microbial and viral metagenomes. *Environ. Microbiol.* **11**: 1752–1766.

Wintle, B.A., Bekessy, S.A., Keith, D.A., Van Wilgen, B.W., Cabeza, M., Schröder, B., *et al*. (2011) Ecological--economic optimization of biodiversity conservation under climate change. *Nature Climate Change*: **7**: 355-359

Woese, C.R., Kandler, O., and Wheelis, M.L. (1990) Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. *Proc. Natl. Acad. Sci. U. S. A.* **87**: 4576–9.

Wommack, K.E., Bhavsar, J., Polson, S.W., Chen, J., Dumas, M., Srinivasiah, S., *et al*. (2012) VIROME: a standard operating procedure for analysis of viral metagenome sequences.

REFERENCES

Stand. Genomic Sci. **6**: 427.

Wood-Charlson, E.M., Weynberg, K.D., Suttle, C.A., Roux, S., and van Oppen, M.J.H. (2015) Metagenomic characterization of viral communities in corals: mining biological signal from methodological noise. *Environ. Microbiol.* **17**: 3440–9.

Wooley, J.C., Godzik, A., and Friedberg, I. (2010) A primer on metagenomics. *PLoS Comput. Biol.* **6**: e1000667.

Wylie, T.N., Wylie, K.M., Herter, B.N., and Storch, G.A. (2015) Enhanced virome sequencing using targeted sequence capture. *Genome Res.* **25**: 1910–1920.

Xiang, L., Li, A., Tian, C., Zhou, Y., Zhang, G., and Ma, Y. (2014) Identification and characterization of a new acid-stable endoglucanase from a metagenomic library. *Protein Expr. Purif.* **102**: 20–26.

Yamagami, T., Matsukawa, H., Tsunekawa, S., Kawarabayasi, Y., Ishino, S., and Ishino, Y. (2015) A longer finger-subdomain of family A DNA polymerases found by metagenomic analysis strengthens DNA binding and primer extension abilities. *Gene* **576**: 690–695.

Yildir, C., Onsan, I., and Kirdar, B. (1998) Optimization of Starting Time and Period of Induction and Inducer Concentration in the Production of the Restriction Enzyme EcoRI from Recombinant Escherichia coli 294. *Turk J Chem* **22**: 221–226.

Yilmaz, S., Allgaier, M., and Hugenholtz, P. (2010) Multiple displacement amplification compromises quantitative analysis of metagenomes. *Nat. Methods* **7**: 943–944.

Yin, J., Li, G., Ren, X., and Herrler, G. (2007) Select what you need: A comparative evaluation of the advantages and limitations of frequently used expression systems for foreign genes. *J. Biotechnol.* **127**: 335–347.

Zablocki, O., Adriaenssens, E.M., and Cowan, D. (2016) Diversity and Ecology of Viruses in Hyperarid Desert Soils. *Appl Env. Microbio* **82**: 770–777.

Zakabunin, A.I., Kamynina, T.P., Khodyreva, S.N., Pyshnaya, I.A., Pyshnyi, D. V, Khrapov, E.A., and Filipenko, M.L. (2011) Gene cloning, purification, and characterization of recombinant DNA ligases of the thermophilic archaea *Pyrococcus abyssi* and *Methanobacterium thermoautotrophicum*. *Mol. Biol.* **45**: 229–236.

Zhao, G., Krishnamurthy, S., Cai, Z., Popov, V.L., Travassos da Rosa, A.P., Guzman, H., *et*

REFERENCES

al. (2013) Identification of Novel Viruses Using VirusHunter -- an Automated Data Analysis Pipeline. *PLoS One* **8**: 1–11.

Zhu, W. and Ito, J. (1994) Family A and family B DNA polymerases are structurally related: evolutionary implications. *Nucleic Acids Res.* **22**: 5177–5183.

(http://www.marketsandmarkets.com/Market-Reports/molecular-biology-enzymes-kits-reagents-market-164131709.html)

(https://courses.lumenlearning.com/boundless-biology/chapter/virus-infections-and-hosts/)

(https://emilybio11.weebly.com /microbiology.html)

(https://i.pinimg.com/736x/6c/1a/1b/ 6c1a1baf2906c6c098c4390e47999944.jpg)

APPENDIX

Appendices A: Chapter 2

Flow Diagram of the Experimental Procedures

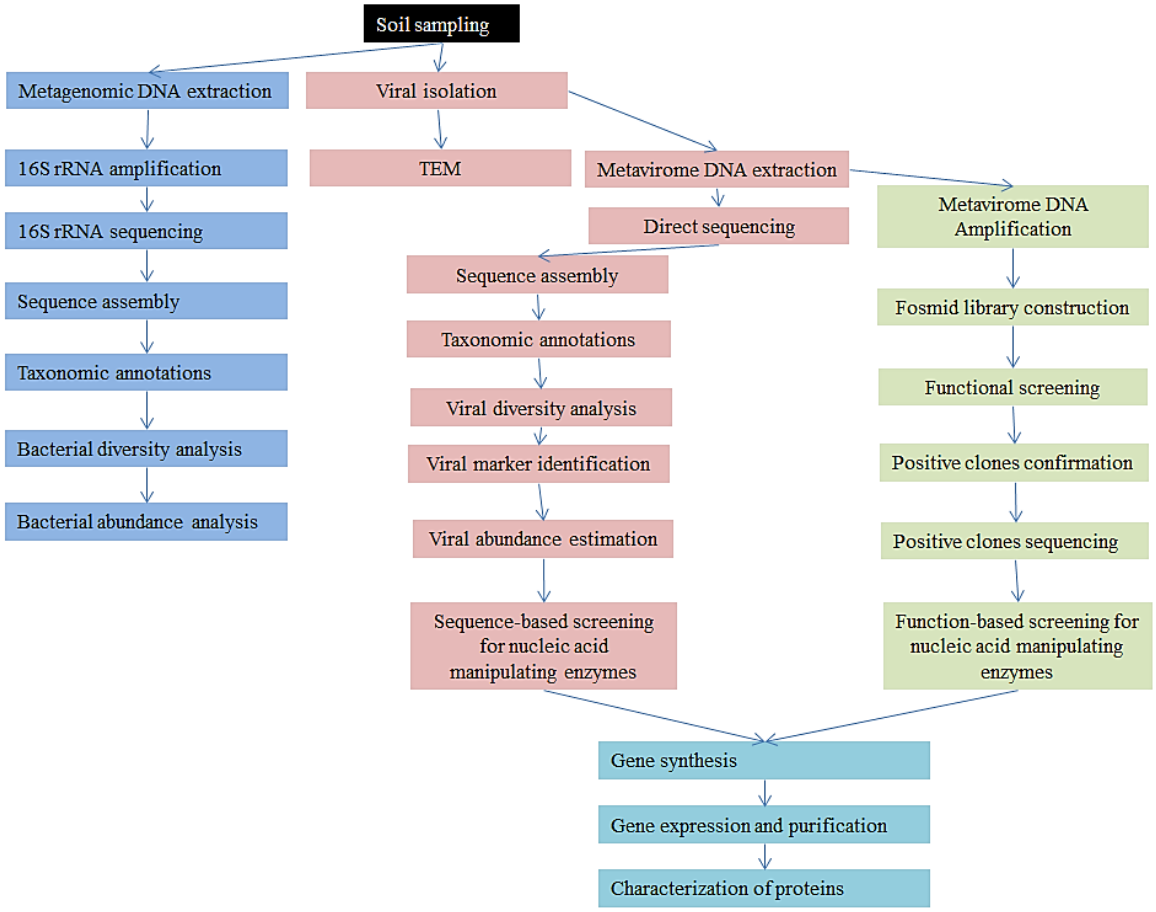


Figure A1: The flow diagram below summarizes the step to step procedures used in this study.

APPENDIX

Appendices B: Chapter 3

Table B1: Chemical and physical properties of the Kogelberg Biosphere Reserve *fynbos* soil samples

Analysis	Unit	KBR 1	KBR 2	KBR 3	AVE
Potassium as K Soluble ions(wet)	mg/kg	<20	<20	<20	<20
Sodium as Na Soluble ions(wet)	mg/kg	<10	<10	<10	<10
Calcium as Ca Soluble ions(wet)	mg/kg	<5	<5	<5	<5
Magnesium as Mg Soluble ions(wet)	mg/kg	<1	<1	<1	<1
Sulphate as SO4 soluble ion (wet)	mg/kg	<5	<5	<5	<5
Chloride as Cl Soluble ions(wet)	mg/kg	<10	<10	<10	<10
pH (Lab) (20°C)		5.2	5.2	5.5	5.3
% moisture		19	10	5	11.3
Aluminium as Al water soluble (wet)	mg/kg	0.62	0.37	*<0.20	~0.4
Iron as Fe water soluble (wet)	mg/kg	*<0.10	*<0.10	*<0.10	*<0.10
Manganese as Mn water soluble (wet)	mg/kg	*<0.10	*<0.10	*<0.10	*<0.10

*the values that are below the limit of detection

APPENDIX

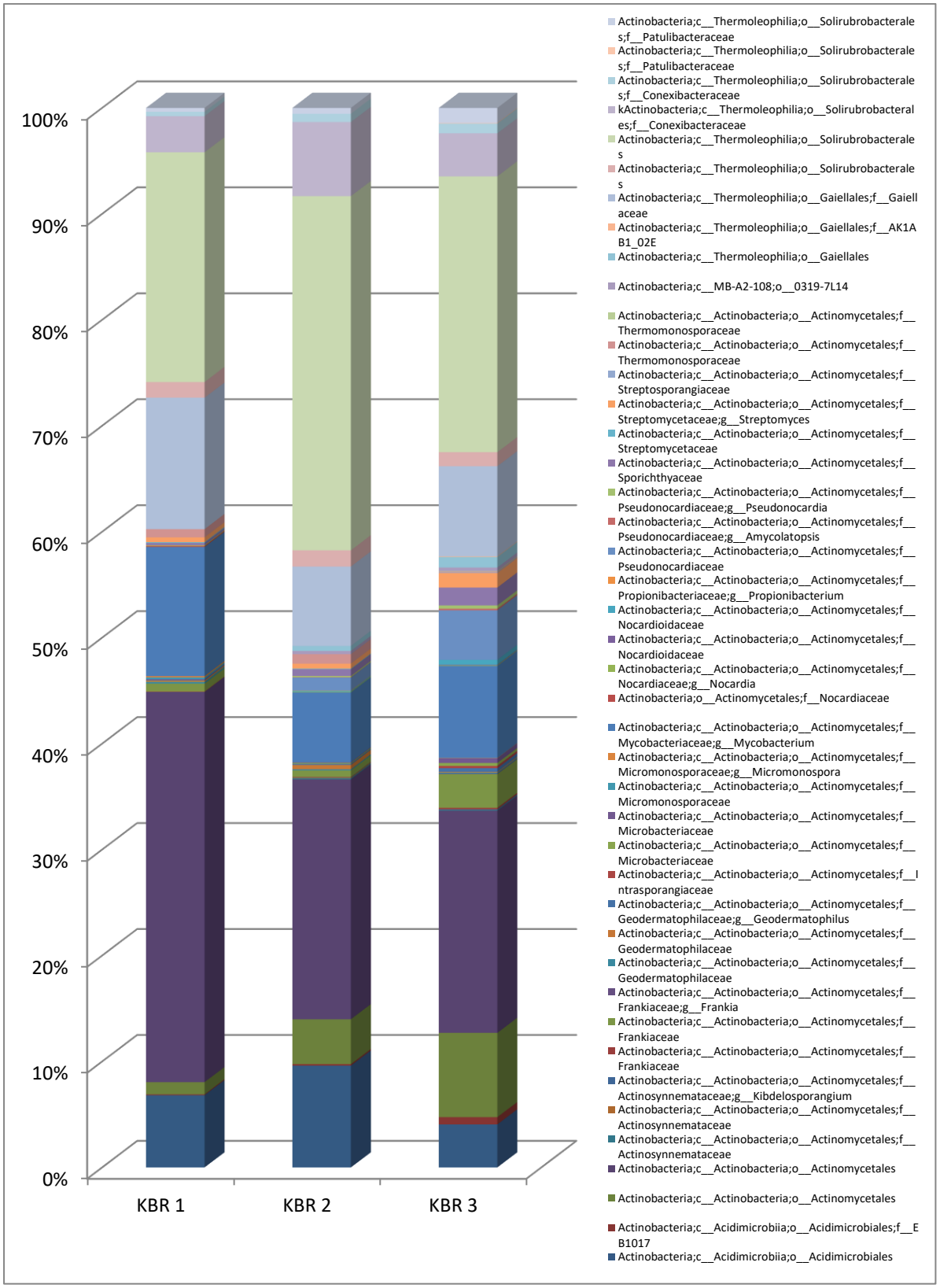


Figure B1: Bar charts represent the taxonomic distribution of *Actinobacteria* phylogenetic groups at the genus level.

APPENDIX

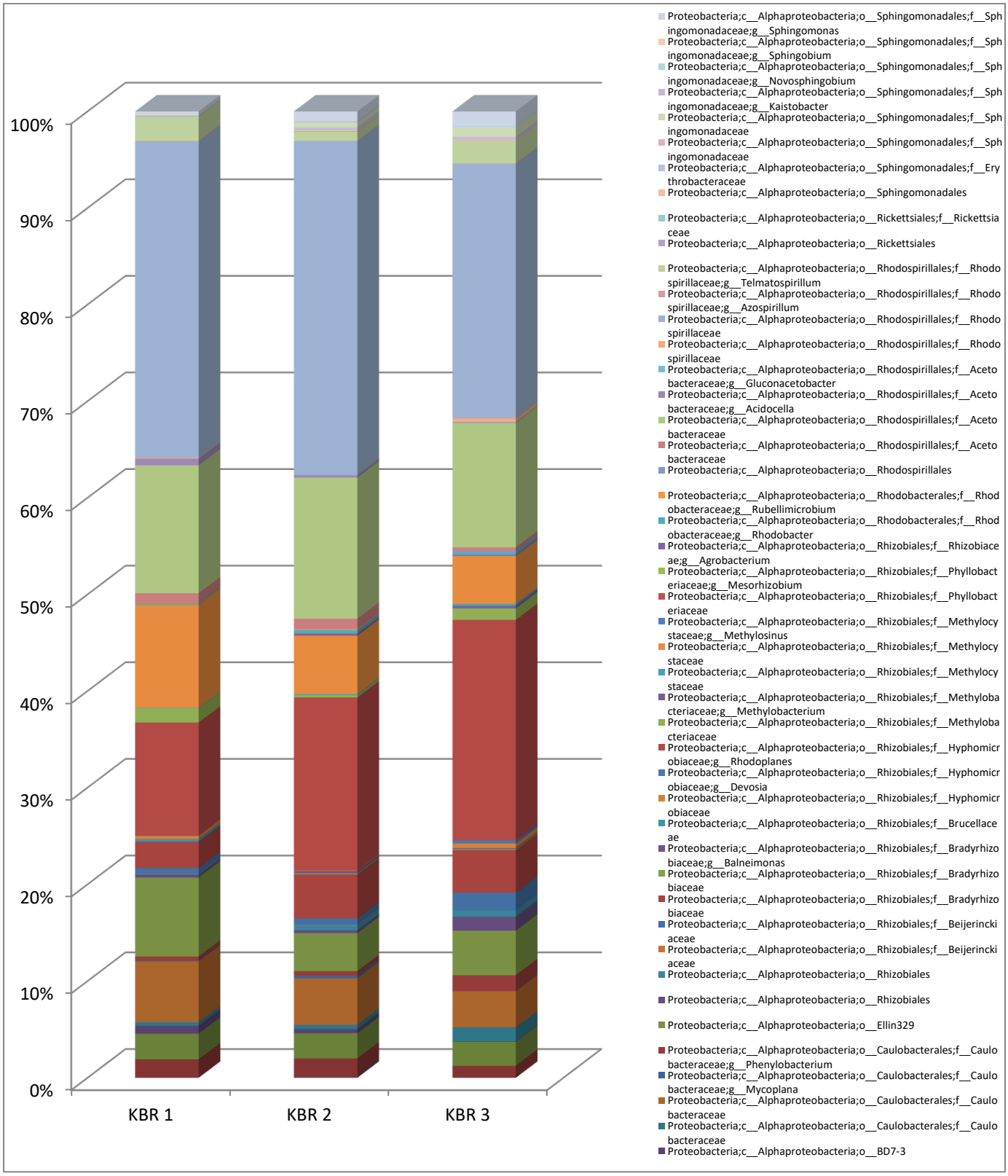


Figure B2: Bar charts represent the taxonomic distribution of *Alphaproteobacteria* phylogenetic groups at the genus level.

APPENDIX

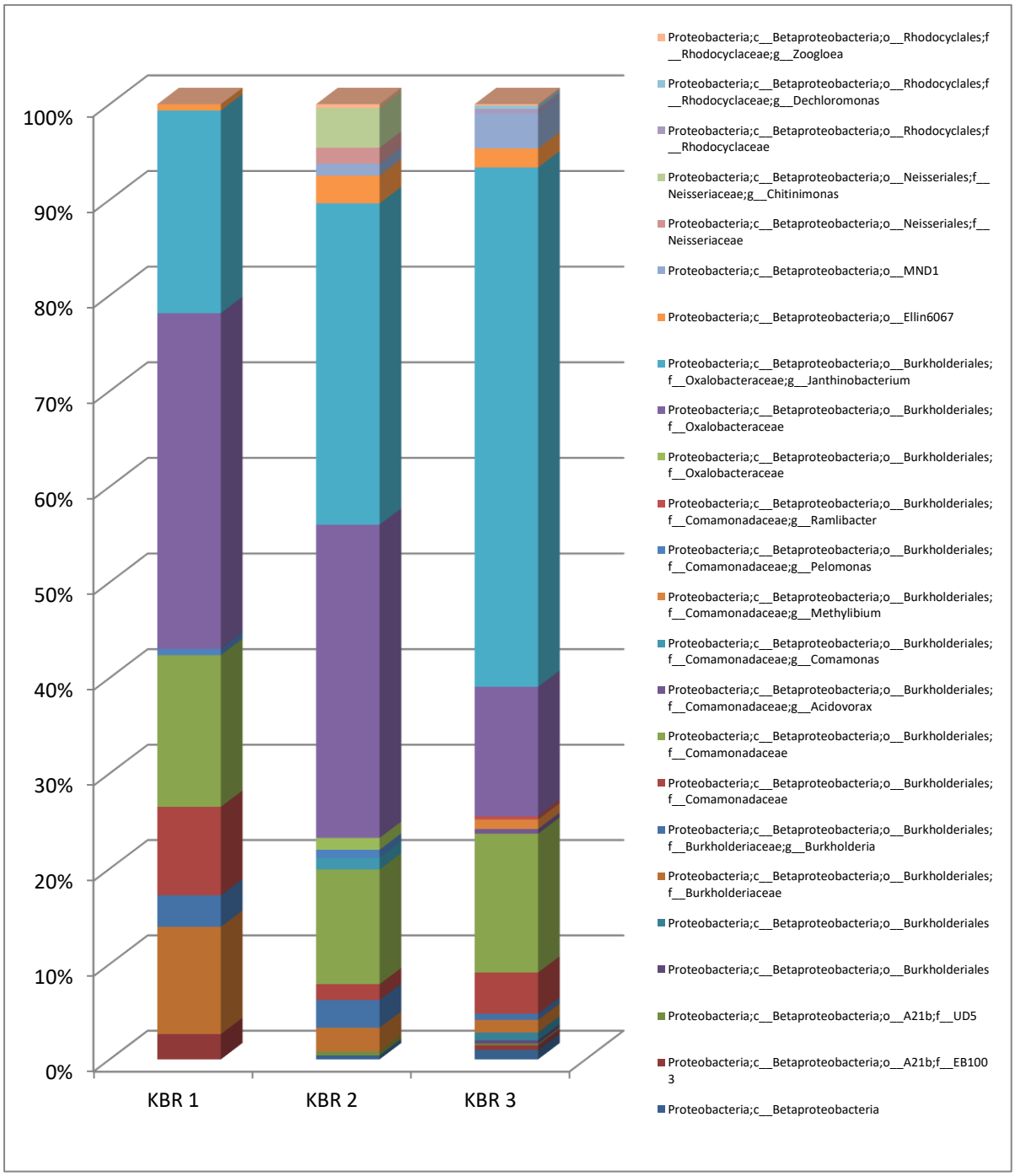


Figure B3: Bar charts represent the taxonomic distribution of *Betaproteobacteria* phylogenetic groups at the genus level.

APPENDIX

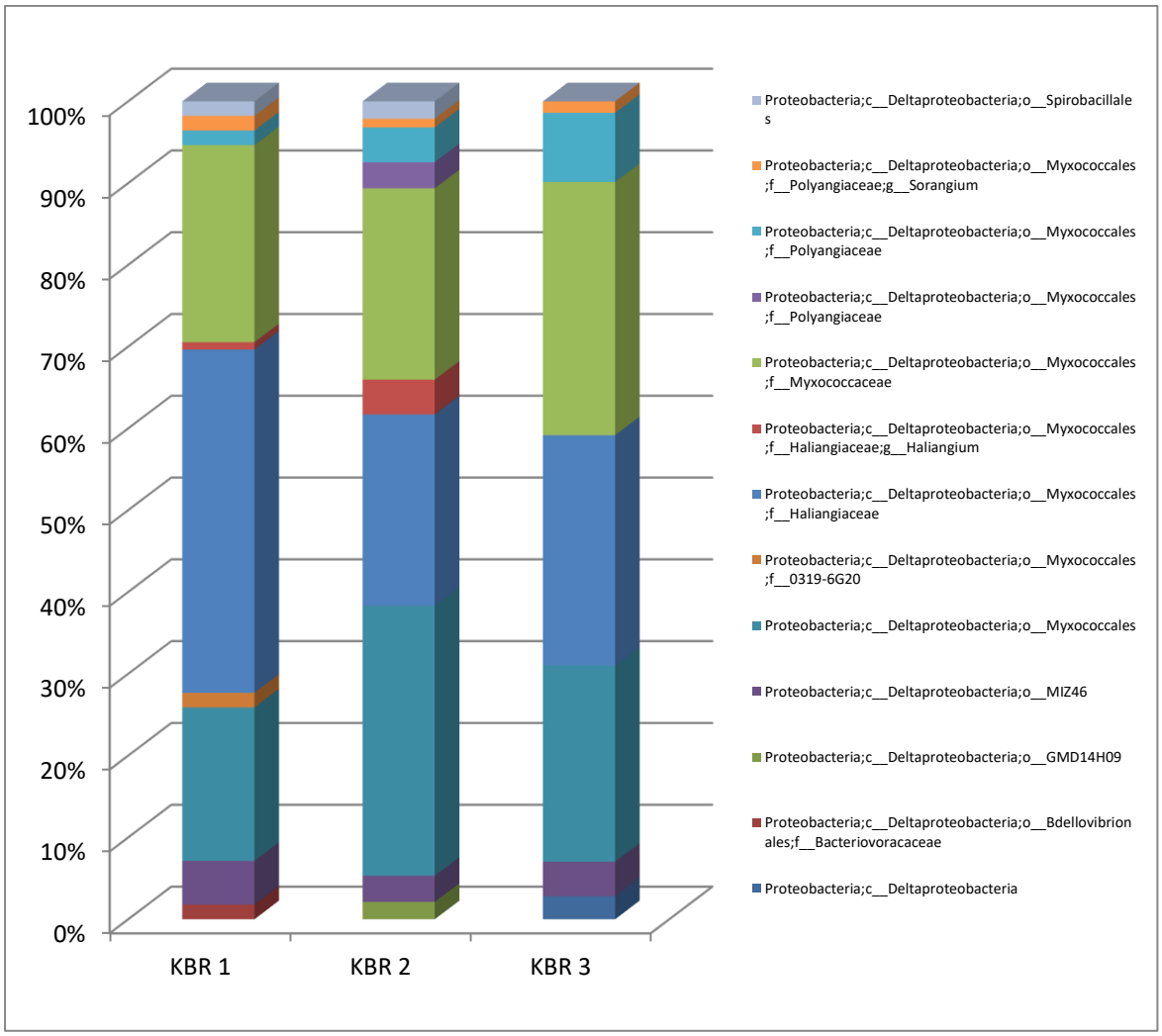


Figure B4: Bar charts represent the taxonomic distribution of *Deltaproteobacteria* phylogenetic groups at the genus level.

