A metagenomic investigation of phage communities from South African deep mines

By

NOBALANDA MABIZELA

Submitted in fulfilment of the requirements for the degree of

Philosophiae Doctor

In the

Department of Microbial, Biochemical and Food Biotechnology

Faculty of Natural and Agricultural Sciences

University of the Free State

Bloemfontein

South Africa

November 2009

Promoters: Prof. Derek Lithauer

Prof. E van Heerden

I Nobalanda Betty Mabizela, student number; 2000017794 declare that I have received copy right from the University of the Free State of the dissertation entitled: **A metagenomic investigation of phage communities from South African deep mines.**

.....

Signature November 2009

Acknowledgements

I would like to send my special thanks to Prof. D. Litthauer, for believing in my potential. Thank you very much for constructive criticism; it has helped me to grow as a scientist. Once again thank you for everything I really appreciate it.

To Prof. E van Heerden and Prof. K Albertyn thank you very much for your time and assistance.

I would like to express my gratitude to everyone who contributed in making this project a success. Friends; Godfrey, Nathlee and Kamini for moral support and constructive discussions we used to have on our projects. Members of the Extreme Boiochemistry thank you very much for everything and troubles of course, I have really grown as individual and in terms of working with people. Everyone in the department of Biotechnology thank you very much.

To Mpho Mokoena thank you very much for waiting and postponing our wedding so that this project can be completed. I will always be grateful for the support, understanding and care you give me.

Family; father (Mabizela), sister (Nomshado), brothers (Vusimuzi and Bongani) thank you for your support and understanding through out my studies.

I would like to thank National Research Fund (NRF) for the financial assistance.

The Oppenheimer Trust for funding the research visit to Spain

Thank you God for bringing me this far I would not have made it without your grace and mercy.

This dissertation is dedicated to the following people

Mpho Mokoena, Stephen Mabizela, Nomshado Mabizela, Vusimuzi Mabizela and Bongani Mabizela

Table of contents

Chapter 1

Literature Review: Phage Metagenomics

Chapter 2

Uncultured phages from Loch Logan pond, Bloemfontein, South Africa: Optimization of phage isolation and detection

Chapter 3

Uncultured T4-like and T7-like phages from four South African deep mines

Chapter 4

Sequencing of viral communities from South African deep gold mines

Chapter 5

Expression of novel phage proteins from a Beatrix mine phage metagenome

Chapter 6

List of tables

List of figures

Chapter 1

Literature Review: Phage Metagenomics

1.1. Introduction

Microorganisms the comprise majority of planet's biological diversity, however approximately 99% of environmental microorganisms cannot be cultured by standard techniques, and they are distantly related to the cultured ones (Riesenfeld *et al.*, 2004). These uncultured microorganisms include phages; prokaryotes and small eukaryotes. Most environmental microorganisms are viruses, specifically bacteriophages which must be cultured on microbial hosts. The number of phages is estimated at ~10^{30} in the oceans alone (Rohwer, 2003) and the sequence data obtained as part of the Sorcerer II Global Ocean Expedition (GOS) revealed a high abundance of viral sequences, representing approximately 3% of the total predicted proteins, indicating that only a small fraction phage metagenomes have been completely sequenced (Williamson et al., 2008). The diversity of phages is likely to be great therefore more complete genome sequences are necessary to fully comprehend the genetic diversity, evolution of phages and their ability for genetic mobilization/exchange or horizontal gene transfer.

The study of environmental phages using classical methods is hindered by the fact that phages must infect a specific host before culturing and most microbes in the environment cannot be grown under standard laboratory conditions (Hambly and Suttle, 2005). Metagenomics or culture-independent methods are therefore necessary to understand the genetic diversity, population structure, and ecological roles of the majority of phages. These technologies have the potential to answer fundamental questions in microbial ecology and have provided access to genetic information available in environmental samples (Cowan, *et al.*, 2005). Metagenomic techniques range from PCR cloning; library construction and sequencing of DNA segments and then carrying out a comparative analysis with appropriate database in order to identify unculturable microorganisms. PCR cloning and sequencing of 16S rDNA is common in estimating the diversity of prokaryotes from the environment, the 16S product enables comparative sequence analysis for the identification and classification of bacteria and archaea (Lane *et al.*, 1985). However, unified molecular taxonomy is impossible for phages as they lack a universally conserved gene or sequence (Maniloff, 1995). This necessitated genome-based taxonomy for phages based on the overall sequence similarity of completely sequenced phages.