APPENDIX

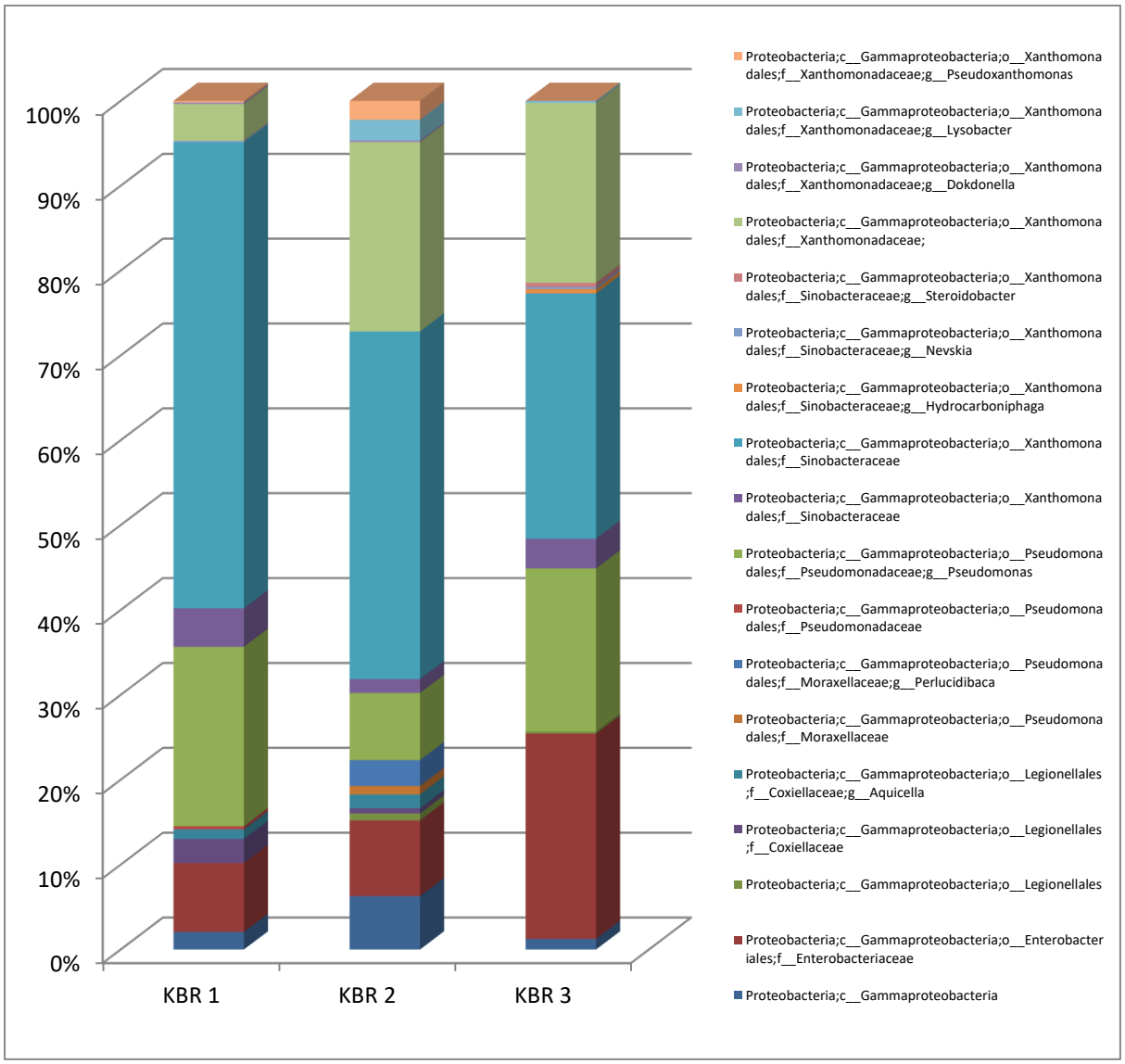


Figure B5: Bar charts represent the taxonomic distribution of *Gammaproteobacteria* phylogenetic groups at the genus level.

APPENDIX

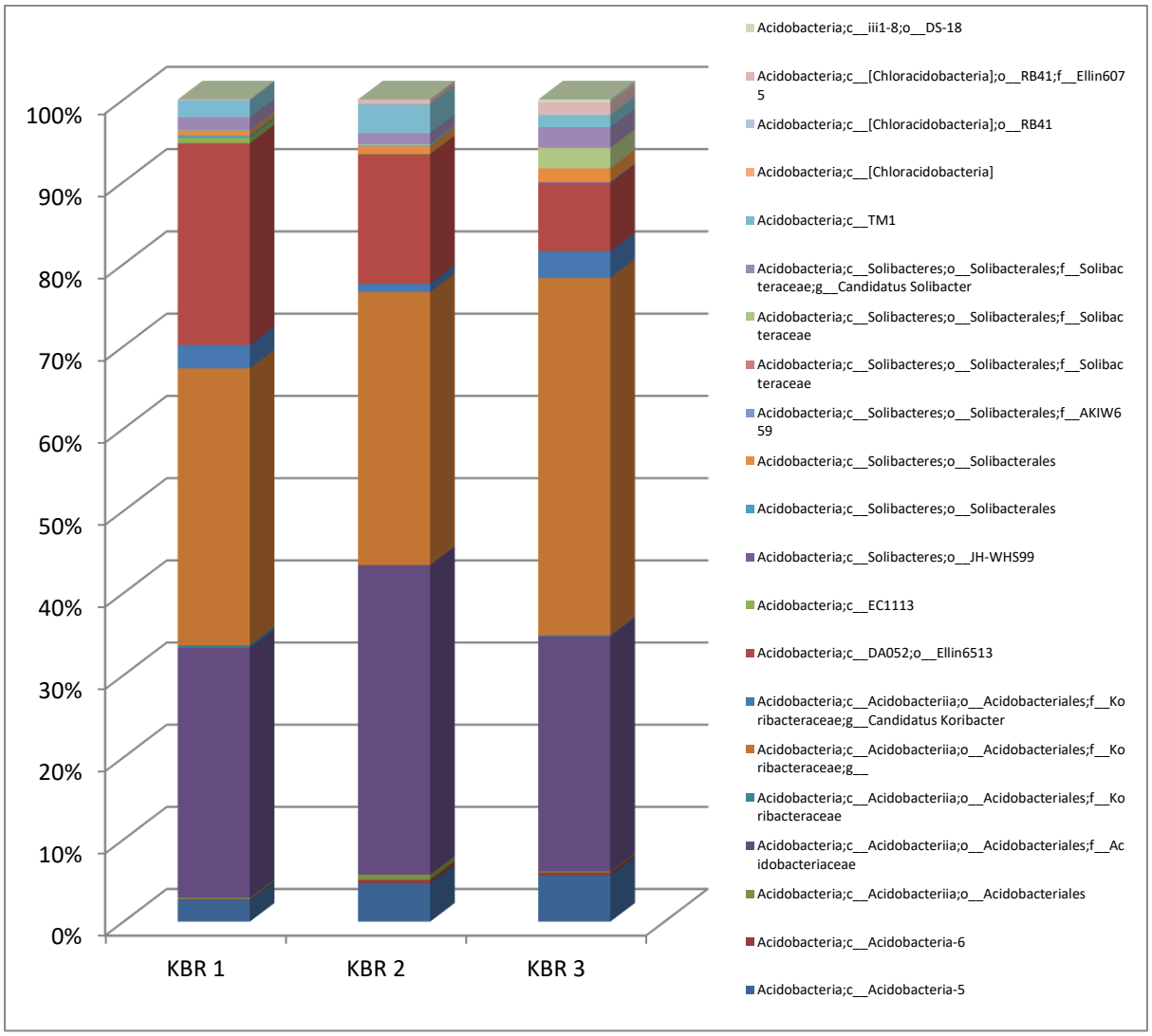


Figure B6: Bar charts represent the taxonomic distribution of *Acidobacteria* phylogenetic groups at the genus level.

APPENDIX

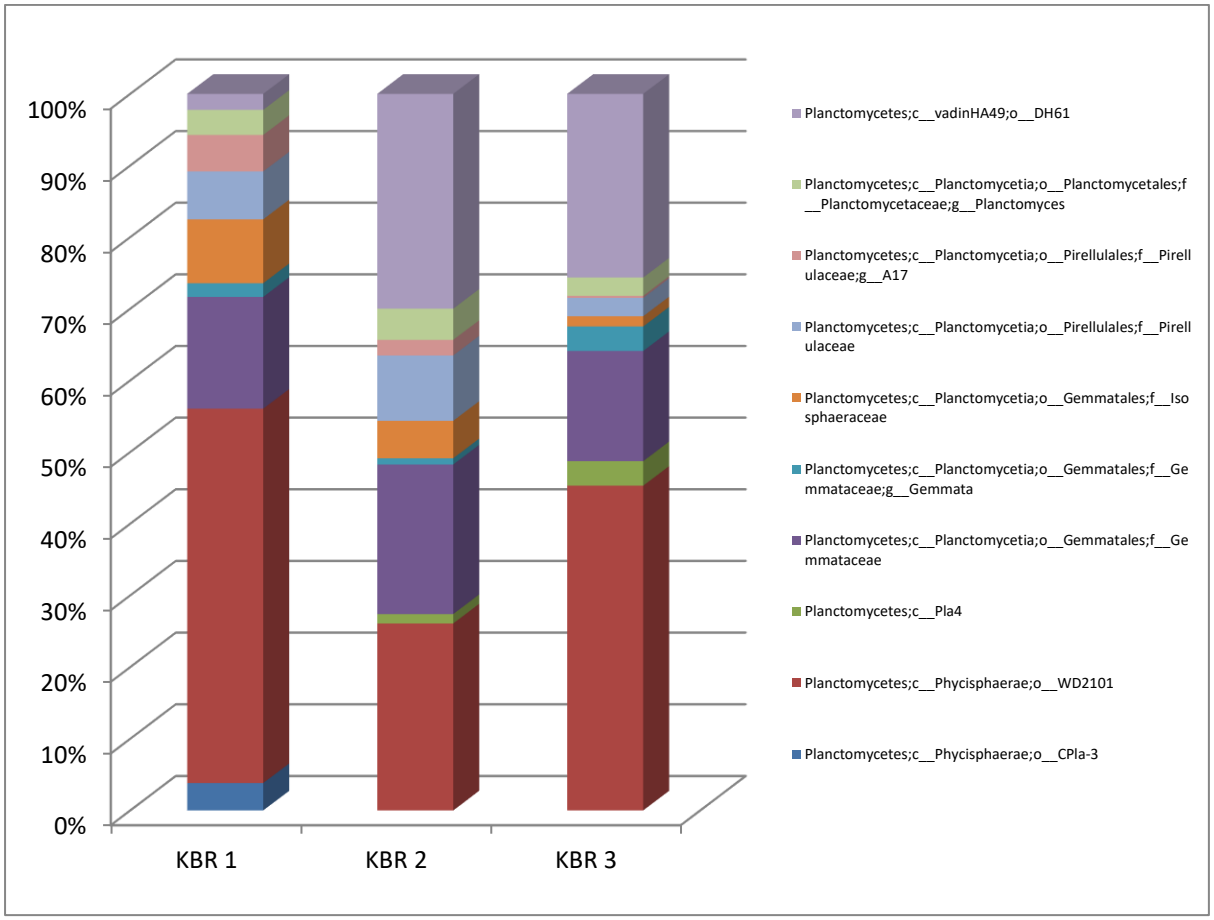


Figure B7: Bar charts represent the taxonomic distribution of *Plantomycetes* phylogenetic groups at the genus level.

APPENDIX

Appendices C: Chapter 4

SUPPLEMENTARY PAPER

Exploring Viral Diversity In A Unique South African Soil Habitat

Jane Segobola[1,2], Evelien Adriaenssens[2], Tsepo Tsekoa[1], Konanani Rashamuse[1] and Don Cowan[2]

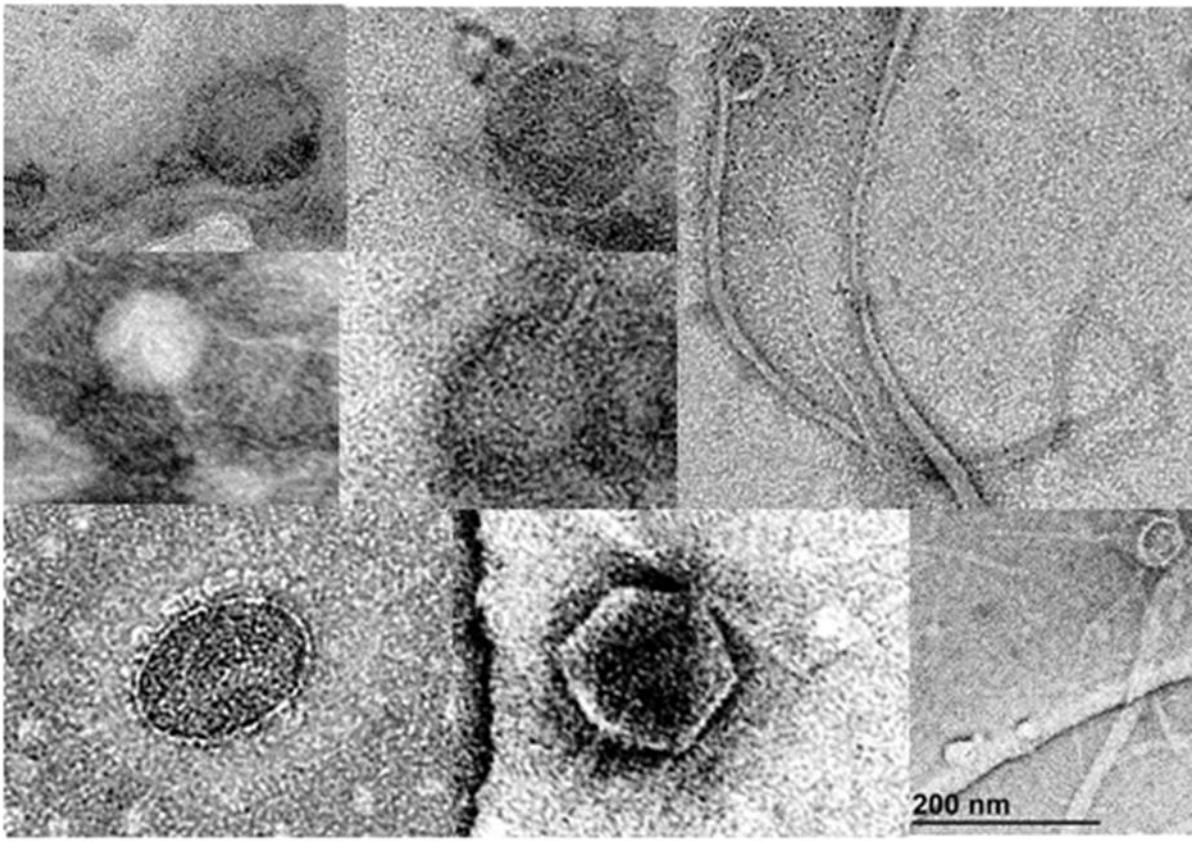
[1]Biosciences Unit, Council for Scientific and Industrial Research (CSIR), Pretoria, South Africa

[2] Centre For Microbial Ecology and Genomics, University Of Pretoria, Pretoria, South Africa

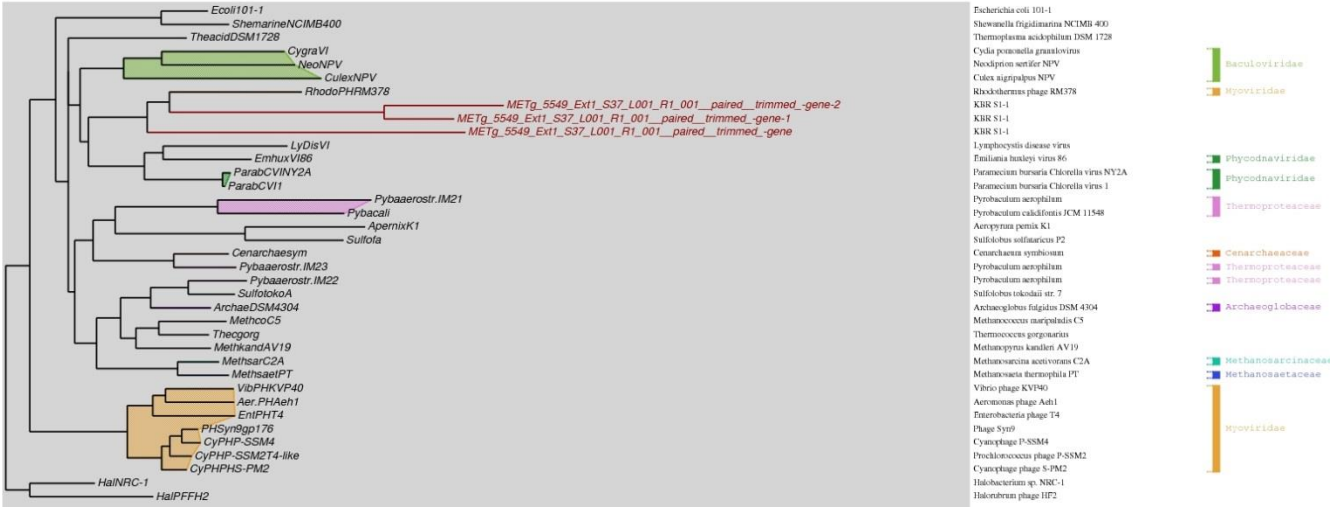
Corresponding Author: : Don Cowan, Centre for Microbial Ecology and Genomics, Department of Genetics, University of Pretoria, Hatfield 0028, Pretoria, South Africa. Tel: +27 (12) 420 5873, don.cowan@up.ac.za

APPENDIX

SUPPLEMENTARY FIGURES:



Supplementary Figure S1: Transmission electron micrographs of the viral particles obtained from KBR. Scale bars correspond to 200 nm. Particles were negatively stained with 2% uranyl acetate. Virus particles observed either belong to the families *Siphoviridae, Myoviridae* or are of an undetermined shape (VLP).



APPENDIX

Supplementary Figure S2: Phylogenetic tree of Pol B Marker gene amino acid sequences of the Kogelberg Biosphere reserve sequences – Red. Reference sequences are coloured according to their taxonomy: *Baculoviridae* – light green, *Myoviridae* – yellow, *Phycodnaviridae* – Dark green, *Thermoproteaceae* – Pink, *Cenarcheaceae* – Mustard, *Archaeoglobaceae* – Purple, *Methanosarcinaceae* – Turquoise, and *Methanosaetaceae* – Blue. Sequences were aligned with de novo assembly using CLC genomics workbench version 6.0.1 (CLC, Denmark) and visualised with MetaVir server (Roux, Tournayre, *et al.*, 2014). Each tree is computed with 100 bootstraps, and the resulting values are indicated for each node. Scale bar indicates the number of substitutions per site.