Groups of phages can now be detected from the environment using a different conserved gene or loci specific for a phage group or family (Rohwer and Edwards, 2002, Angly *et al.*, 2006).

Other methods that have been applied to examine the diversity from the environment are denaturing gradient gel electrophoresis (DGGE), pulse field gel electrophoresis (PFGE) and shotgun library constructions. DGGE is used within the same viral groups and a band on the gel can correspond to a different cluster of the same phage group (Diez *et al.*, 2000). Whereas with the PFGE bands on the gel corresponds to different phages, as the method can separate DNA according to their size whereby each band on the gel correspond to a phage genome (Wommack *et al.*, 1999).

Discovery of novel/new genes is a fundamental goal in all metagenomics projects, regardless of whether genome sequences can be assembled or not. Therefore the use of PCR and DGGE is inadequate for the discovery of novel viruses from the environment because they rely on the conserved sequences that are present within viral ge nomes (Breitbart and Rohwer, 2005). These methods can only be used preliminarily to identify the presence of phages from the environments. Until recently construction of phage shotgun libraries have been used to study the diversity of phages from the environment, by first cloning viral DNA into a transcription free vector followed by sequencing of the clones. The method is still useful except that is time consuming and limited sequencing data is obtained from this approach as compared to the data obtained with pyrosequencing. In contrast to the Sanger sequencing this method does not require cloning of DNA in a vector. Due to its effectiveness, pyrosequencing is slowly replacing the use of shotgun libraries especially in phage metagenomics. Pyrosequencing can result in millions of bases in a week and the technique has been applied successfully to study the diversity of microbes as well as phages from the oceans and other unculturable samples (Angly *et al.*, 2006, Williamson *et al.*, 2008).

1.2. Bacteriophages (phages): A definition

Bacteriophages are viruses that infect bacteria and they have single- or double-stranded DNA or RNA genomes that range in size from a few thousand to half a million base pairs (bp) (Madigan, *et al.*, 1997). The genome can be linear or circular and it is packaged inside a protective coat called the capsid, which surrounds the genetic material. The caspid is made up of morphological subunits called capsomers, which consist of protomers and some also contain lipids and structures such as tails and spikes (Grabow, 2001). Each phage specifically targets a

certain bacterium, or several bacteria as its host, and they cannot infect the cells of organisms more complex than bacteria because the surface properties of these cells are not susceptible to the phage's invasion. Phages can be lytic (virulent) or temperate (lysogenic) depending on the relationship established between the phage and their respective specific hosts.

1.2.1. *Lysogenic (temperate) phages*

These phages are able to establish a symbiotic relationship with bacterial hosts they infect. After adsorption and infection, the phage genome integrates into the host chromosome and becomes latent, persisting as a prophage (Ackermann and Dubow, 1987). Bacteria carrying prophages are described as lysogenic, and they have the potential to produce phages and eventually lyse the host cells. The cycle of a lysogenic virus infection extends over several replications of the infected host cell. Prophage can enter the lytic cycle through the process called induction (Lewin, 1997).

1.2.2. *Lytic phages*

Lytic phages always infect a cell from the outside without integrating their genetic information into the genome of the host. After replication newly produced phages are released by bursting (or lysing) the cell (Ackermann and Dubow, 1987). There are two fundamentally different strategies for host cell lysis, used by phages. Most double stranded DNA phages synthesize an endolysin, which degrades the peptidoglycan layer of bacterial cell wall. In addition, phages encode holins, and these proteins facilitate lysis through their pore forming ability thereby allowing the endolysin to access the peptidoglycan (Young, 1992). Alternatively, single stranded DNA phages with smaller genomes encode proteins that interfere with bacterial enzymes involved in the peptidoglycan biosynthesis. Therefore cell lysis occurs through the collapse of the bacterial cell wall from the osmotic pressure from within, influenced by the impaired peptidoglycan synthesis (Bernhardt *et al.,* 2001).

1.3. Infection of host cells

Multiplication of phages in bacterial hosts proceeds in the following main steps: attachment, penetration, replication and packaging.

1.3.1. *Attachment*

The first step in the infection process is the adsorption of the phage to the bacterial cell and this is facilitated by specific surface structures called receptor sites, where phages attach. The nature of these receptors varies with different phages and they can be cell wall lipopolysaccharides or proteins. Teichoic acids, flagella, and pili can serve as receptors, and variation in receptor properties is partly responsible for phage host preferences (Spinelli *et al.*, 2006, Wendlinger *et al.*, 1996). This step is mediated by the tail fibers or by some analogous structure on phages that lack tail fibers and it is reversible. The tail fibers attach to specific receptors on the bacterial cell and the host specificity of the phage is usually determined by the type of tail fibers that a phage has (Tétart *et al.*, 1996). The receptors are on the bacteria for other purposes and phages have evolved to use these receptors for infection. Attachment of a phage to the bacterium via the tail fibers is weak and reversible, therefore the components of the base plate mediates irreversible binding of phage to a bacterium. The irreversible binding of the phage to the bacterium results in the contraction of the sheath and the hollow tail fiber is pushed through the bacterial envelope (Calendar, 1988). Phages that do not have contractile sheaths use other mechanisms to get the phage particle through the bacterial envelope. Some phages have enzymes that digest various components of the bacterial envelope.

In cases of non-tailed phages other mechanisms are used to attach to the host. Examples include the Ff class of single-stranded DNA filamentous bacteriophages which infect *Escherichia coli* containing the information for the F conjugative plasmid. Infection is initiated by the binding of one end of the phage to the tip of the conjugative pilus. Recognition of the pilus tip is the function of the amino-terminal portion of the phage protein (pIII), a minor capsid protein found at one end of the phage particle (Marvin, 1998). Binding of the phage is thought to be followed by retraction of the pilus, bringing the pIII end of the particle near the surface of the bacteria. Once at the cell surface, most or all of the capsid protein integrates into the bacterial cytoplasmic membrane and the DNA is translocated into the cytoplasm (Click and Webster, 1997).

1.3.2. *Penetration (Nucleic Acid Injection)*

During injection nucleic acid from the head passes through the hollow tail and enters the bacterial cell. Usually, the only phage component that actually enters the cell is the nucleic acid, and the remainder stays on the outside of the bacterium (Sambrook *et al.*, 1989).

1.3.3. *Replication*

Linear double stranded DNA phages carry a molecule with complementary single-stranded termini 12 nucleotides in length (cohesive, *cos* termini), and after infection the *cos* sites associate by base pairing. These nicks are rapidly sealed by the host's DNA ligase to generate a closed circular DNA molecule that serve as the template for transcription at the early phase of infection (Chauthaiwale *et al*, 1992). During the later stages of infection DNA replication results in multiple copies of the circular genome of the bacteriophage, a terminase enzyme is responsible for the excision of a single genome at the *cos* site creating linear DNA.

In cases of ssDNA phages, the DNA must be converted to a double-stranded form before either replication or transcription can occur. When the phage DNA enters the host, it is immediately copied by the bacterial polymerase to form a double-stranded DNA, the replicative form or RF. The replicative form then directs the synthesis of more RF copies, mRNA and copies of the DNA genome. The filamentous ssDNA bacteriophages behave quite differently in many respects from other ssDNA phages. The fd phage, during replication a replicative form is first synthesized and then transcribed. Phage-coded proteins then aid in replication of the phage DNA by use of a modified rolling circle mechanism, in which pI1 cleaves the positive strand of RF DNA at the positive-strand origin, and host enzymes extend the 3' end of the nick, generating a new positive strand (Russel, 1995).

1.3.4. *Packaging*

The assembly of viral proteins and nucleic acids into mature and biologically active virions involves a diversity of macromolecular interactions. After capsid formation, structural and packaging proteins must interact with viral nucleic acids. These interactions may confer packaging specificity, spatially organize the genome, enhance particle stability, or contribute directly to capsid quaternary structure. Typically, packaging proteins are extremely basic, neutralizing the negative charges associated with the genome. There are different strategies for packaging viral genome, and they include filling of the pre-formed capsid structures with previously synthesized nucleic acid or material that is being synthesized during packaging (Mindich, 2004). In other cases the genomic RNA or DNA is transported into an assembled polyhedral particle (Catalano, 2000). During the late stages of infection in icoshedral phages the head assembly and DNA replication converge in preparation for packaging. The head assembly pathway produces a mature empty prohead, and the DNA replication pathway result in a head-to-tail DNA, called a concatemer. The terminase links the two pathways by recognizing the viral

DNA, making the endonucleolytic cut and joining it to the prohead through the specific interactions (Yang and Catalano, 2003). The above reactions are ATP-dependent (Mitchell *et al.*, 2002).