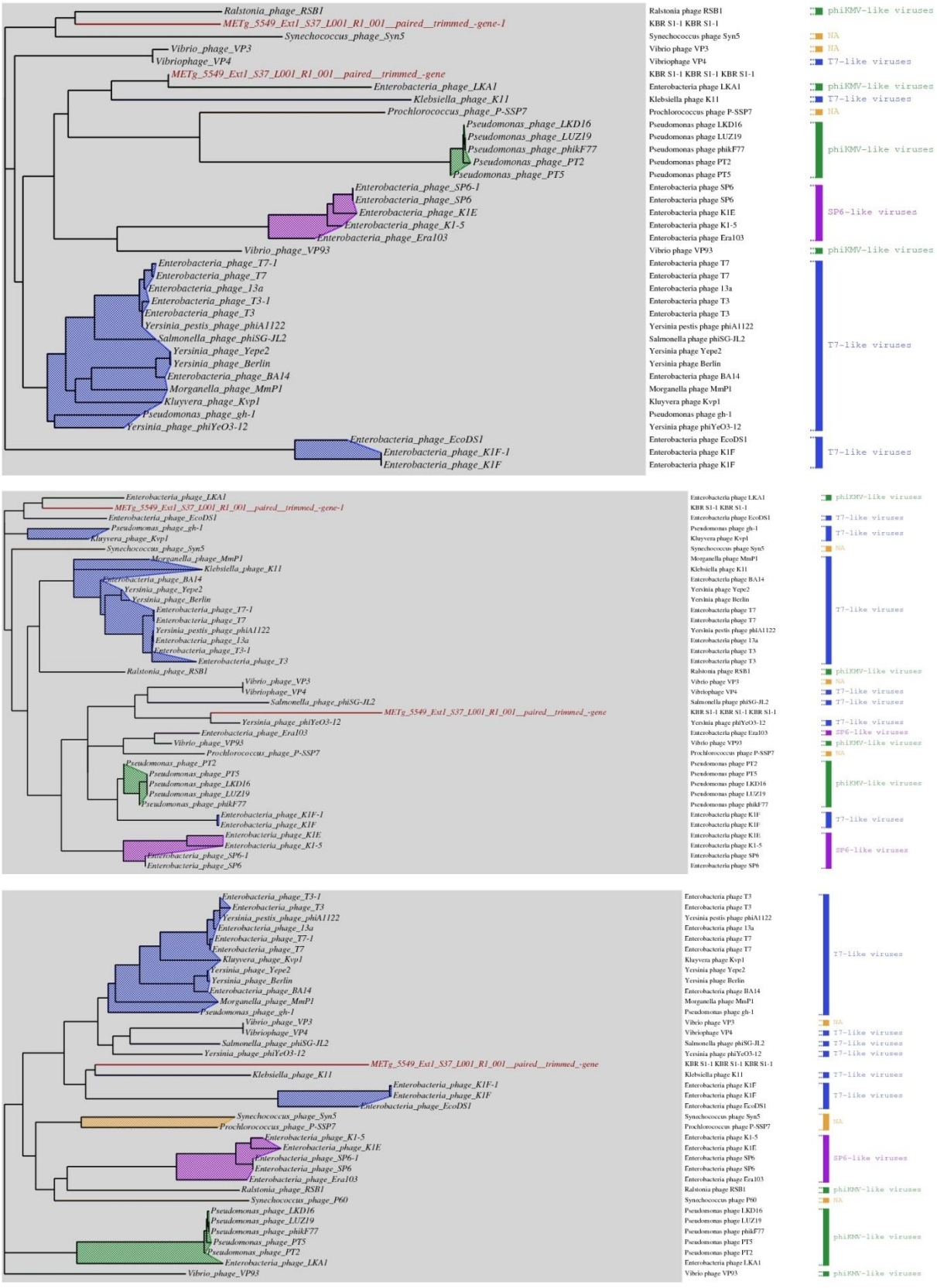
APPENDIX

Supplementary Figure S3: Phylogenetic tree of Pol B1 Marker gene amino acid sequences of the Kogelberg Biosphere reserve sequences – Red. Reference sequences are coloured according to their taxonomy: *Podoviridae* – Pink, *Ampullariidae* – Purple,

APPENDIX

***Tectiviridae* – green, *Adenoviridae* – Blue, Not assigned any Family – Yellow. Sequences were aligned with de novo assembly using CLC genomics workbench version 6.0.1 (CLC, Denmark) and visualised with MetaVir server (Roux, Tournayre, *et al.*, 2014). Each tree is computed with 100 bootstraps, and the resulting values are indicated for each node. Scale bar indicates the number of substitutions per site.**

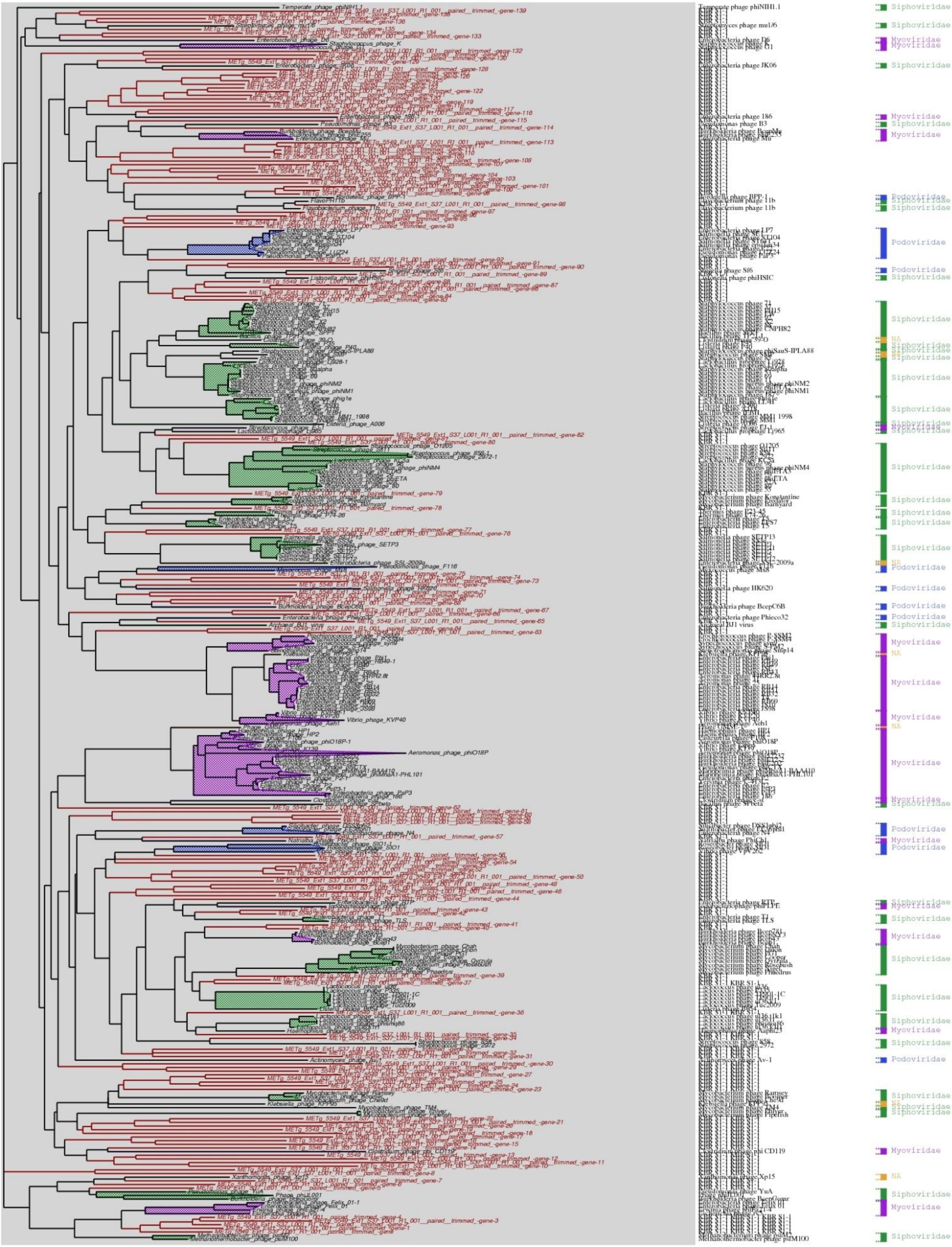
APPENDIX



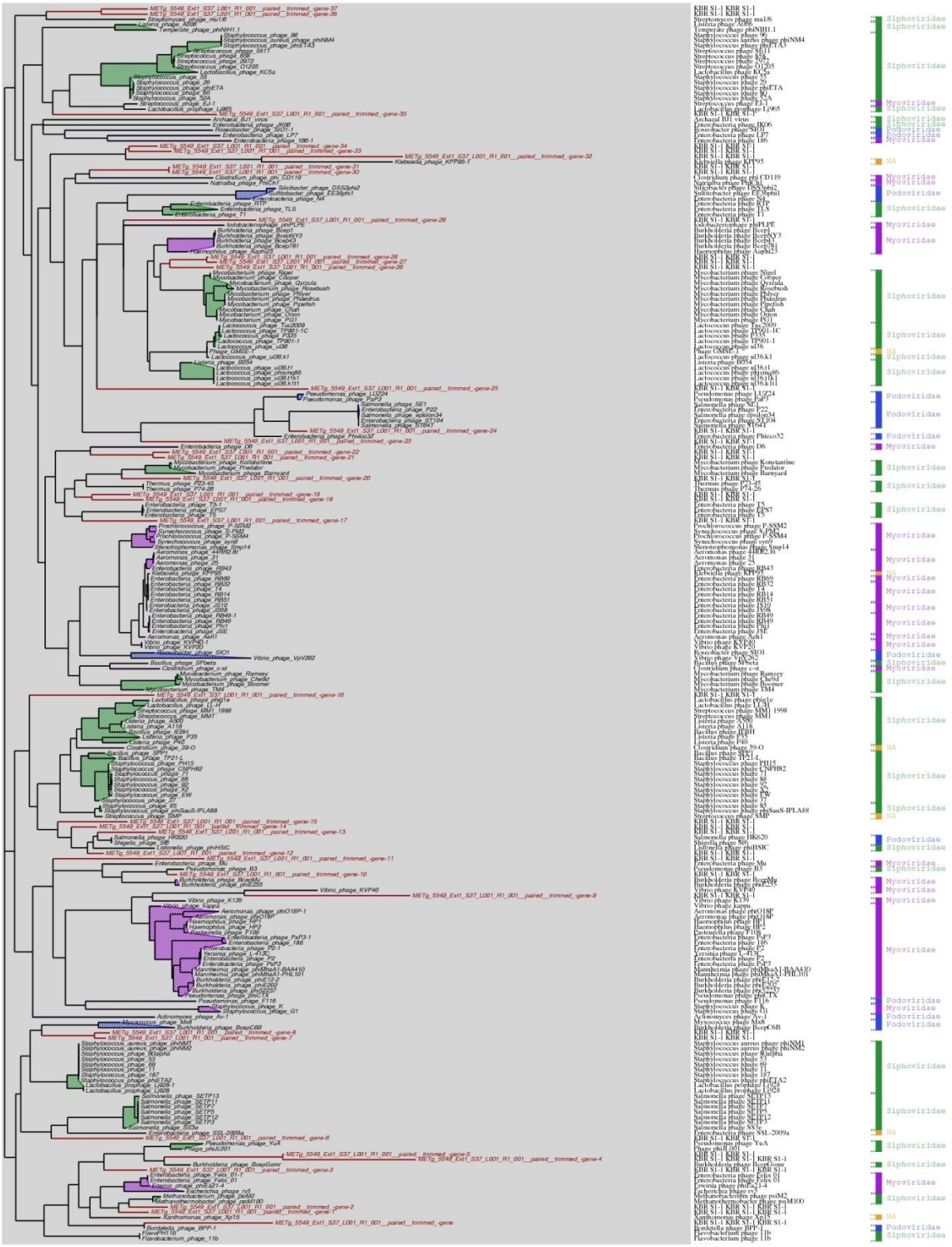
Supplementary Figure S4: Phylogenetic tree of T7gp17 Marker gene amino acid sequences of the Kogelberg Biosphere reserve sequences – Red. Reference sequences are

coloured according to their taxonomy: phiKMV-like viruses – green, T7-like viruses – Blue, SP6-like viruses – Purple, not assigned any family. Sequences were aligned with de novo assembly using CLC genomics workbench version 6.0.1 (CLC, Denmark) and visualised with MetaVir server (Roux, Tournayre, *et al.*, 2014). Each tree is computed with 100 bootstraps, and the resulting values are indicated for each node. Scale bar indicates the number of substitutions per site.

APPENDIX



APPENDIX



Supplementary Figure S5: Phylogenetic tree of TerL Marker gene amino acid sequences of the Kogelberg Biosphere reserve sequences – Red. Reference sequences are coloured according to their taxonomy: *Siphoviridae* – Green, *Myoviridae* – Purple, *Podoviridae* –

APPENDIX

Blue, Not assigned family – yellow Sequences were aligned with de novo assembly using CLC genomics workbench version 6.0.1 (CLC, Denmark) and visualised with MetaVir server (Roux, Tournayre, *et al.*, 2014). Each tree is computed with 100 bootstraps, and the resulting values are indicated for each node. Scale bar indicates the number of substitutions per site.

Supplementary Table S1: Comparisons of functional and taxonomic analysis between VIROME and MetaVir

Functional analysis	VIROME	MetaVir
Hypothetical proteins	1359	46457
glycoside hydrolase	124	83
gp11	11	39
YapH protein	1	2
HNH homing endonuclease	6	26
Helicase	231	277
Primase	77	93
pyrophosphatase	2	1
DNA polymerase	417	287
terminase large subunit	95	316
Lysine	19	95
endonuclease	117	198
structural protein	61	101
phage-related protein	52	7

APPENDIX

Taxonomic analysis	VIROME	MetaVir
Caudovirales	1118	25922
Phycodnaviridae	20	499
Ampullaviridae	1	173
Mimiviridae	1	110

Supplementary Table S2: Sample description of taxonomic abundance of the 10 largest contigs

Contigs name	Contigs length	#of predicted genes	Predicted genes	#Hypothetical proteins
5549_Ext1_S37_L001_R1_001__paired__trimmed__p_414	47854	63	terminase, glucosaminidase, phage_integrase	57
5549_Ext1_S37_L001_R1_001__paired__trimmed__p_53	44760	58	putative terminase large subunit, Peptidase_S74, endolysin, Lipase, putative DNA polymerase, putative DNA helicase putative adenine methyltransferase	48
5549_Ext1_S37_L001_R1_001__paired__trimmed__p_458	42564	61	putative endonuclease, putative terminase large subunit	52
5549_Ext1_S37_L001_R1_001__paired__trimmed__p_185	42057	53	putative phage terminase large subunit, putative peptidase, putative lytic tail protein, putative endolysin, putative DNA	27

			methyltransferase, putative DNA helicase, putative RecB family exonuclease, putative primase, putative DNA polymerase	
5549_Ext1_S37_L001_R1_001__paired__trimmed__p_9	41177	68	putative endolysin, putative terminase large subunit	53
5549_Ext1_S37_L001_R1_001__paired__trimmed__p_90	40336	49	putative terminase large subunit, putative endoprotease, putative major capsid protein, putative endolysin, putative DNA polymerase	33
5549_Ext1_S37_L001_R1_001__paired__trimmed__p_259	36396	45	Pectate_lyase, Endonuclease, putative terminase large subunit	39
5549_Ext1_S37_L001_R1_001__paired__trimmed__p_645	36310	87	RNA_ligase, Endonuclease, putative deoxycytidylate deaminase, putative PseT polynucleotide 5'-kinase and 3'-phosphatase, putative RNA ligase	67
5549_Ext1_S37_L001_R1_001__paired__trimmed__p_929	34830	43	putative phage tail fiber-like protein, putative tail fiber protein, putative N-acetylmuramoyl-L-alanine amidase, putative DNA polymerase B region, putative tRNA ribotransferase, putative GTP cyclohydrolase, putative glutamine amidotransferases class-	21

			II, putative organic radical activating enzyme	
5549_Ext1_S37_L001_R1_001__paired__trimmed__p_416	30995	34	putative Mu-like prophage I protein, putative tape measure protein, Pectate_lyase, putative DNA-binding protein	24

Supplementary Table S3: Selected Biomes for viral composition comparison with KBR

Table discription	Sample name	Project ID	Sample type	References
KBR (s)	KBR sample 1	5549	soil	Current study
AOS (s)	Antarctic open soil contigs	2473	soil	(Zablocki *et al.*, 2015)
RF(s)	RF Peru	4906	Soil	(N Fierer *et al.*, 2007)
AH	Antarctic hypolith contigs	2472	hypolith	(Zablocki *et al.*, 2015)
NH	Namib hypolith contigs	2186	hypolith	(Adriaenssens *et al.*, 2015)
Far (fw)	Far-T4 Lake Pavin	5127	fresh water	(S. Roux *et al.*, 2015)
LB (fw)	Lake Bourget	1327	fresh water	(Roux *et al.*, 2012)
LP (fw)	Lake Pavin	1328	fresh water	(Roux *et al.*, 2012)

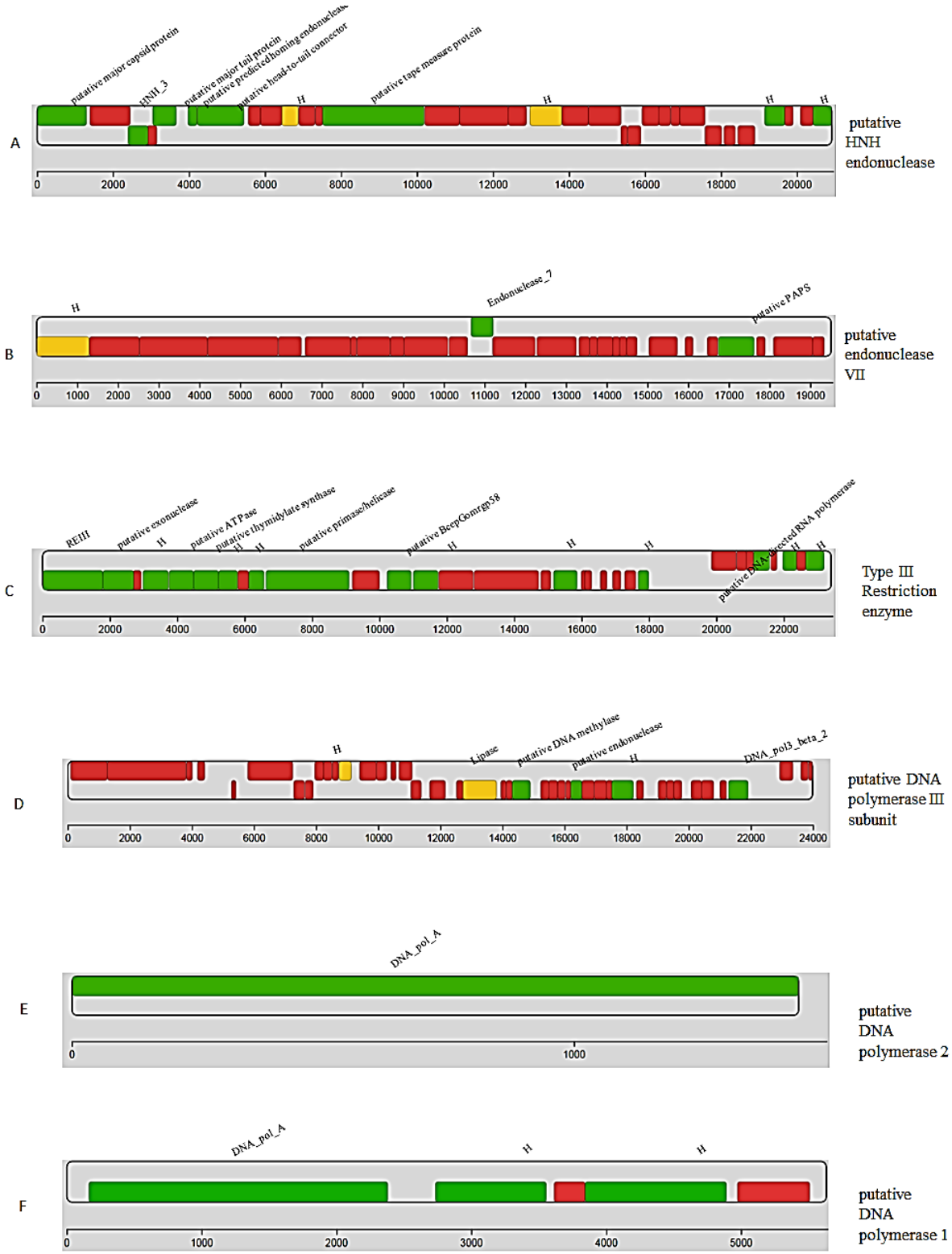
APPENDIX

57th (fw)	57th_St_05-Jun-13	3305	fresh water	(Siobhan C. Watkins *et al.*, 2016)
M1 (fw)	Montrose_05-Jun-13-1	3306	fresh water	(Siobhan C. Watkins *et al.*, 2016)
M2 (fw)	Montrose_25-Jun-13-2	3307	fresh water	(Siobhan C. Watkins *et al.*, 2016)
SP (p)	Salted pond	25	pond	(Rodriguez-Brito *et al.*, 2010)
ALOHA (ds)	ALOHA_station_deep_abyss	3816	deep sea	(Angly *et al.*, 2006)
B47 (ds)	B47_Bohai_Sea_Sep_2010	5754	sea	(Angly *et al.*, 2006)
P	P._acuta_2012	2315	unknown	(Wood-Charlson *et al.*, 2015)
Sup05	Sup05_prophage_contigs	3232	unknown	(Roux, Hawley, *et al.*, 2014)
VS	VirSorter_curated_dataset	5062	unknown	(Simon Roux *et al.*, 2015)
10eld	10_eld_contigs		unknown	(M.-S. Kim *et al.*, 2011)

APPENDIX

Appendices D: Chapter 5

Appendices D1: Sequence screening contig maps



APPENDIX

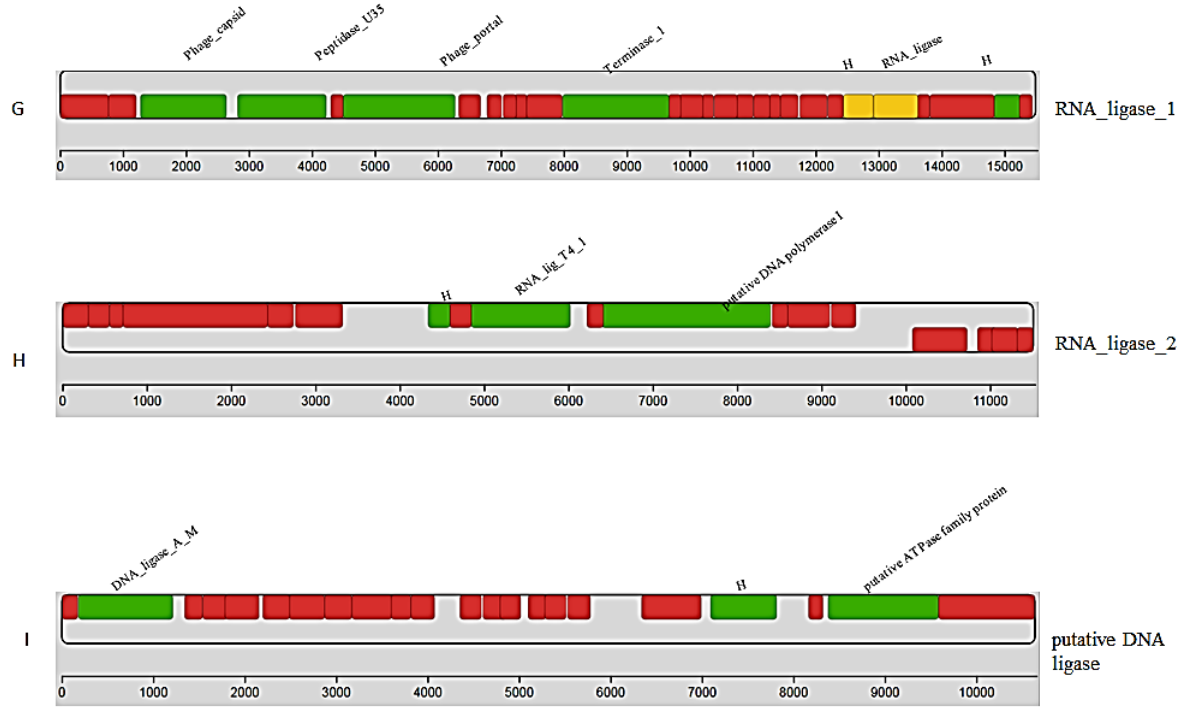
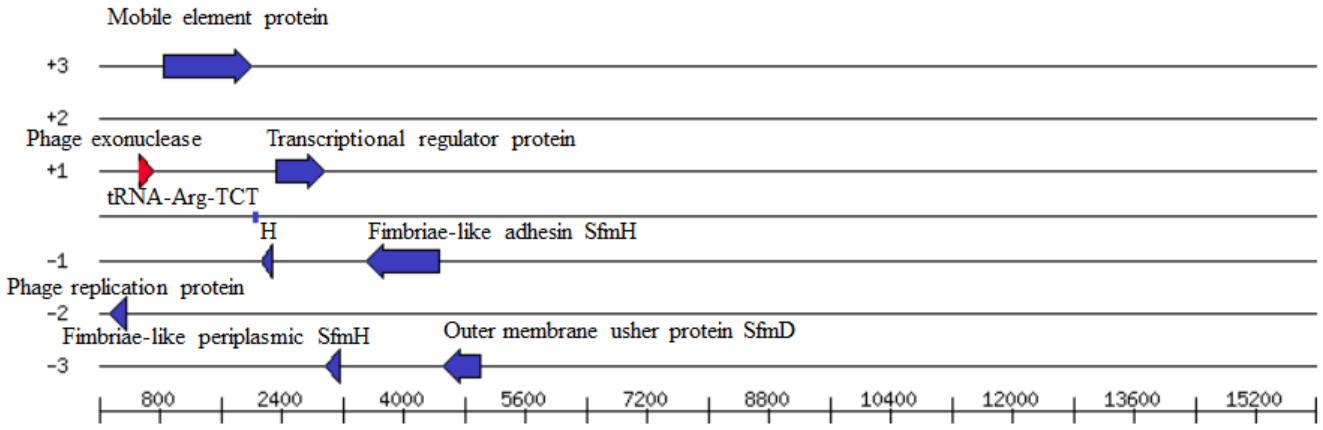


Figure D1 (A, B, C, D, E, F, G, H, I): Contigs maps of the Kogelberg Biosphere Reserve. Each map is drawn using RaphaelSvg, and presents the different predicted genes and their affiliation. The affiliations are assessed from a BLASTp comparison against RefseqVirus (threshold of 10-3 on e-value and 50 on bitscore), and an hmmscan comparison to the PFAM database (26.0, July 2012; threshold of 30 on score). Each gene is coloured according to its level of annotation: Green for genes with a Refseq significant hit, Yellow for genes with no Refseq hit but a PFAM affiliation, and Red for genes with neither Refseq nor PFAM hit.