1.4. Classification

Phages are the most abundant groups of organisms in the biosphere, and are capable of infecting a large diversity of bacterial hosts. However, they have proven difficult to classify, because of their genetic variation, phages with similar morphologies, modes of replication, and overall genomic architectures may be completely unrelated at the nucleotide level. Classification based on their host range or available life-cycles has led to conflicting conclusions regarding the origin and evolution of phages. Groups of phages related to each other by common gene organization and some degree of sequence similarity do exist, and evidence for horizontal transfer among phage genes have been reported (Rokyta *et al.*, 2006).

The taxonomy of viruses is therefore based upon two main criteria, which are morphological features as well as nucleic acid material. In addition the following characteristic features; mechanism of replication and assembly also forms important part of phage classification (Maniloff and Ackermann, 1998). The current ICTV phage classification includes one order, *Caudovirales* or tailed phages; seventeen families and three floating group (Ackermann, 2007).

More than 96% of phages are tailed belonging to the order *Caudovirales* (Ackermann, 2000) and has been assigned into three families based on the tail morphology (specifically tail length), replication and assembly of phages (Maniloff and Ackermann, 1998). The families are, *Myoviridae*, *Podoviridae* and *Siphoviridae* with contractile tails, short tail stubs and long tails, respectively. The family *Myoviridae* is characterized by a double-stranded DNA genome, an icosahedral capsid, and a contractile tail with associated base plate and extended tail fibers (Ackermann and Krisch, 1997). Detailed descriptions on other families are indicated on Table 1.

Viruses infecting archaea are also classified as phages, and have been classified into seven families, primarily, on the basis of their unusual or unique morphotypes, and this classification is reinforced by the genomic properties. Phages infecting crenarchaea are mostly dsDNA and have morphotypes that have not previously been observed among dsDNA viruses of bacteria and euryarchaeota (Rachel *et al.*, 2002) even though there are some exceptions. The crenarchaeal viruses that have unique virion structures are the droplet-shaped virions of the *Guttaviridae*, the bottle-shaped virions of the *Ampullaviridae*, and the two-tailed virion of the *Bicaudaviridae* (Haring *et al.*, 2004, 2005).

Most known viruses of Euryarchaeota resemble tailed dsDNA bacteriophages, with icosahedral heads and helical tails, contractile or non-contractile, and, accordingly, have been assigned to the families *Myoviridae* and *Siphoviridae*, respectively (Haring *et al.*, 2005).

Table 1.1: Overview of phage families (modified from Ackermann, 2006)

Shape	Family of phages	Genome	Characteristic features
Tailed	*Myoviridae* (A-1,2,3)	Linear dsDNA	Contractile tail, isometric head
	Siphoviridae (B-1,2,3)	Linear dsDNA	Long and non-contractile tail, isometric head
	Podoviridae (C-1,2,3)	Linear dsDNA	Short and non-contractile tail, isometric head
Polyhedral	*Microviridae*	Circular ssDNA	Icosahedral capsid
	Corticoviridae	Circular supercoiled dsDNA	Icosahedral capsid with internal lipid layer
	Tectiviridae	Linear dsDNA	Icosahedral capsid with pseudotail
	Leviviridae	Linear ssRNA	Poliovirus-like with icosahedral capsid
	Cystoviridae	Segmented with three molecules of linear dsRNA	Enveloped, icosahedral capsid, lipids,
Filamentous	*Inoviridae genus* (Inovirus/Plectrovirus)	Circular ssDNA	Long and short rods with helical symmetry
	Lipothrixviridae	Linear dsDNA	Enveloped filaments, lipids
	Rudiviridae	Linear dsDNA	Helical rods
Pleomorphic	*Plasmaviridae*	Circular supercoiled dsDNA	Enveloped, lipids, no capsid
	Fuselloviridae	Circular supercoiled dsDNA	Enveloped, lipids, no capsid
	Salterprovirus	Circular supercoiled dsDNA	Lemon-shaped
	Guttaviridae	Circular supercoiled dsDNA	Droplet-shaped
	Ampullaviridae	Linear dsDNA*	Bottle-shaped
	Bicaudaviridae	Circular dsDNA*	Two-tailed
	Globulaviridae	Linear dsDNA*	Paramyxovirus-like

* Not yet classified by ICTV

1.5. Phages and Functions

1.5.1. *Ecological functions*

Among key roles of viruses in aquatic ecosystems is their potential effect on community composition, structure and diversity. Phages affect microbial evolution by killing specific microbes; hence they are a major source of diversity. Thus, viruses may control populations and have the ability to either maintain or drastically alter bacterial and cyanobacterial community composition (Williamson *et al.*, 2005). As mortality agents of heterotrophic and photosynthetic microbes, they affect the cycling of carbon and nutrients.