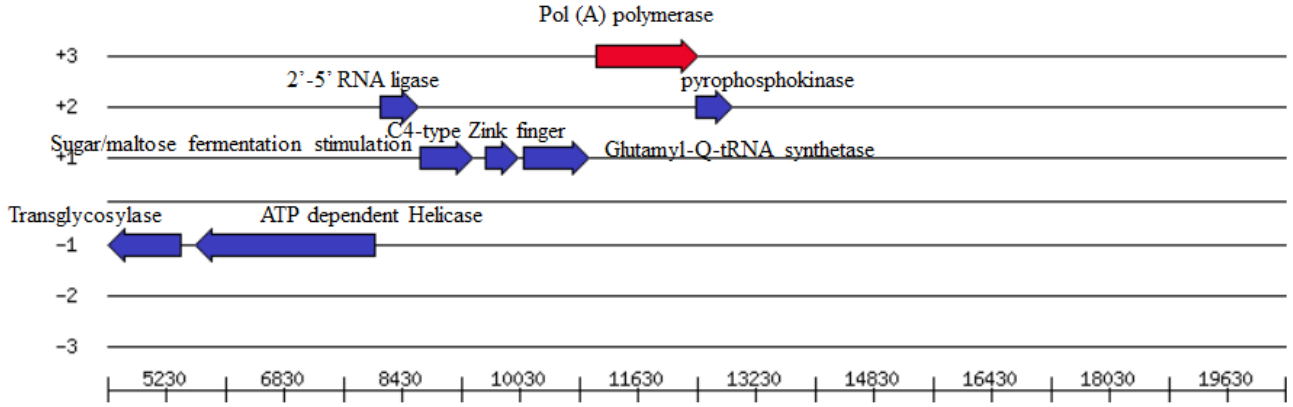
APPENDIX

Appendices D2: Functional screening contig maps

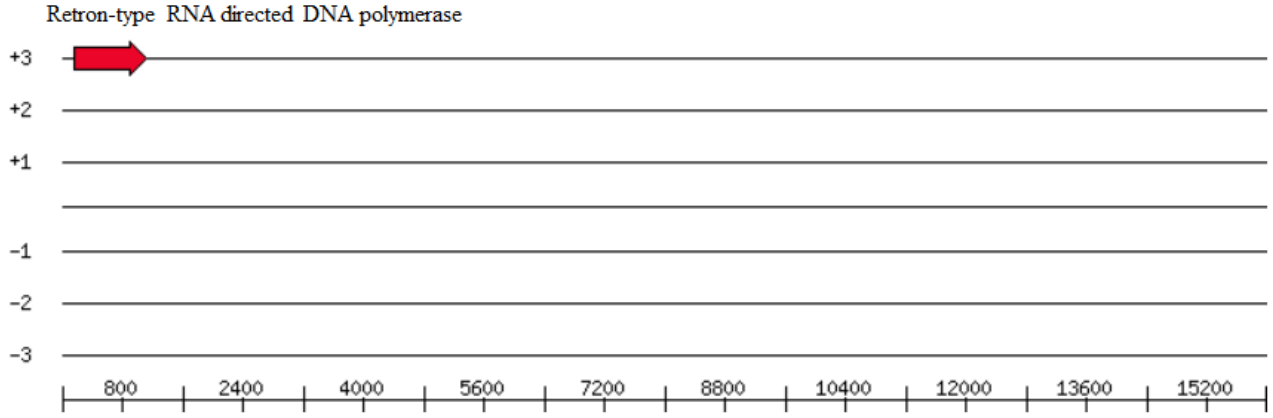
1. Contig 213



2. Contig 126

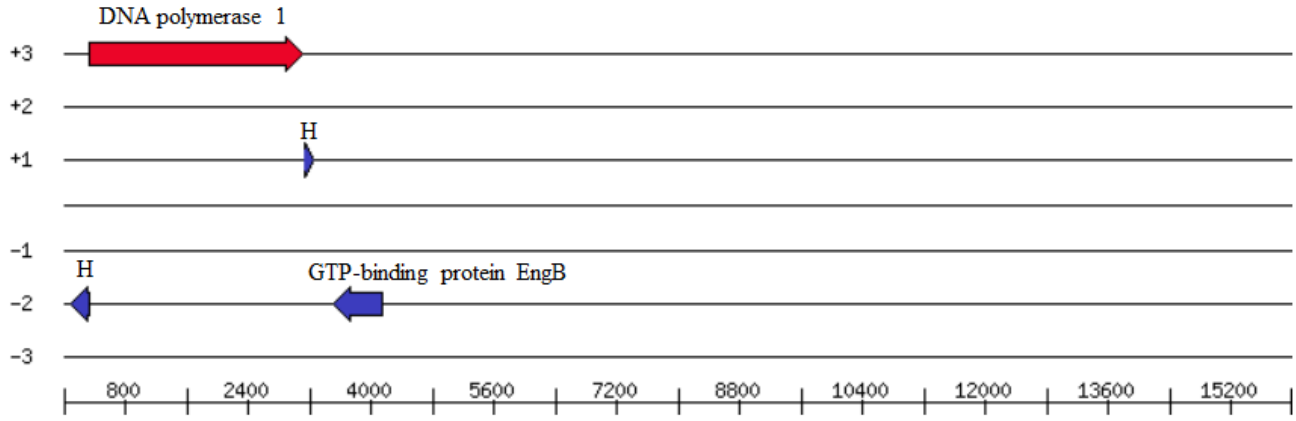


3. Contig 539



4. Contig 101

APPENDIX



5. Contig 70

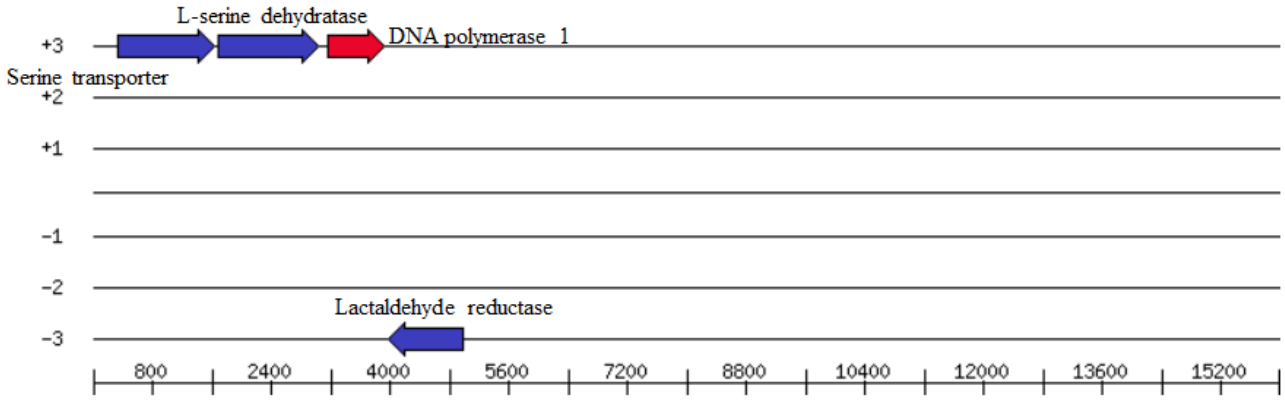


Figure D2 (A, B, C, D, E, F, G, H, I): Contigs maps of the Kogelberg Biosphere Reserve *fynbos* soil metavirome sequences from functional screening. Each map presents the different predicted genes and their affiliation. The affiliations are assessed from a SEED database using RAST server. Each gene encoding putative 5'-3' exonuclease domain is coloured red. Blue colour is for other genes flanking the putative 5'-3' exonuclease domain of the DNA polymerase 1.

APPENDIX

Section D3: Expression optimisation SDS-PAGE gels

Expression optimisation time profiles growth conditions are as follows: LB medium, at 37, 25, 30 and 16°C, 1mM IPTG over 24 hr growth period), sampled at the following time intervals (T0, T1 =1h, T2=2h, T3=3h, T4 =4h, T5=5h, T6=6h and overnight = ON) resulted in no detectable protein bands of the expected size in soluble intracellular fractions. However, for the following expression constructs *PolB, HNHc, RNALig2, RE, PolA2* and *DNAlig* proteins bands corresponding with the expected molecular mass were detected in the insoluble fractions. There were no detectable protein bands in both soluble and insoluble *E. coli* fractions for *RNALig1, E7* and *PolA1* constructs (Figure D5, D6, D7 and D8).

Table D3: Representation of expression optimisation conditions

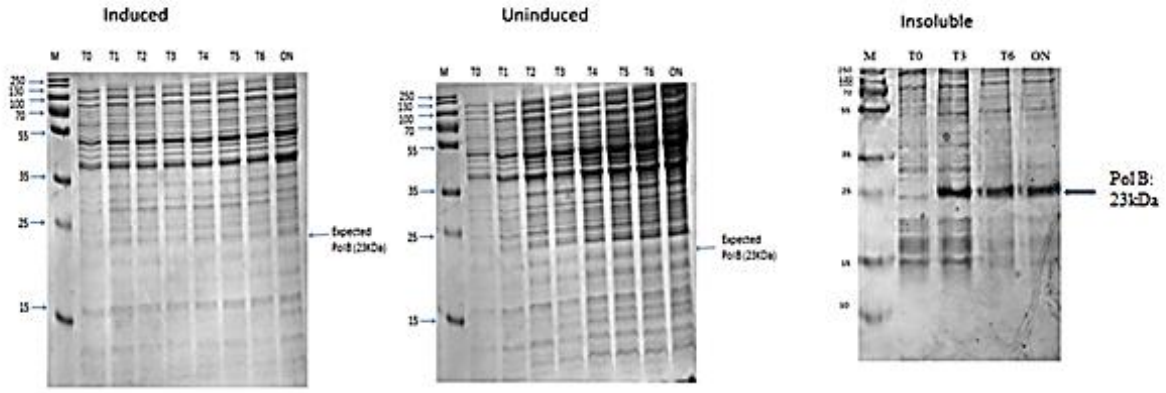
Conditions		pET20B(+)_PolB1		pET20(+)_HNHc		pET20b(+)_RNALig1		pET28B(+)_RNALig2		pET28B(+)_RE		pET28B(+)_E7		pET28B(+)_PolA1		pET30_PolA2		pET30_DNALig1	
		S	I	S	I	S	I	S	I	S	I	S	I	S	I	S	I	S	I
Standard conditions	37°C, 1mM IPTG, LB, BL21DE3	-	+	-	+	-	-	-	+	-	+	-	-	-	-	-	+	-	+
Temperature	30°C	-	+	-	+	-	-	-	+	-	+	-	-	-	-	-	-	-	+
	25°C	-	+	-	+	-	-	-	+	-	+	-	-	-	-	-	-	-	+
	18°C	-	+	-	+	-	-	-	+	-	+	-	-	-	-	-	-	+	+
IPTG (mM)	1mM	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+
	0.5mM	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+
	0.2mM	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+
	100uM	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+

Media	2×YT	-	-	-	-	-	-	-	+	-	-	-	-	+	+	-	+	+	+
	EnPresso ® B	-	-	-	-	-	-	-	+	-	+	-	+	+	+	-	-	+	+
***E. coli* host**	BL21 DE3	-	+	-	+	-	-	-	+	-	+	-	-	+	+	-	+	+	+
	BL21 DE3 pLysS	-	-	-	-	-	-	-	+	-	-	-	-	-	+	-	+	+	+
	BL21 AI	-	-	-	-	-	-	-	+	-	+	-	-	+	+	-	-	+	+
Tags	MBP-tag	NA	Na	Na	Na	Na	Na	Na	Na	Na	Na	Na	Na	+	+	Na	Na	+	+

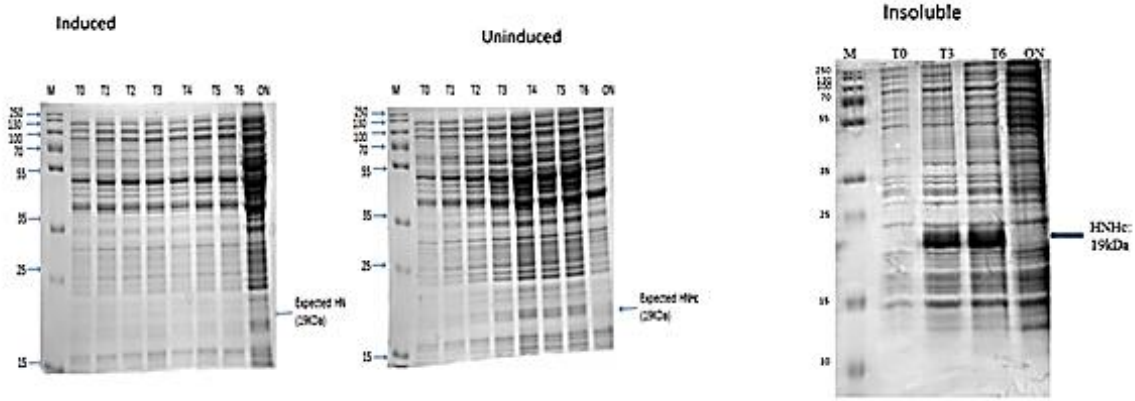
*NA means not applicable.

Expression @ 37 °C, 1mM IPTG

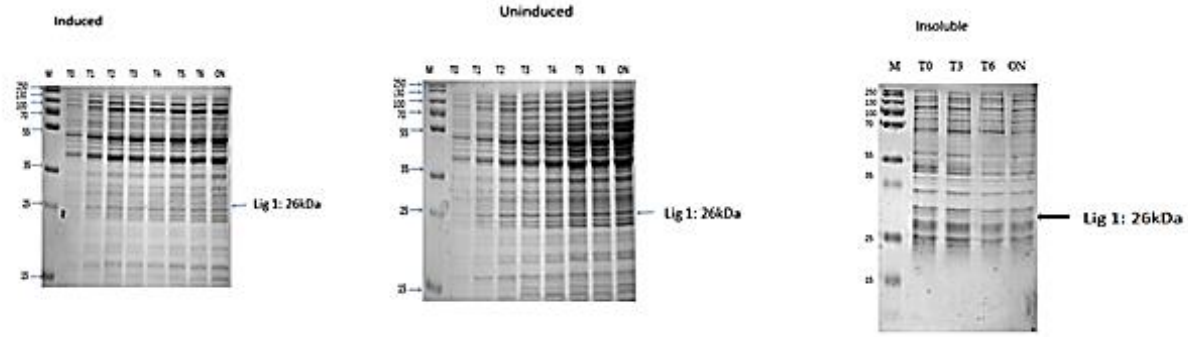
A: Pol B1



B: HNHc

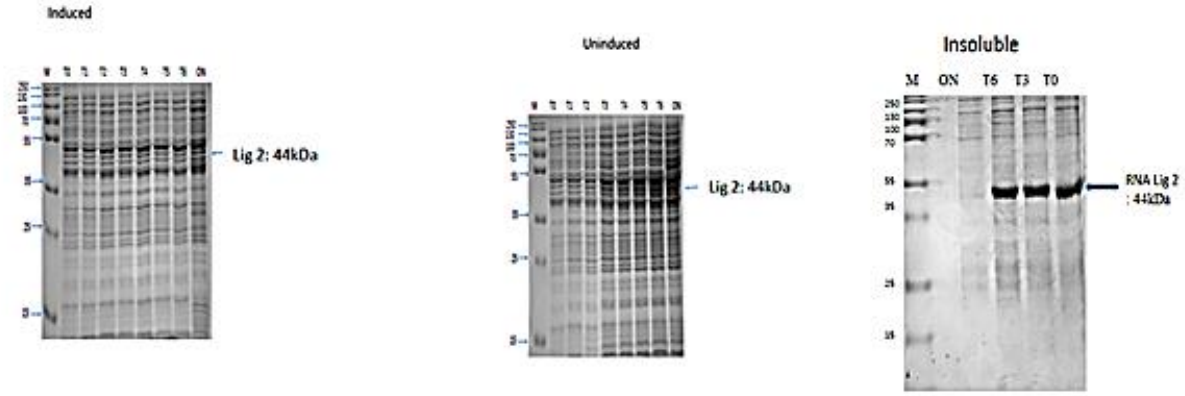


C: RNA Lig 1

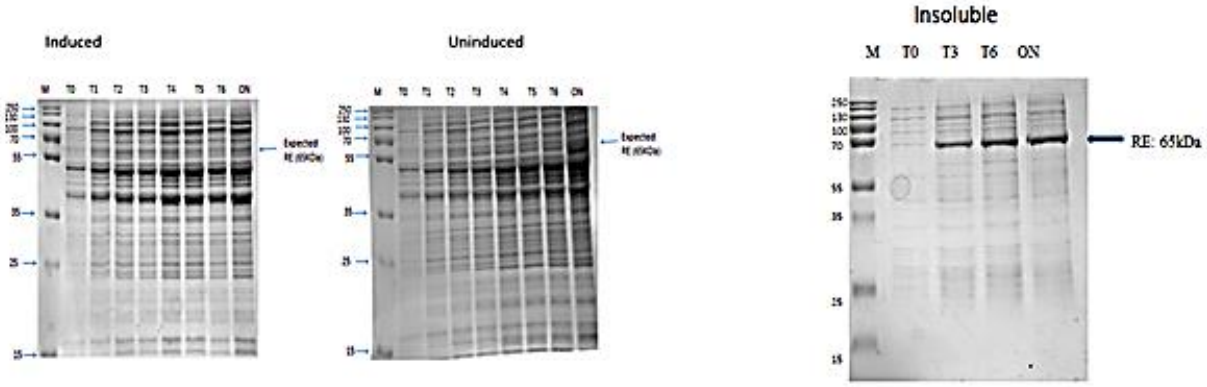


APPENDIX

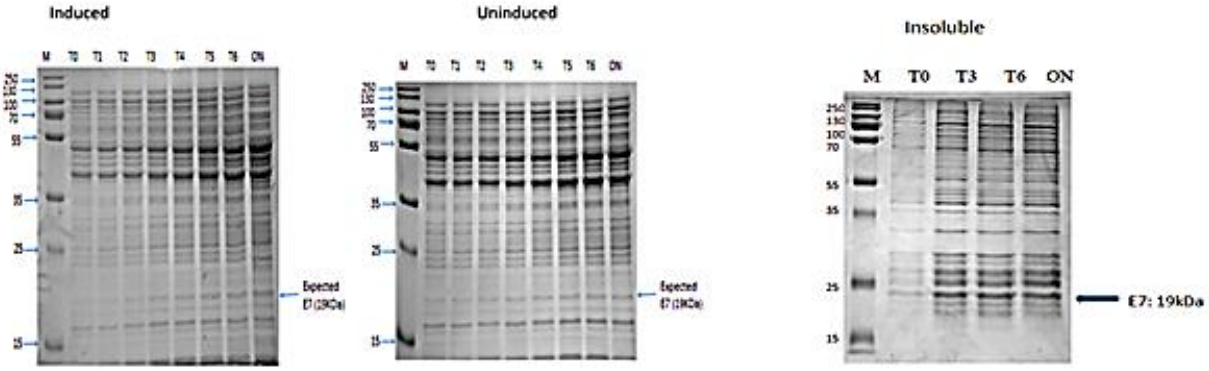
D: RNA Lig 2



E: RE

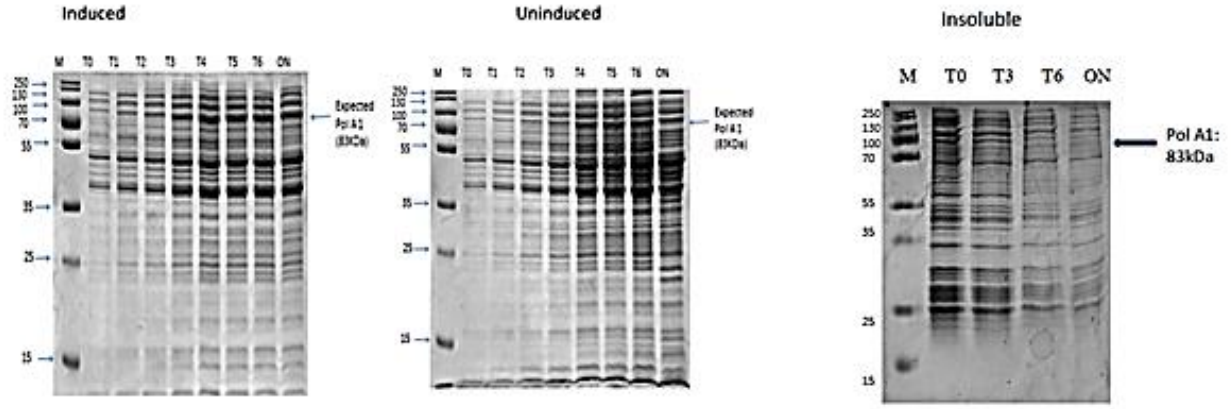


F: E7

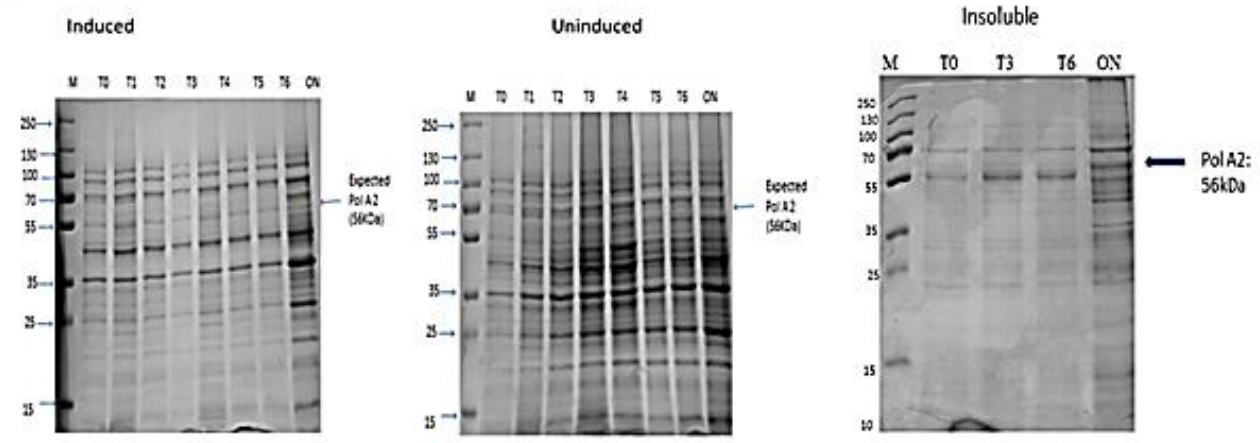


APPENDIX

G: Pol A1



H: Pol A2



I: DNA Lig

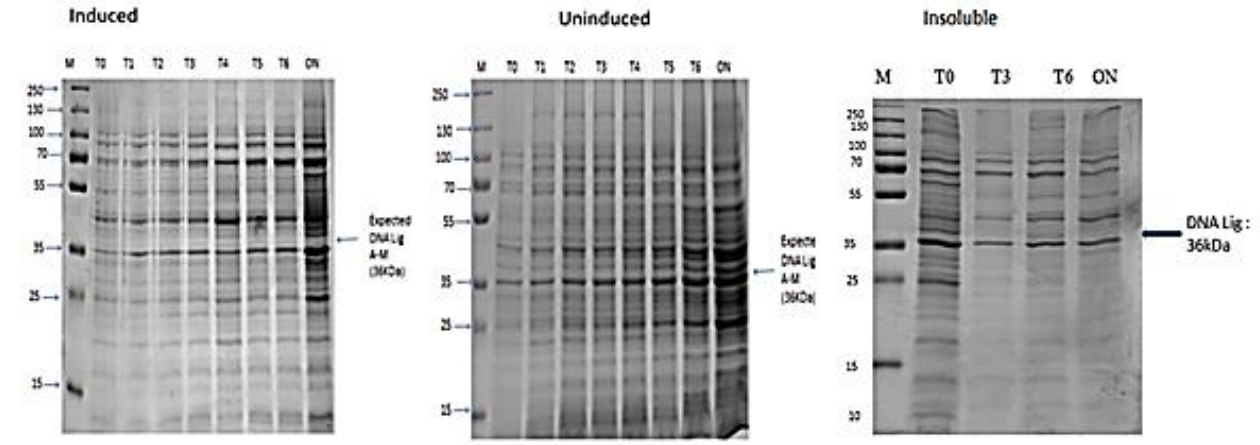
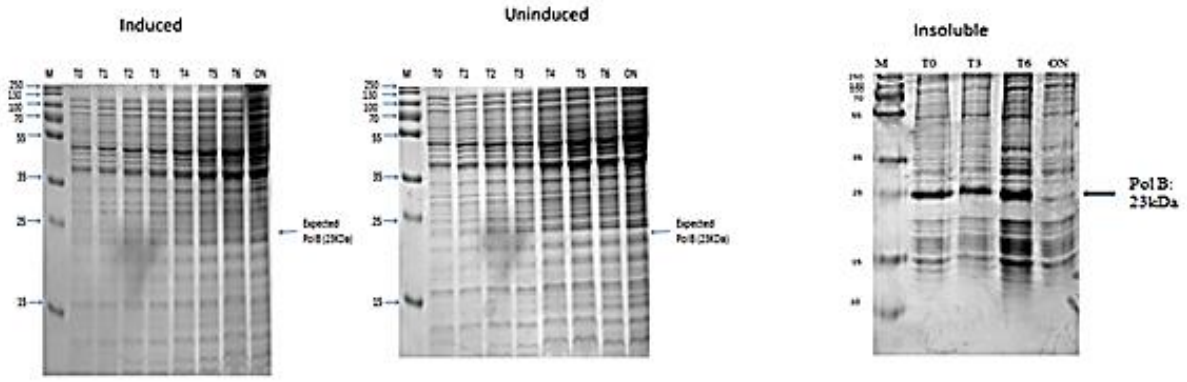


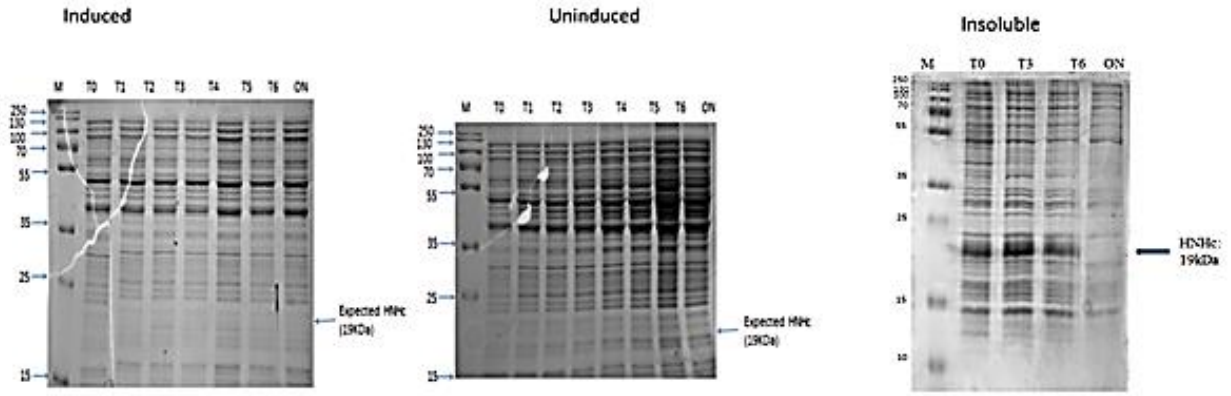
Figure D3 (A, B, C, D, E, F, G, H and I): SDS-PAGE analysis of recombinant genes expressed in LB at 37°C with 1mM IPTG concentration. A: pET20B(+)_PolB1, B: pET20(+)_HNHc, C: pET20b(+)_RNALig1, D: pET28B(+)_RNALig2, E: pET28B(+)_RE, F: pET28B(+)_E7, G: pET28B(+)_PolA1, H: pET30_PolA2 and I: pET30_DNALig1 recombinant genes with the expected protein band corresponding to a size of 23 kDa (Pol B), 19 kDa (HNH), 26 kDa (Lig 1), 44 kDa (Lig 2), 65 kDa (RE), 19 kDa (E7), 83 kDa (Pol A1), 56 kDa (Pol A2) and 36 kDa (DNAlig) indicated by an arrow.

Expression @ 30°C, 1mM IPTG

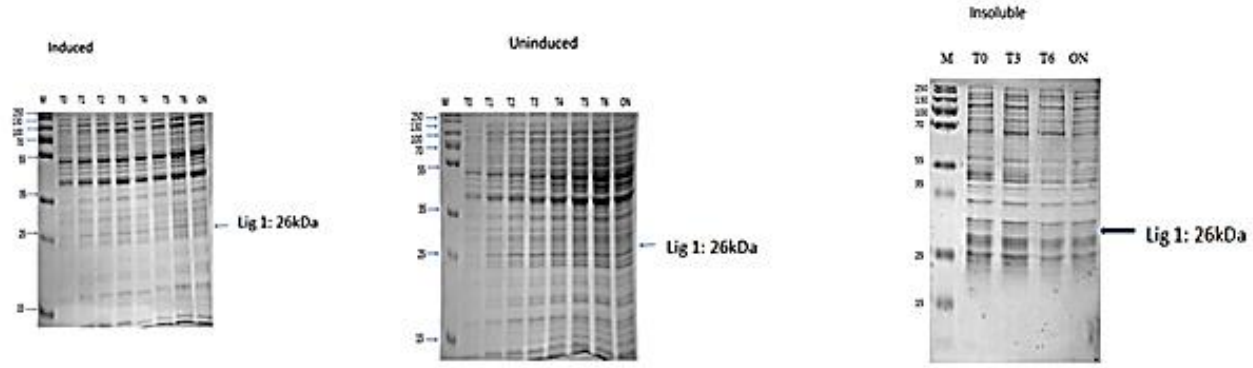
A: Pol B1



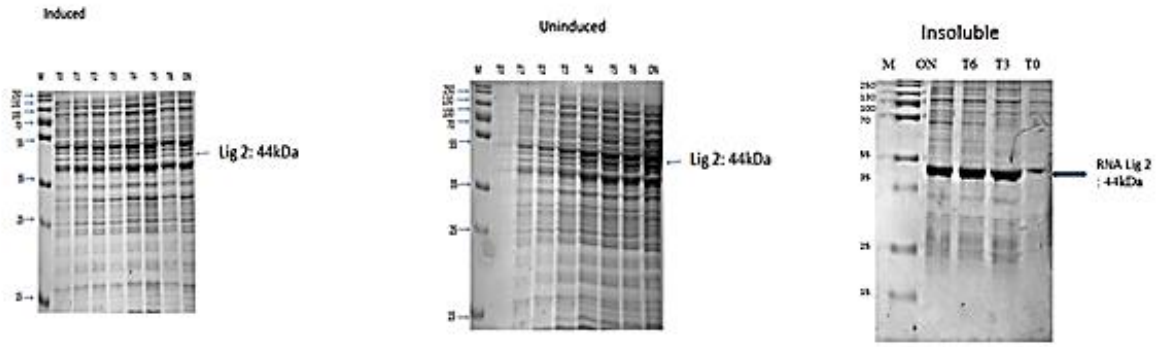
B: HNHc



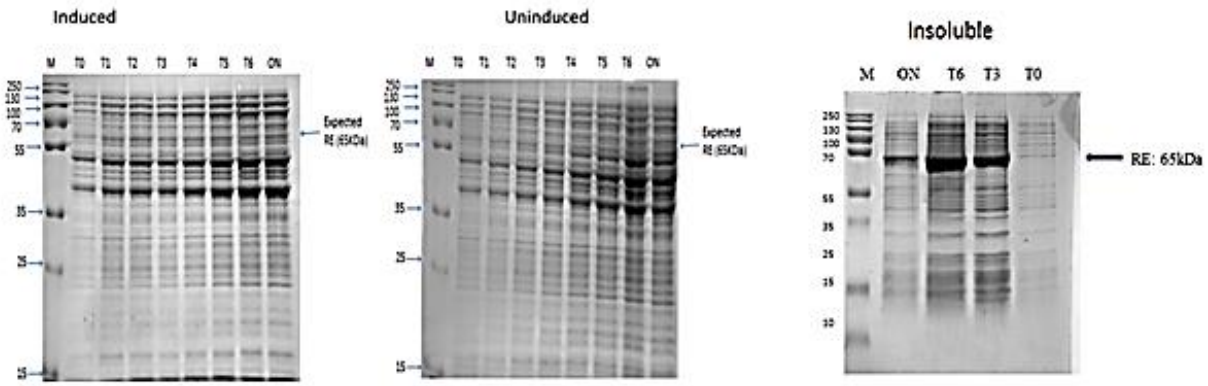
C: RNA Lig 1



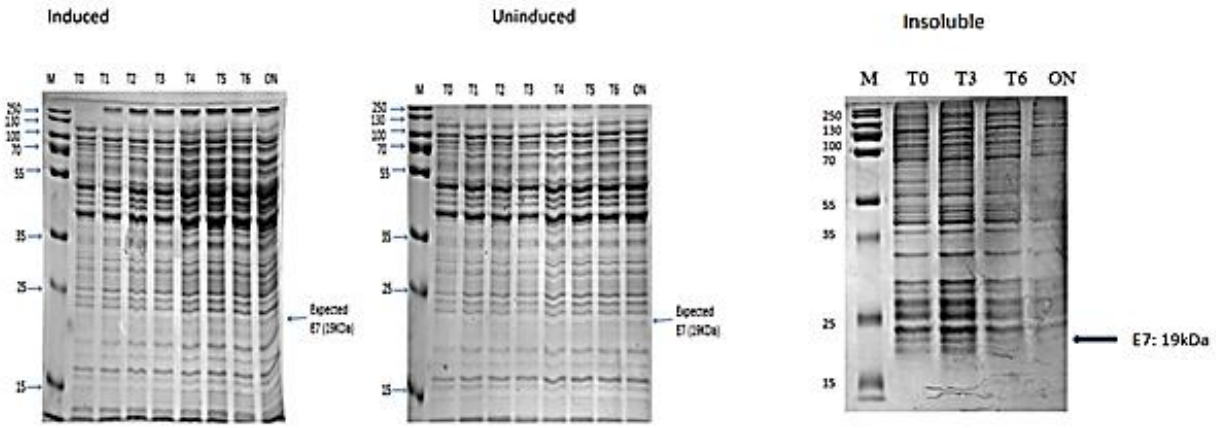
D: RNA Lig 2



E: RE

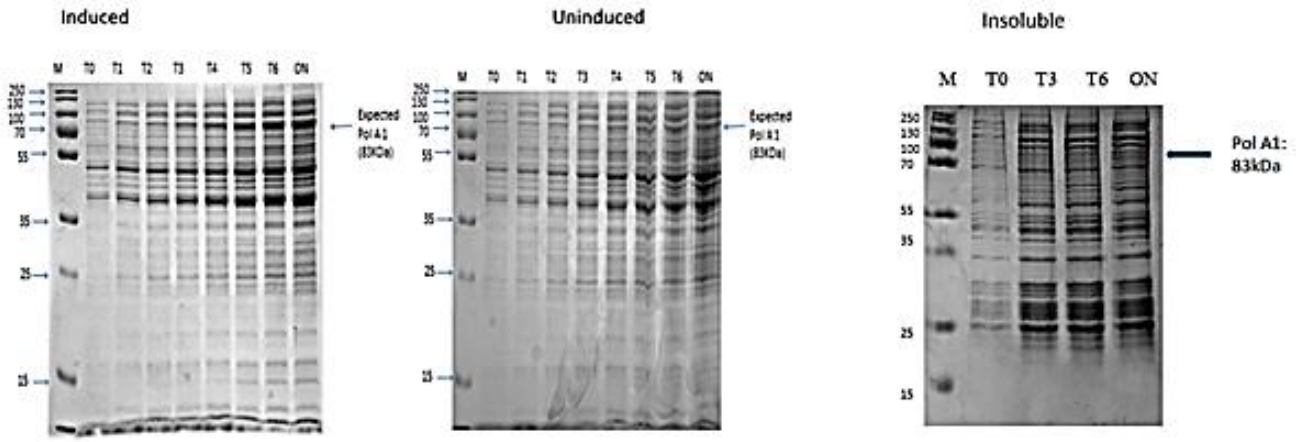


F: E7

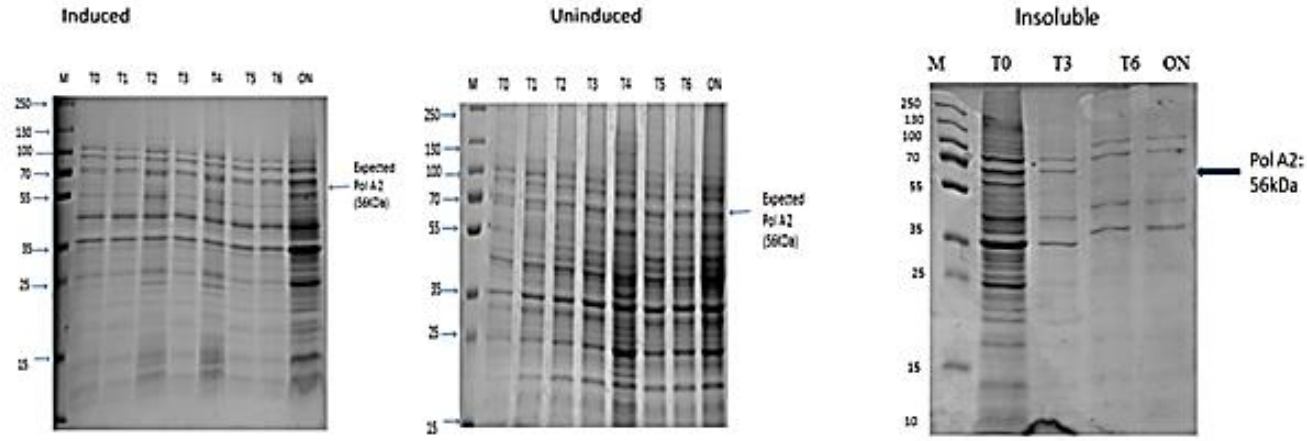


APPENDIX

G: Pol A1



H: Pol A2



I: DNA Lig

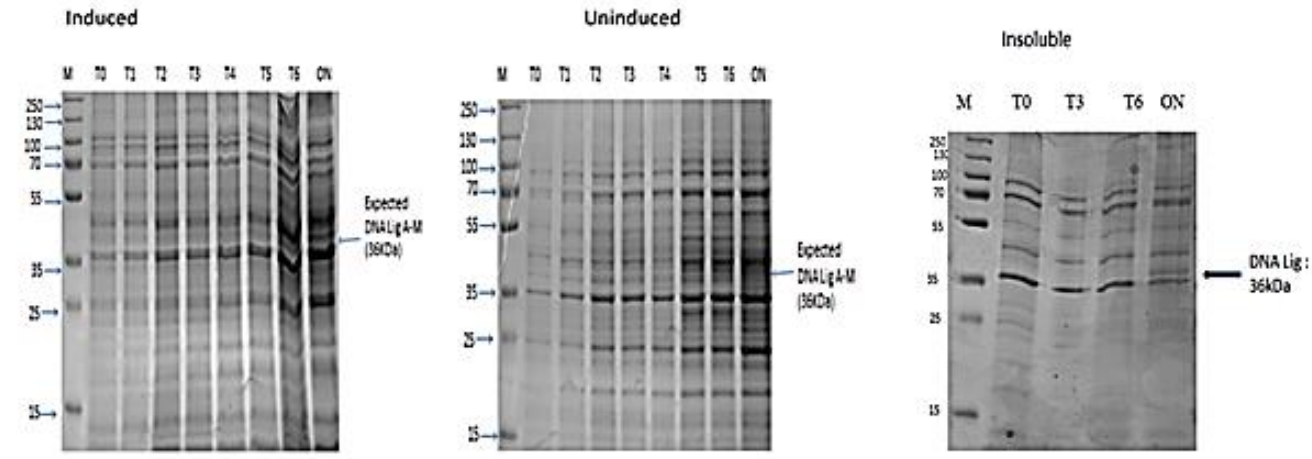
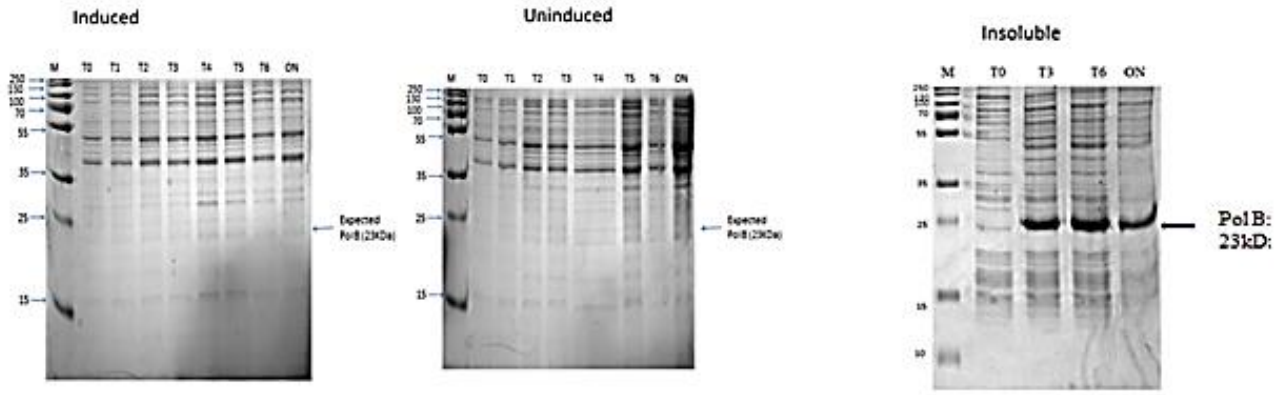


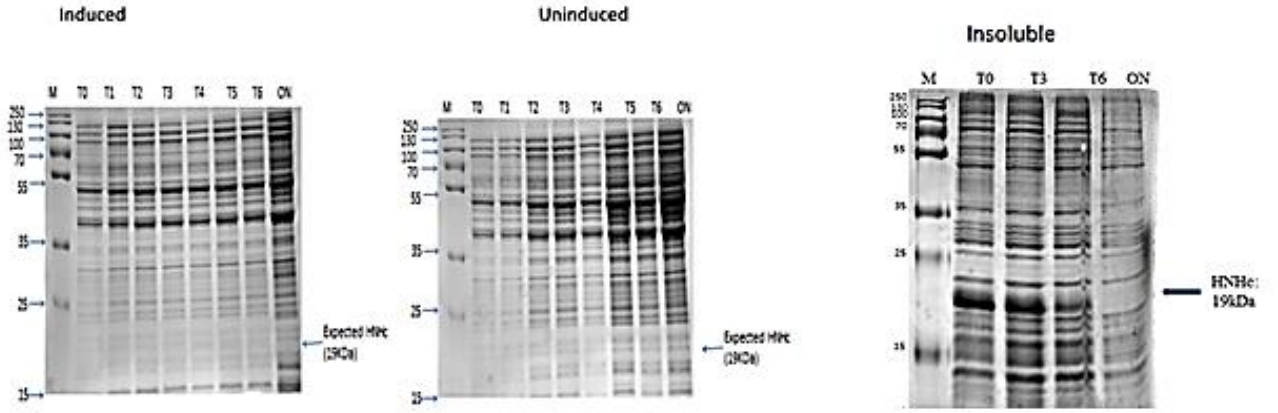
Figure D4 (A, B, C, D, E, F, G, H and I): SDS-PAGE analysis of recombinant genes expressed in LB at 30°C with 1mM IPTG concentration. A: pET20B(+)_PolB1, B: pET20(+)_HNHc, C: pET20b(+)_RNALig1, D: pET28B(+)_RNALig2, E: pET28B(+)_RE, F: pET28B(+)_E7, G: pET28B(+)_PolA1, H: pET30_PolA2 and I: pET30_DNALig1 recombinant genes with the expected protein band corresponding to a size of 23 kDa (Pol B), 19 kDa (HNH), 26 kDa (Lig 1), 44 kDa (Lig 2), 65 kDa (RE), 19 kDa (E7), 83 kDa (Pol A1), 56 kDa (Pol A2) and 36 kDa (DNAlig) indicated by an arrow.

Expression @ 25 C, 1mM IPTG

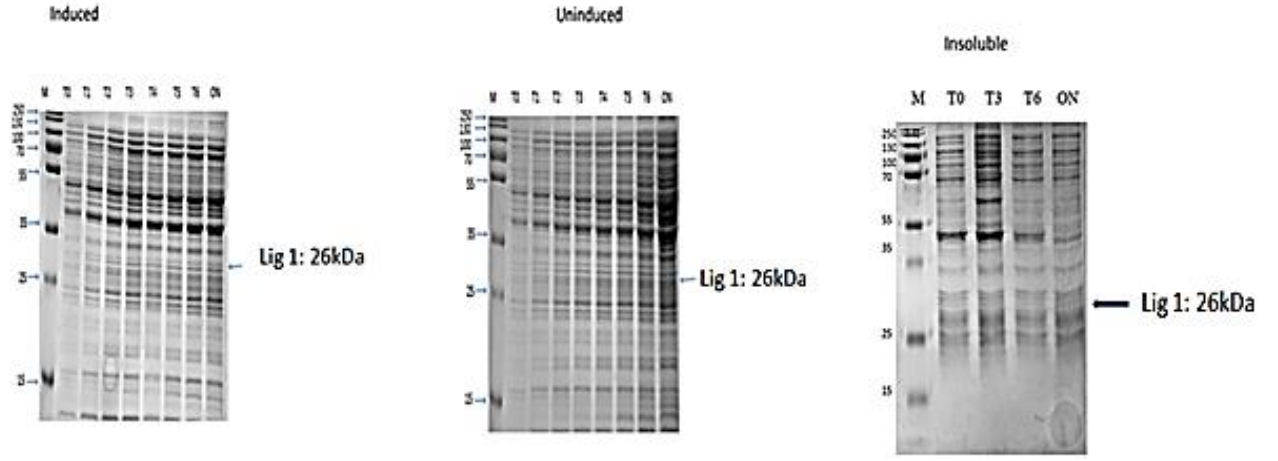
A: Pol B1



B: HNHc

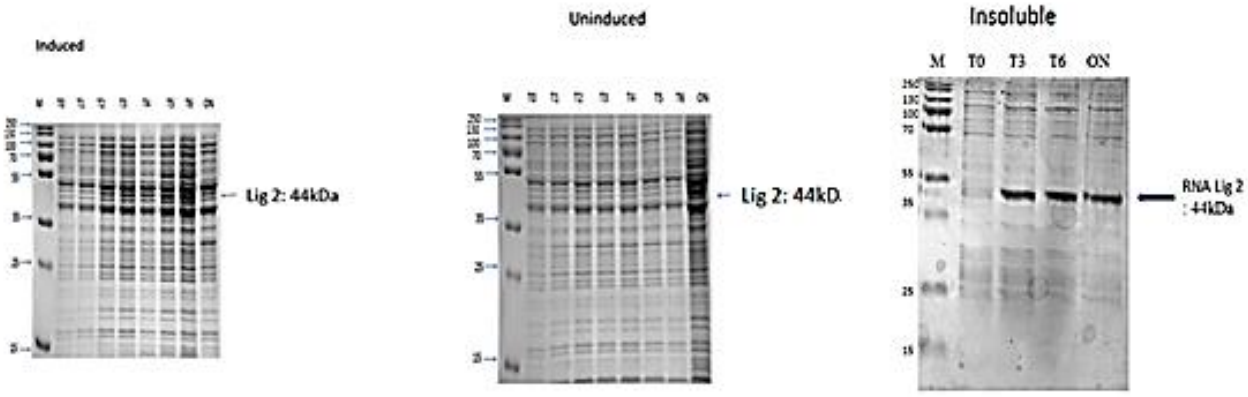


C: RNA Lig 1

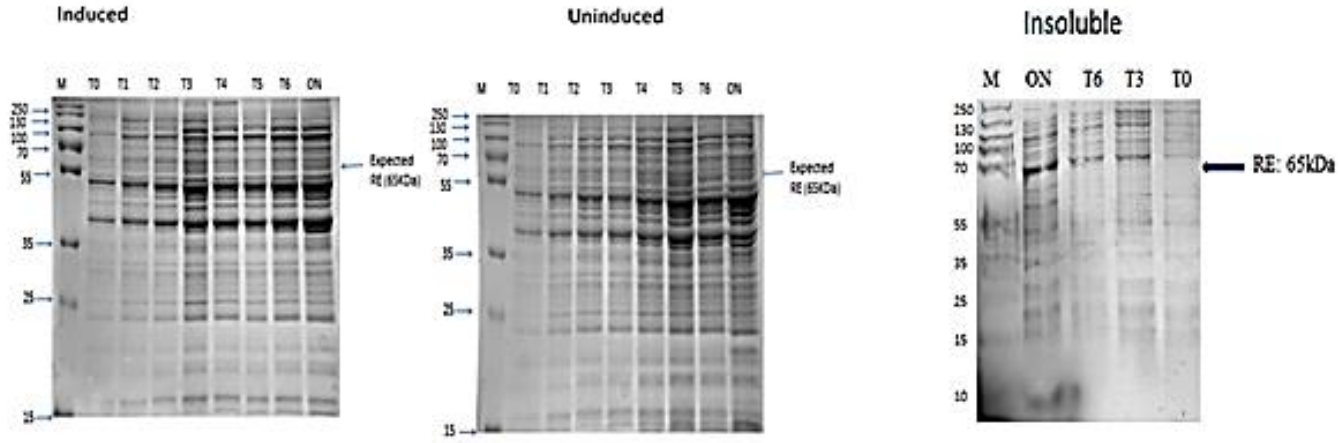


APPENDIX

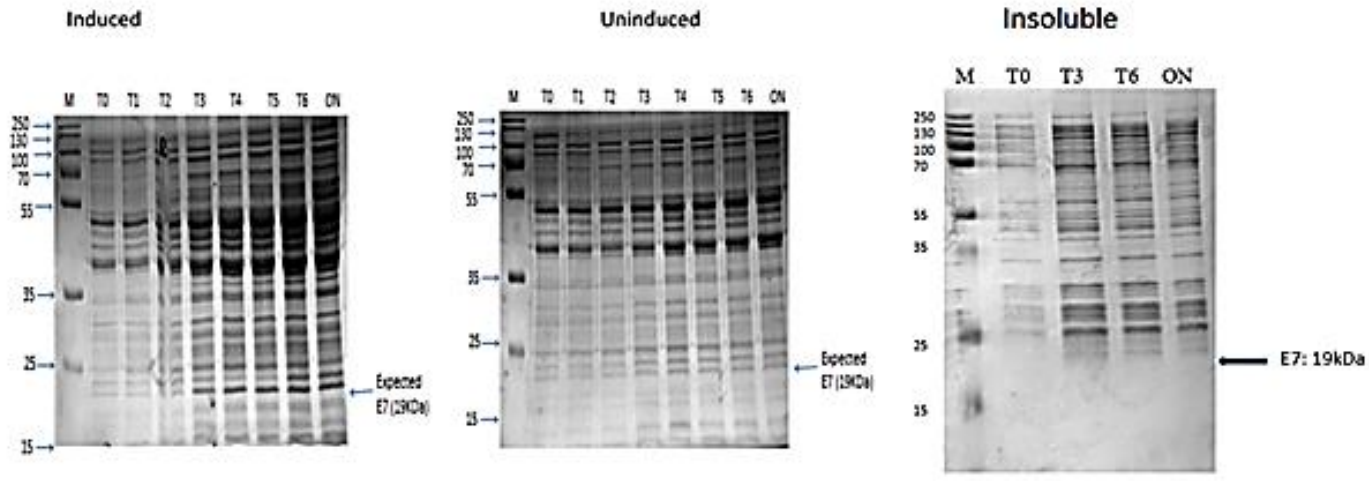
D: RNA Lig 2



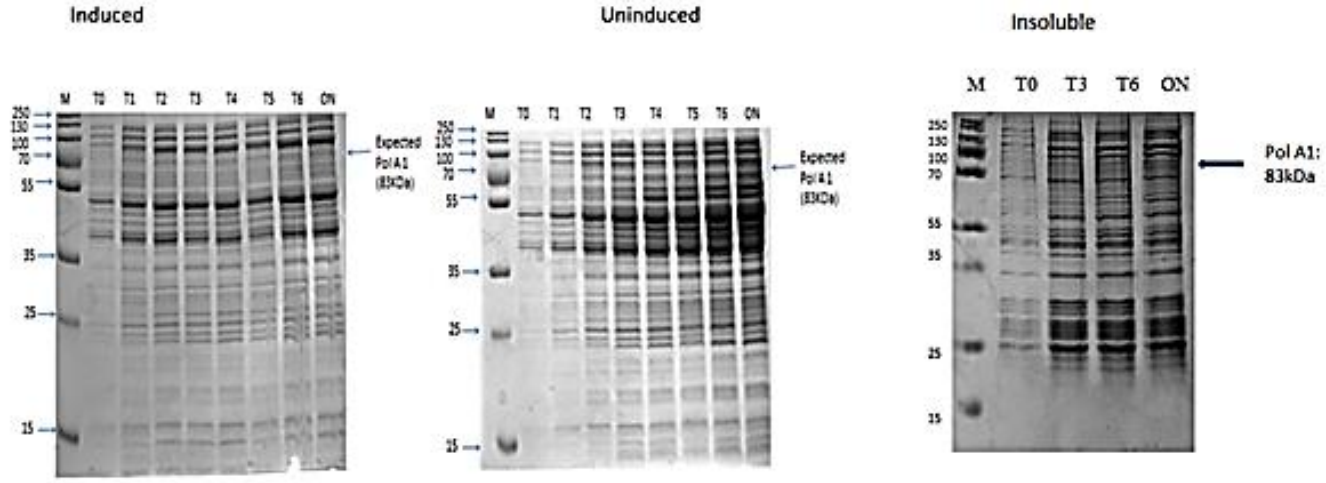
E: RE



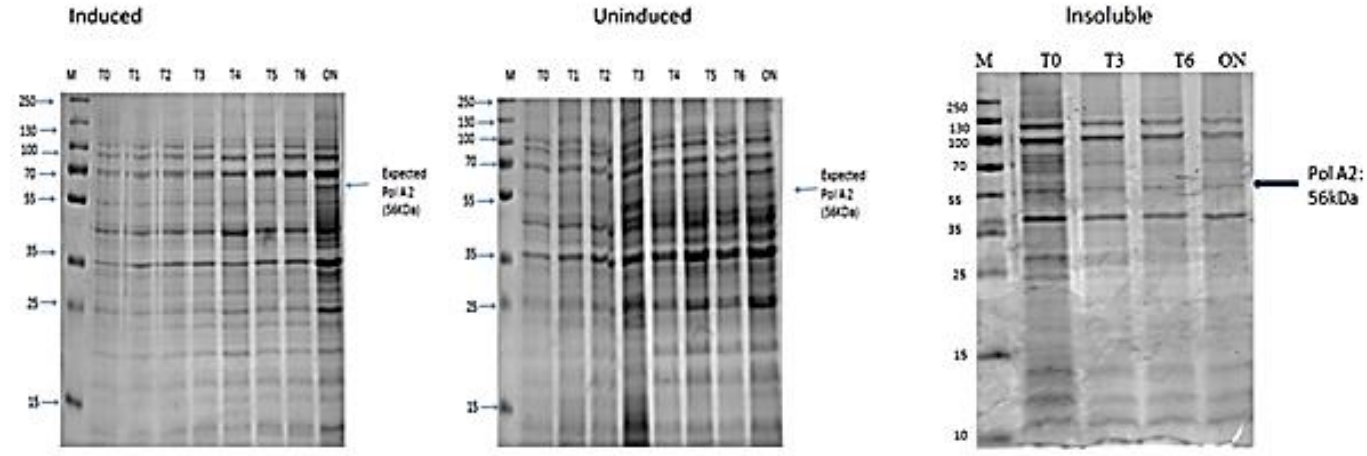
F: E7



G: Pol A1



H: Pol A2



I: DNA Lig

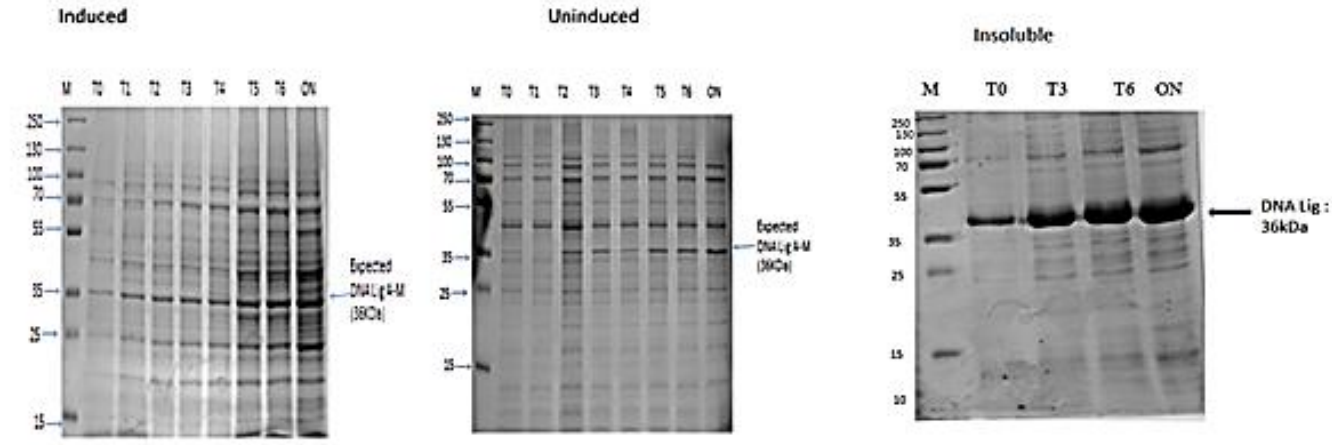
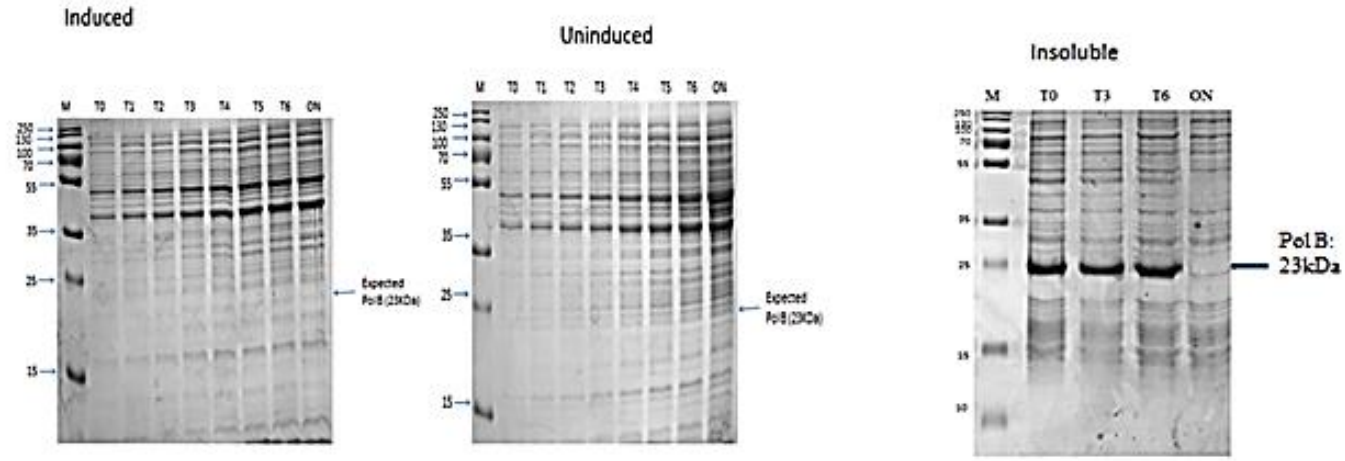


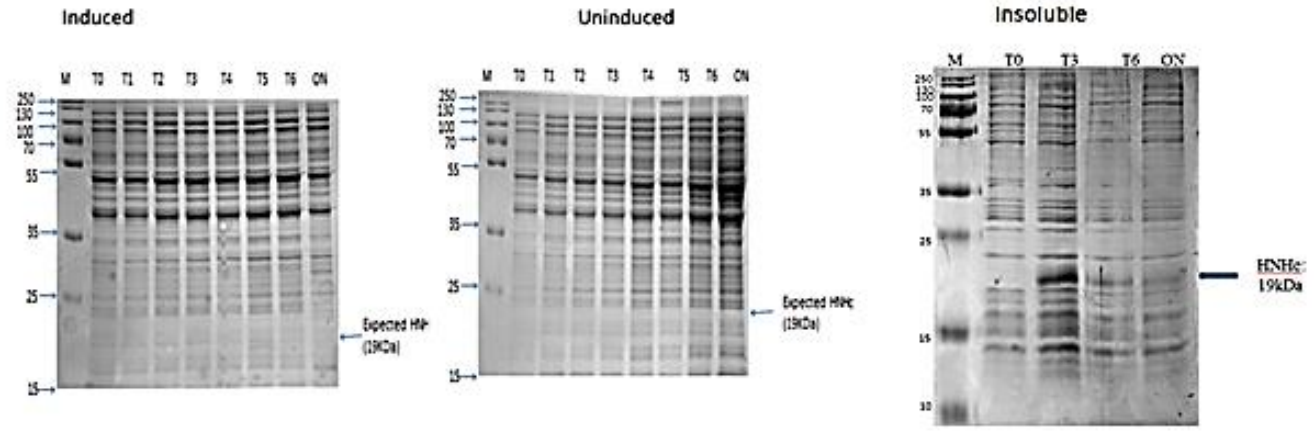
Figure D5 (A, B, C, D, E, F, G, H and I): SDS-PAGE analysis of recombinant genes expressed in LB at 25°C with 1mM IPTG concentration. A: pET20B(+)_PolB1, B: pET20(+)_HNHc, C: pET20b(+)_RNALig1, D: pET28B(+)_RNALig2, E: pET28B(+)_RE, F: pET28B(+)_E7, G: pET28B(+)_PolA1, H: pET30_PolA2 and I: pET30_DNALig1 recombinant genes with the expected protein band corresponding to a size of 23 kDa (Pol B), 19 kDa (HNH), 26 kDa (Lig 1), 44 kDa (Lig 2), 65 kDa (RE), 19 kDa (E7), 83 kDa (Pol A1), 56 kDa (Pol A2) and 36 kDa (DNAlig) indicated by an arrow.

Expression @ 16 C, 1mM IPTG

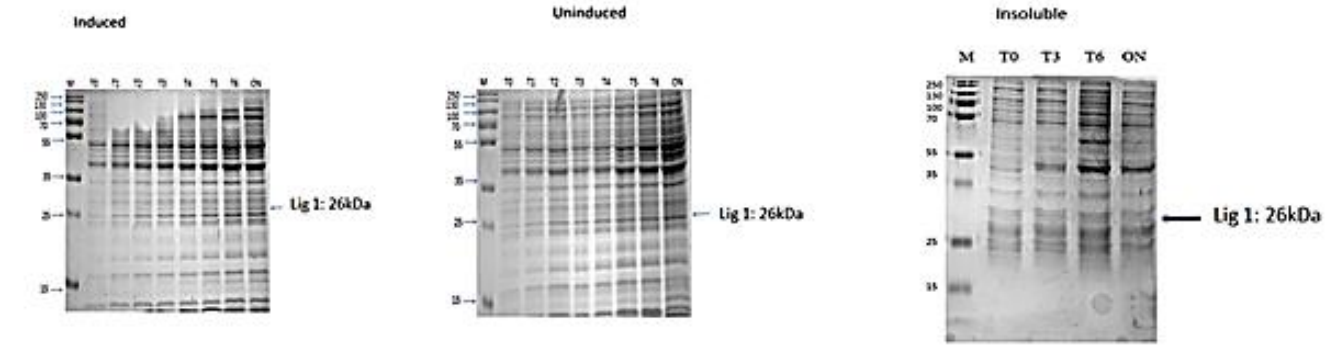
A: Pol B1



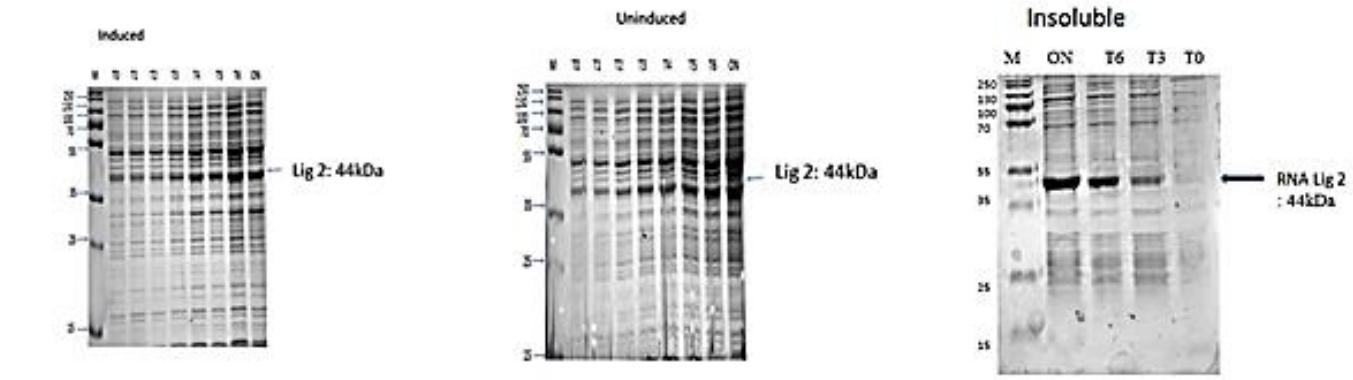
B: HNHc



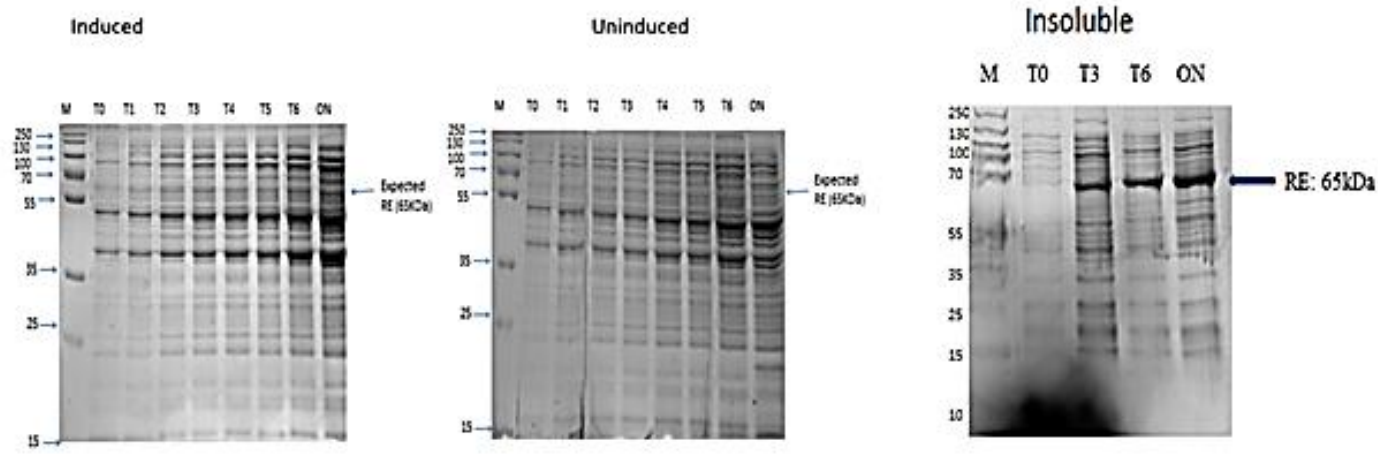
C: RNA Lig 1



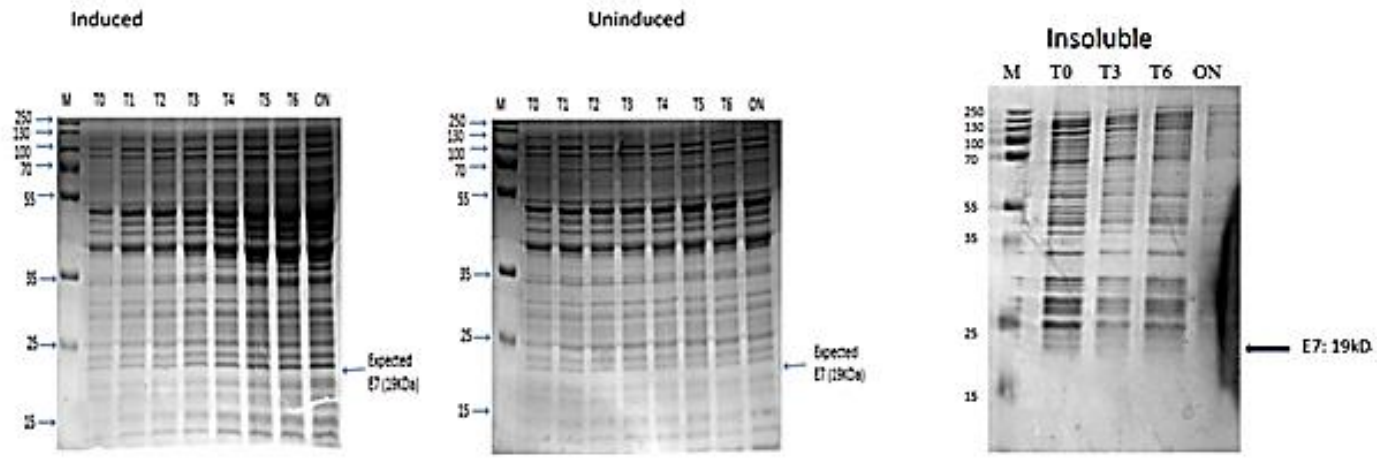
D: RNA Lig 2



E: RE

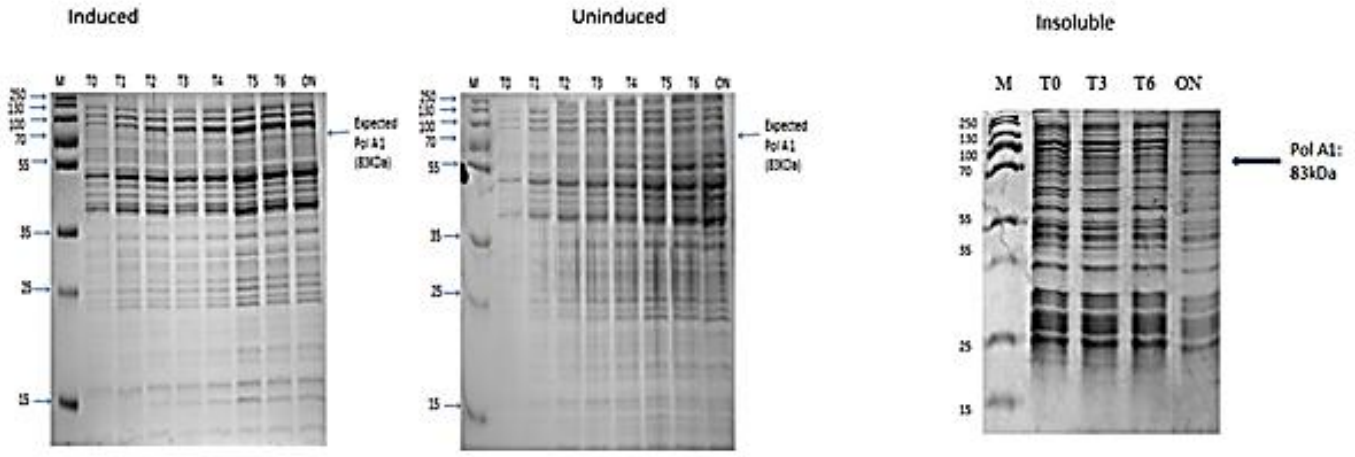


F: E7

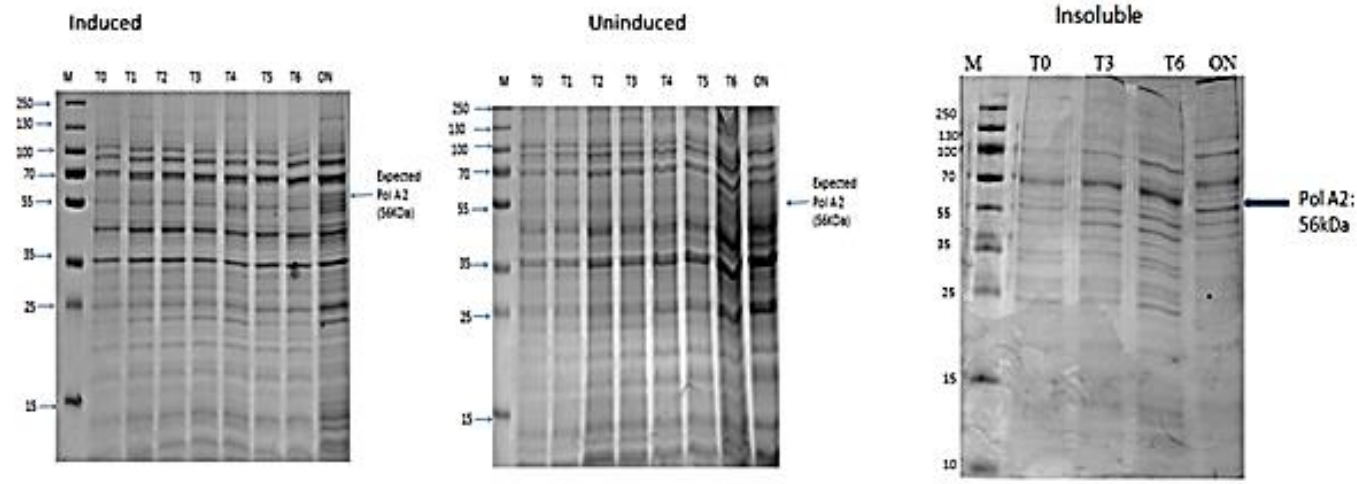


APPENDIX

G: Pol A1



H: Pol A2



I: DNA Lig

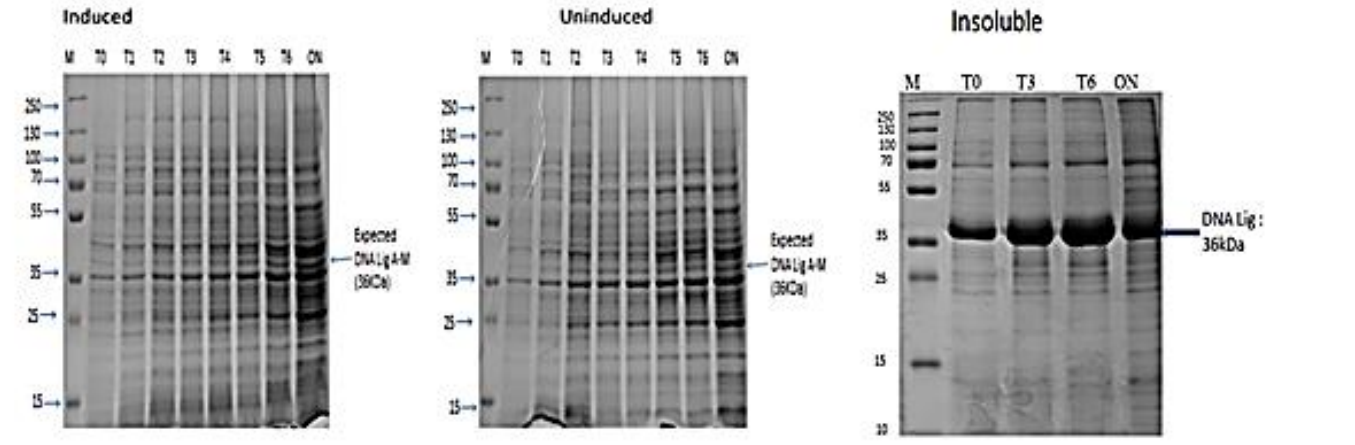


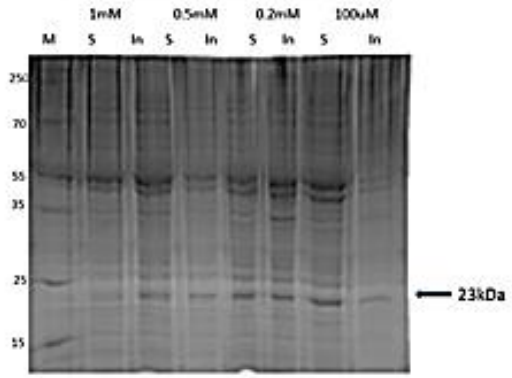
Figure D6 (A, B, C, D, E, F, G, H and I): SDS-PAGE analysis of recombinant genes expressed in LB at 16°C with 1mM IPTG concentration. A: pET20B(+)_PolB1, B: pET20(+)_HNHc, C: pET20b(+)_RNALig1, D: pET28B(+)_RNALig2, E: pET28B(+)_RE, F: pET28B(+)_E7, G: pET28B(+)_PolA1, H: pET30_PolA2 and I: pET30_DNALig1 recombinant genes with the expected protein band corresponding to a size of 23 kDa (Pol B), 19 kDa (HNH), 26 kDa (Lig 1), 44 kDa (Lig 2), 65 kDa (RE), 19 kDa (E7), 83 kDa (Pol A1), 56 kDa (Pol A2) and 36 kDa (DNAlig) indicated by an arrow.

Change in IPTG concentration

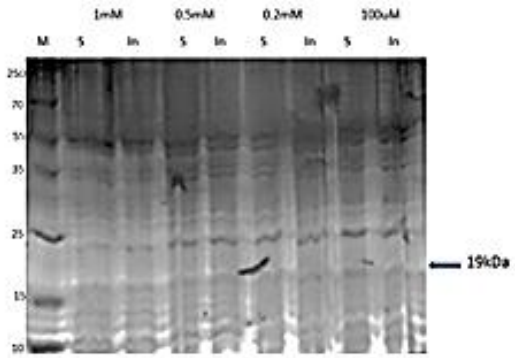
Experiments with variations in the concentration of the inducer (IPTG) from 1mM to 100 µM at 16°C with overnight growth were also attempted. With the exception of the *DNAlig* gene that showed evidence of a soluble protein product (16°C, 100 µM), all the other genes (*PolB, HNHc, RNALig2, RE, PolA2, RNALig1, E7* and *PolA1*) failed to express the proteins (Figure D9). Our results support the observation that induction with IPTG may lead to high levels of recombinant protein that are improperly folded into inclusion bodies (Makrides, 1996; Sohoni *et al.*, 2015), as observed with *DNAlig* where maximum production of insoluble fractions was observed at higher IPTG concentrations. Therefore, lowering the IPTG concentration resulted in an increase in soluble *DNAlig* enzyme production.

APPENDIX

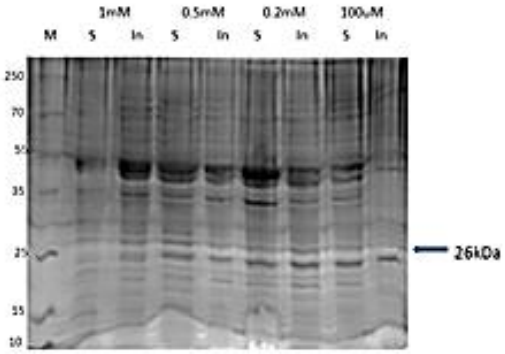
DNA Pol B 1



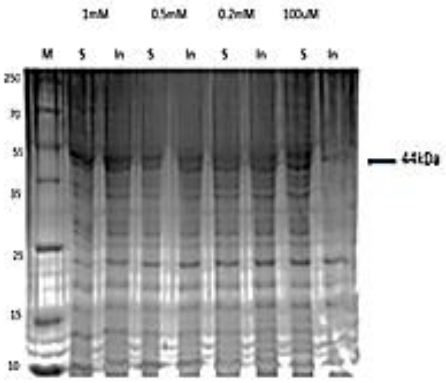
HNHc



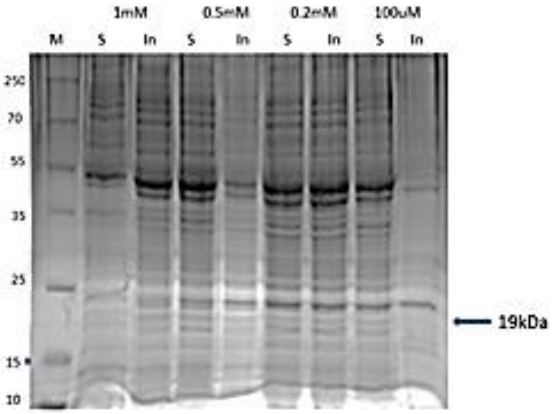
RNA Lig 1



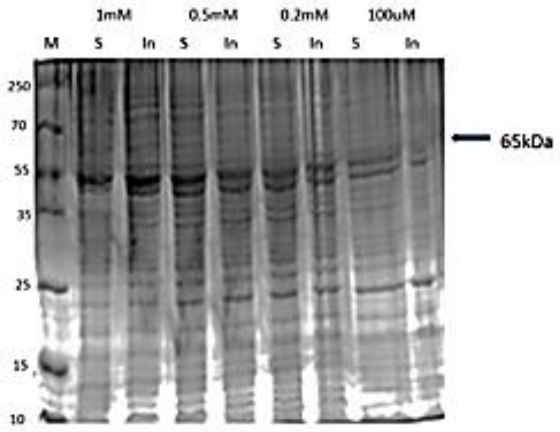
RNA Lig 2



E7



RE



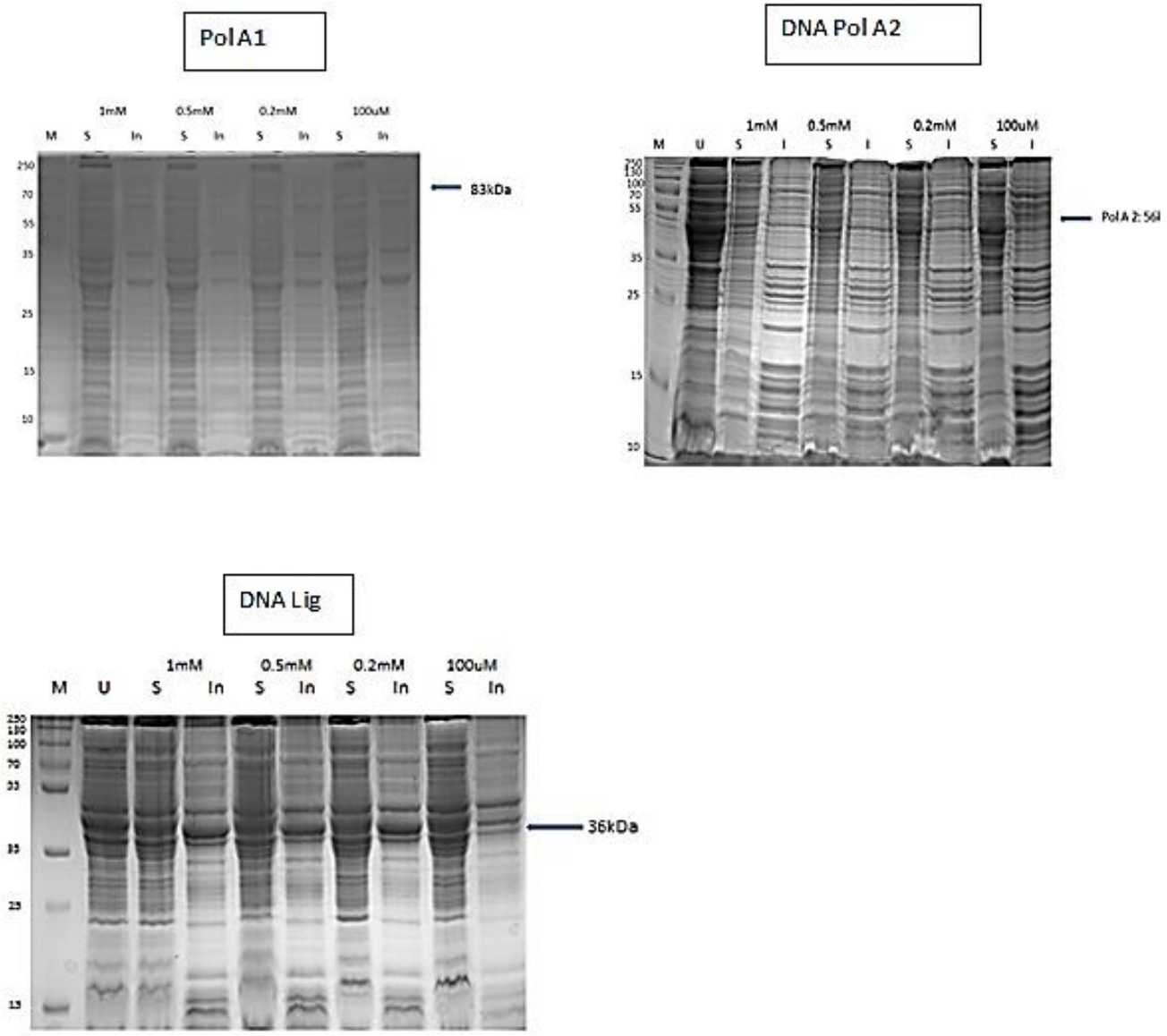
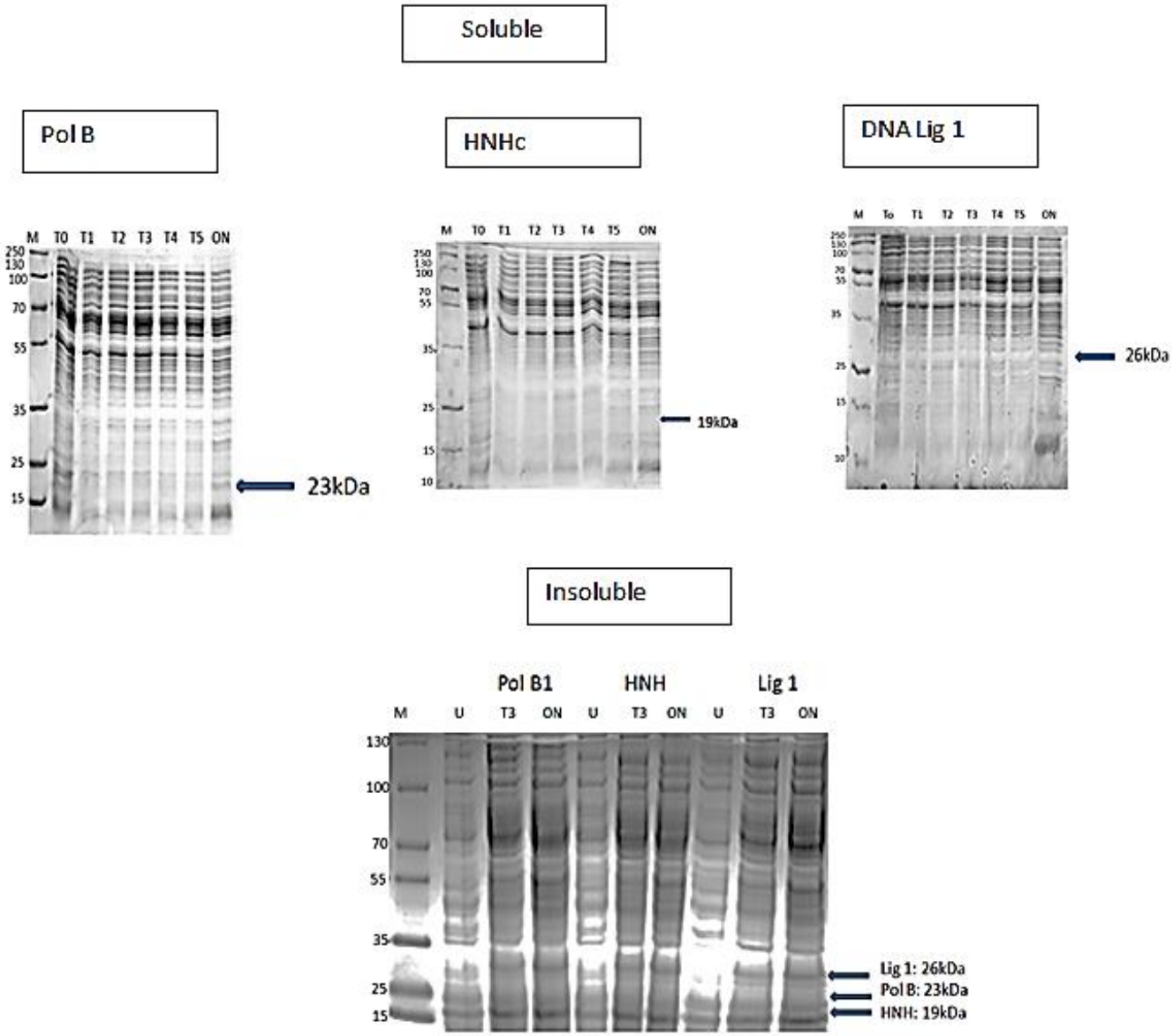


Figure D7: SDS-PAGE analysis of recombinant genes pET20B(+)_PolB1, pET20(+)_HNHc, pET20b(+)_RNALig1, pET28B(+)_RNALig2, pET28B(+)_RE, pET28B(+)_E7, pET28B(+)_PolA1, pET30_PolA2 and pET30_DNALig1 expressed in LB at 16°C with 1mM, 0.5mM, 0.2mM and 100uM IPTG concentrations. Recombinant genes with the expected protein band corresponding to sizes of 23 kDa (Pol B), 19 kDa (HNH), 26 kDa (Lig 1), 44 kDa (Lig 2), 65 kDa (RE), 19 kDa (E7), 83 kDa (Pol A1), 56 kDa (Pol A2) and 36 kDa (DNAlig) indicated by an arrow. Lane M; Marker, Lane U: Uninduced, Lane S: Soluble fraction and Lane IN; insoluble fraction.

Optimisation using different expression hosts

In order to explore alternative methods for enhancing the expression levels of the genes, the use of different *E. coli* host systems was also investigated. *E. coli* BL21 DE3 pLysS SDS-PAGE gels are shown in Figure D10 and *E. coli* BL21 AI cells SDS-PAGE gels are shown in Figure D11.

Optimisation using *E. coli* BL21 DE3 pLysS



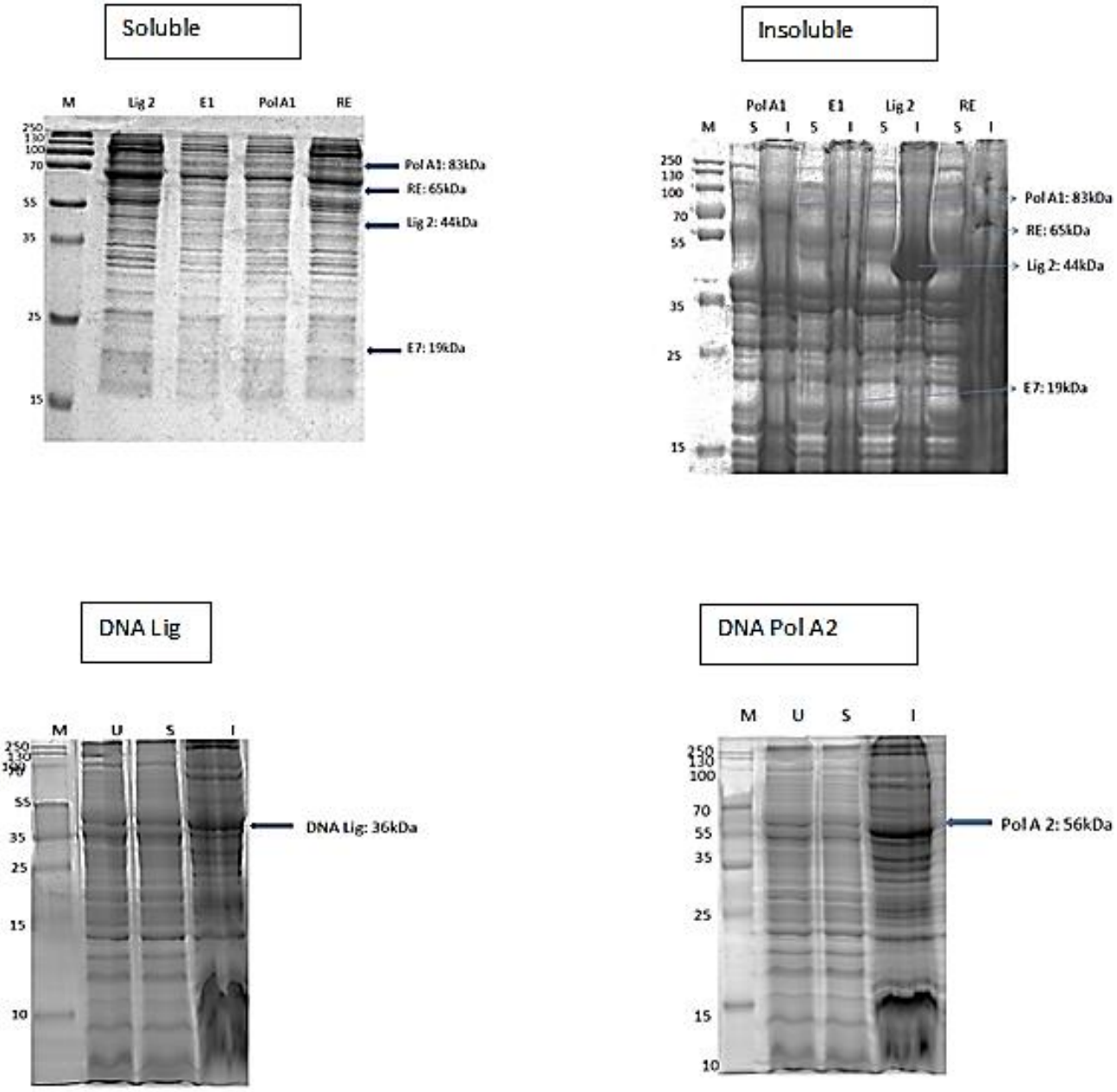
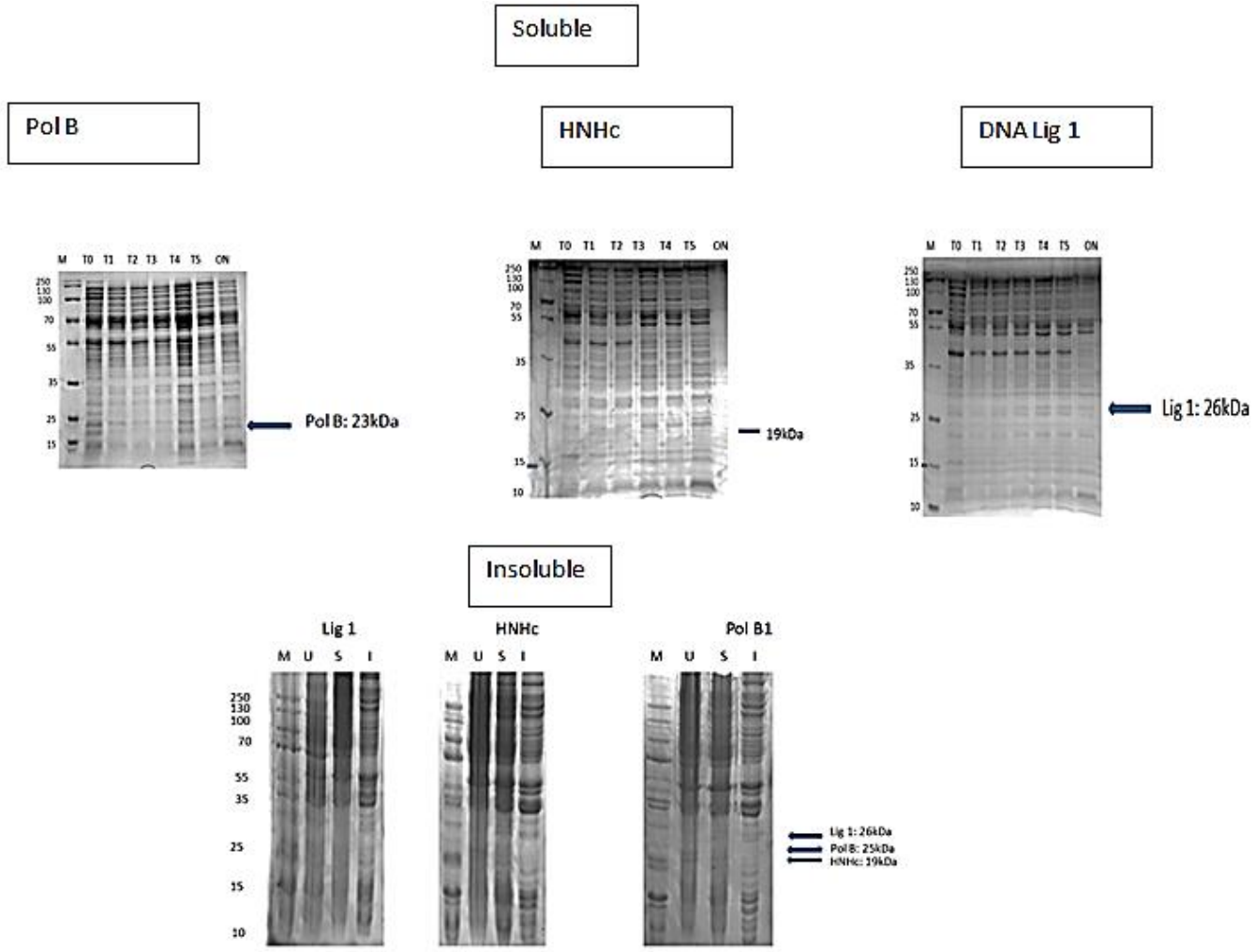


Figure D8: SDS-PAGE analysis of recombinant genes pET20B(+)_PolB1, pET20(+)_HNHc, pET20b(+)_RNALig1, pET28B(+)_RNALig2, pET28B(+)_RE, pET28B(+)_E7, pET28B(+)_PolA1, pET30_PolA2 and pET30_DNALig1 expressed in LB at 16°C with 100uM IPTG concentration using *E. coli* BL21 DE3 pLysS host strain. Recombinant genes with the expected protein band corresponding to sizes of 23 kDa (Pol B), 19 kDa (HNH), 26 kDa (Lig 1), 44 kDa (Lig 2), 65 kDa (RE), 19 kDa (E7), 83 kDa (Pol A1), 56 kDa (Pol A2) and 36 kDa (DNAlig) indicated by an arrow. Lane M; Marker, Lane U: Uninduced, Lane S: Soluble fraction and Lane IN; insoluble fraction

Optimisation using *E. coli* BL21 AI cells



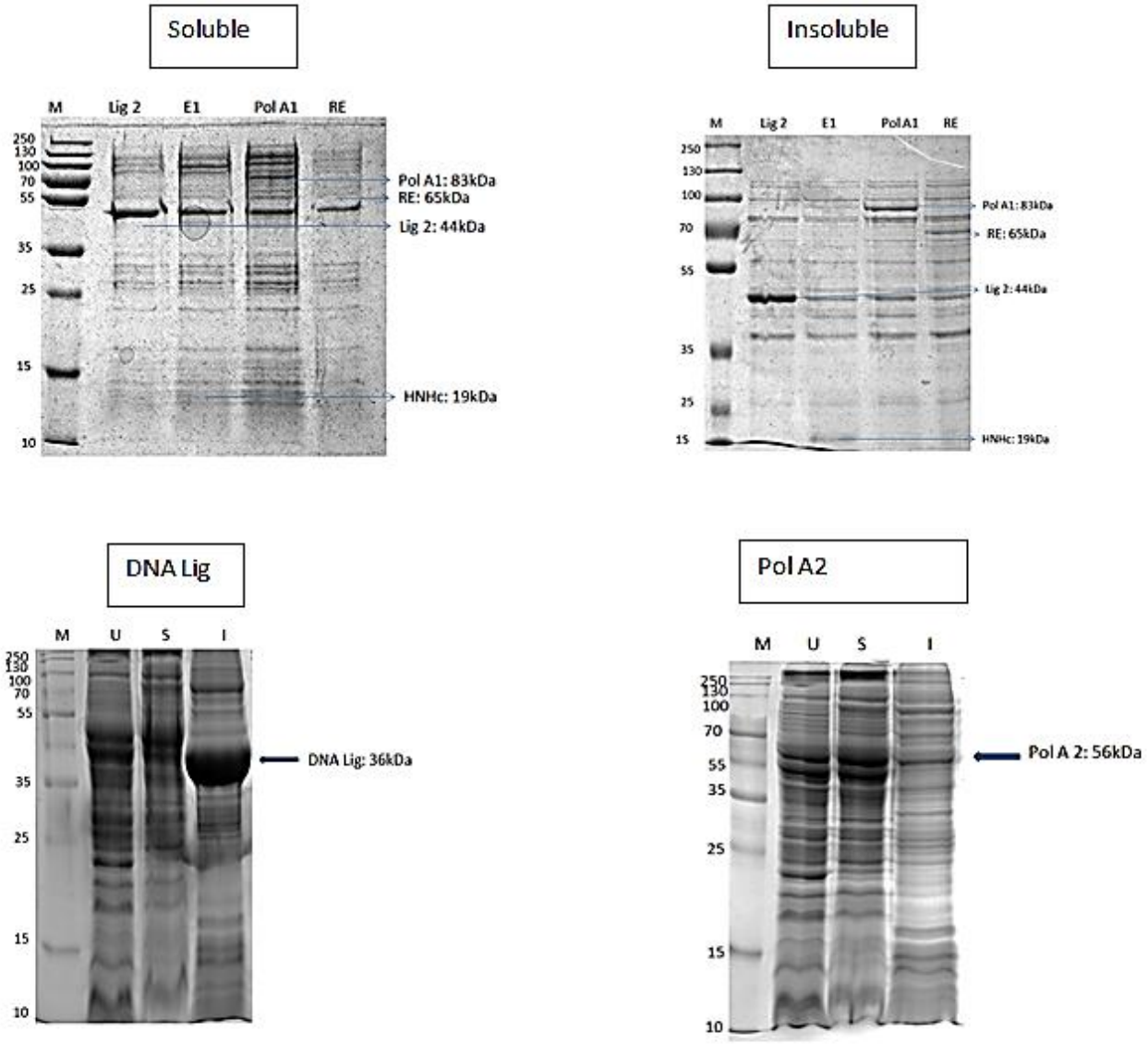
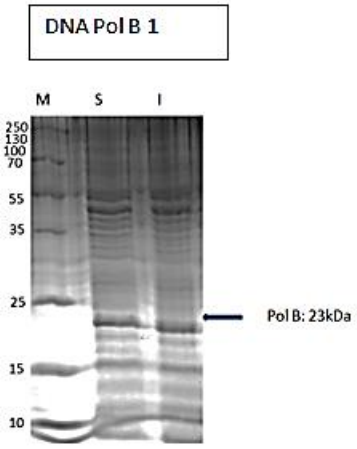


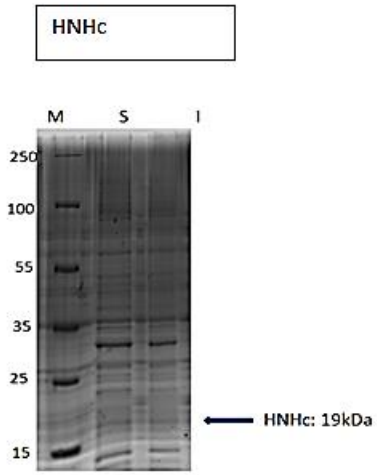
Figure D9: SDS-PAGE analysis of recombinant genes pET20B(+)_PolB1, pET20(+)_HNHc, pET20b(+)_RNALig1, pET28B(+)_RNALig2, pET28B(+)_RE, pET28B(+)_E7, pET28B(+)_PolA1, pET30_PolA2 and pET30_DNALig1 expressed in LB at 16°C with 100uM IPTG concentration using *E. coli BL21 AI* host strain. Recombinant genes with the expected protein band corresponding to sizes of 23 kDa (Pol B), 19 kDa (HNH), 26 kDa (Lig 1), 44 kDa (Lig 2), 65 kDa (RE), 19 kDa (E7), 83 kDa (Pol A1), 56 kDa (Pol A2) and 36 kDa (DNAlig) indicated by an arrow. Lane M; Marker, Lane U: Uninduced, Lane S: Soluble fraction and Lane IN; insoluble fraction.

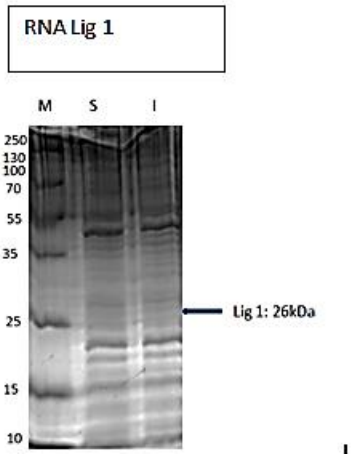
Change in growth medium

Other conditions investigated included changing growth media from LB to 2×YT and EnPresso® B. The use of 2×YT growth medium resulted in no detectable soluble corresponding protein bands for all the constructs. However, the level of insoluble protein for *PolA2*, *RNALig1*, *PolA1* and *DNAlig* significantly increased compared to the LB growth medium (Figure D12 and D13). This medium also produced low levels of soluble *PolA1* and *DNAlig*. The use of the EnPresso® B growth medium (at 16°C, 100µM for 24h) resulted in an expression of the *PolA1* and *DNAlig* soluble proteins.

Optimisation using EnPresso® B growth medium







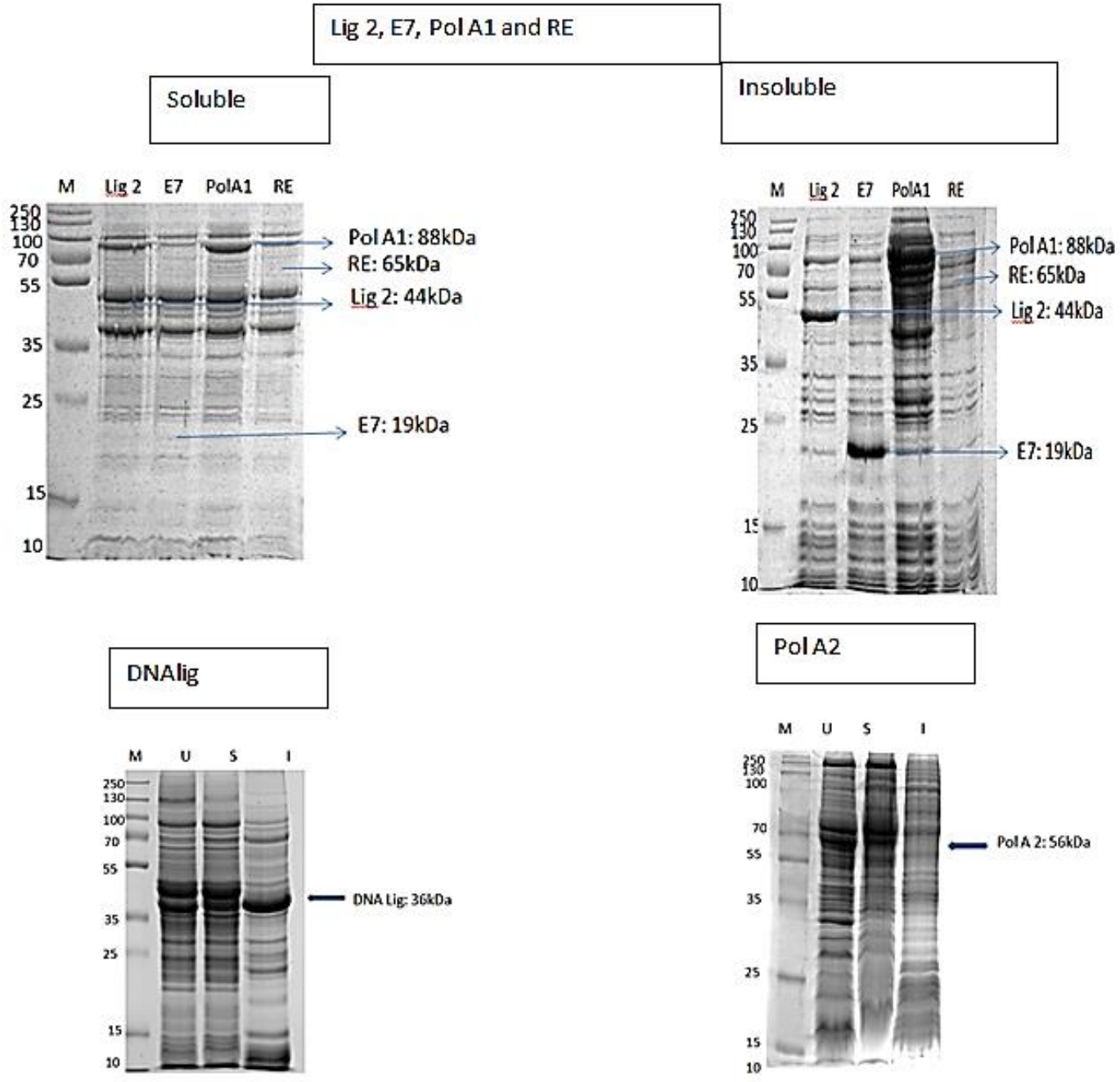


Figure D10: SDS-PAGE analysis of recombinant genes pET20B(+)_PolB1, pET20(+)_HNHc, pET20b(+)_RNALig1, pET28B(+)_RNALig2, pET28B(+)_RE, pET28B(+)_E7, pET28B(+)_PolA1, pET30_PolA2 and pET30_DNALig1 expressed at 16°C with 100uM IPTG concentration using EnPresso® B growth medium. Recombinant genes with the expected protein band corresponding to sizes of 23 kDa (Pol B), 19 kDa (HNH), 26 kDa (Lig 1), 44 kDa (Lig 2), 65 kDa (RE), 19 kDa (E7), 83 kDa (Pol A1), 56 kDa (Pol A2) and 36 kDa (DNAlig) indicated by an arrow. Lane M; Marker, Lane U:

Uninduced, Lane S: Soluble fraction and Lane IN; insoluble fraction 2×YT growth Medium.

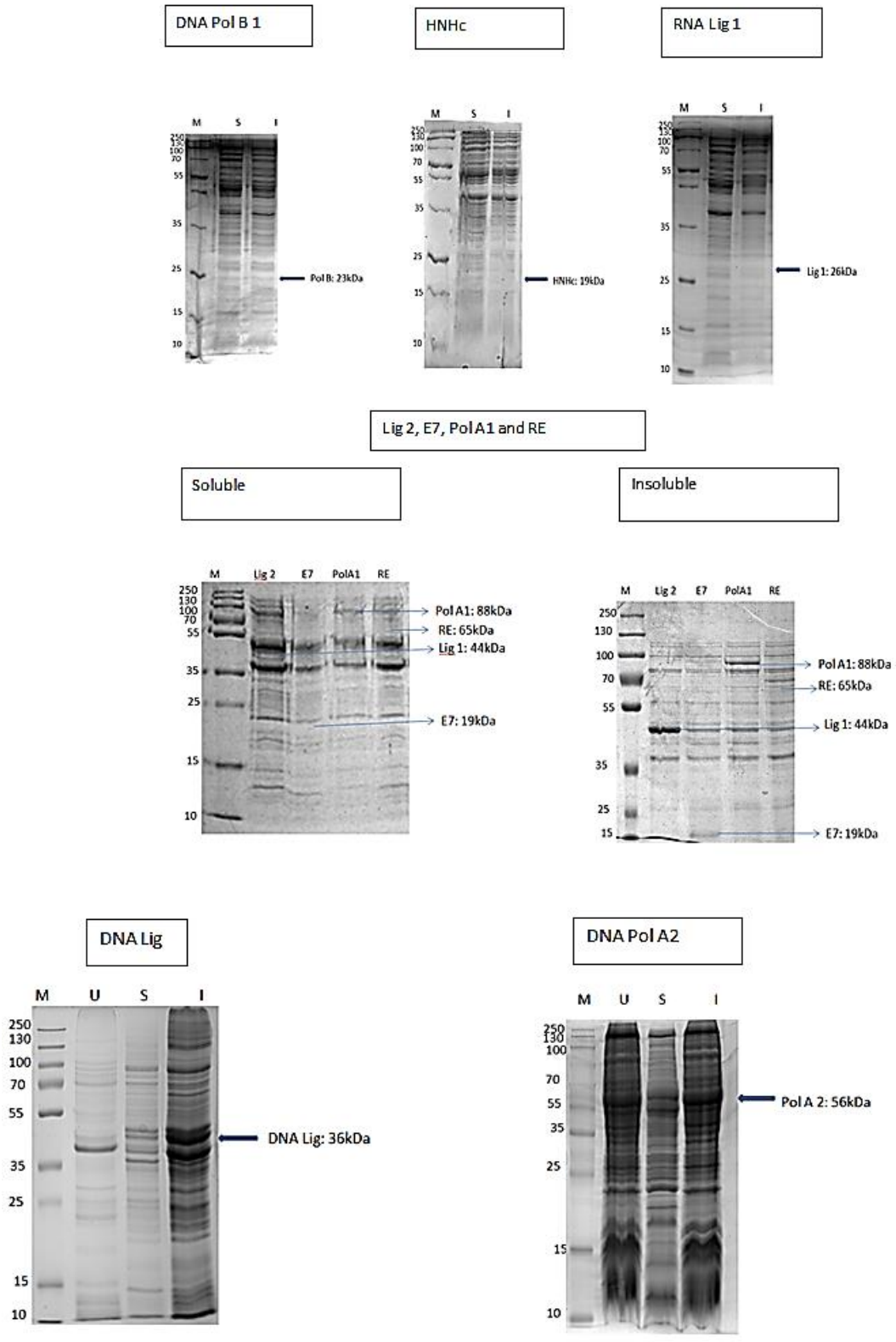


Figure D11: SDS-PAGE analysis of recombinant genes pET20B(+)_PolB1, pET20(+)_HNHc, pET20b(+)_RNALig1, pET28B(+)_RNALig2, pET28B(+)_RE, pET28B(+)_E7, pET28B(+)_PolA1, pET30_PolA2 and pET30_DNALig1 expressed at 16°C with 100uM IPTG concentration using 2×YT growth Medium. Recombinant genes with the expected protein band corresponding to sizes of 23 kDa (Pol B), 19 kDa (HNH), 26 kDa (Lig 1), 44 kDa (Lig 2), 65 kDa (RE), 19 kDa (E7), 83 kDa (Pol A1), 56 kDa (Pol A2) and 36 kDa (DNAlig) indicated by an arrow. Lane M; Marker, Lane U: Uninduced, Lane S: Soluble fraction and Lane IN; insoluble fraction