Temperate phages play a major role in the evolution of bacterial genomes and the generation of microbial diversity. They mediate rearrangements of bacterial chromosomes (Nakagawa *et al.*, 2003), transmit non-viral genes by transduction and alter the phenotype of their host through lysogenic conversion (Canchaya *et al.*, 2003). This is evident in many non-pathogens and pathogens whereby the latter encode exotoxin genes usually from phage origin (Davis *et al.*, 2002).

1.5.2. *Therapeutic applications*

The emergence of antimicrobial resistance among a multitude of bacterial and fungal pathogens has become a critical problem in modern medicine, and phage therapy has therefore been proposed as a natural alternative approach to conventional antibiotics. The therapy involves the use of lytic phages to specifically kill pathogenic bacteria as an alternative to antibiotics (Clark and March, 2006). One characteristic that allows phages to be useful in this area is the fact that, they infect specific a bacterium or several types of related species of bacteria. Susceptibility to lysis by a particular phage may be the only apparent phenotypic difference between two bacterial strains and may be the only means by which a strain causing an outbreak of disease can be recognized. This observation is the basis of phage typing; hence lysogenic phages cannot be used for this application as they may not lyse the bacterial cell and might therefore introduce virulence genes.

Phages are effective in combating infections caused by a variety of pathogens in humans (Sandeep, 2006). Examples include *Listeria monocytogenes* which is a food-borne pathogen responsible for listeriosis, a frequently fatal infection resulting from the ingestion of food contaminated with this bacterium. Virulent phage, P100 can infect and kill a majority of *Listeria*

monocytogenes strains and therefore can be used to treat listeriosis and also be used as food additive (Carlton *et al.*, 2005, Hagens and Loessner, 2007).

Capsular polysaccharides are virulence factors of many pathogenic bacteria (Taylor and Roberts, 2005). They are hydrated polymer gels, which provide a thick layer protecting bacterial host from harsh environments and immune defense, by masking underlying surface structures. Capsules result in resistance against lysis, which is a crucial step in the development of systemic infections. Phages that encode capsule depolymerases can penetrate the capsule and gain access to the outer membrane (Stummeyer *et al.*, 2006). Hence phages encoding the gene can be used in the development of treatment for these pathogens.

Though lysogenic phages cannot be used on the above applications, they have been used to deliver DNA encoding bactericidal proteins to the bacteria (Westwater *et al.*, 2003). In addition genetically engineered filamentous phage proved to be an efficient and nontoxic viral delivery vector to the brain, offering an obvious advantage over other mammalian vectors (Frenkel and Solomon, 2002).

1.5.3. *Biotechnological functions*

Due to their unique biology, both filamentous and double stranded *E. coli* phages have been exploited as useful cloning vectors (Jones *et al.*, 1986). At first the major obstacle with the use of lambda as the cloning vector was the presence of multiple recognition sites for a number of restriction enzymes in its genome and the other problem was the size required for efficient packaging. This necessitated development of lambda derivatives with one or two sites of a specific recognition enzyme per genome at the nonessential regions of the genome (Murray and Murray, 1974). Limits on the size of the DNA that can be packaged into phage particles has given rise to two different types of cloning vectors, insertional vectors for small DNA fragments and replacement vectors for large DNA (Chauthaiwale *et al.*, 1992). During cloning two arms are produced by restriction enzymes and then joined to the ends of the insert DNA (Dunnl and Blattner, 1987). As an example phage P1 has been used as a Vector for Tn5 Insertion Mutagenesis (Quinto and Bender, 1984) and Fosmid vectors which are involved in cloning of large inserts (Lee *et al.*, 2004).

Phage display if a powerful technology for selecting and engineering polypeptides with novel functions, and it involves fusion of phage coat genes to the DNA encoding these polypeptides. Upon expression, the coat protein fusion is incorporated into new phage particles that are

assembled in the periplasmic space of the bacterium. Expression of the gene fusion product and its subsequent incorporation into the mature phage coat results in the ligand being presented on the phage surface, while its genetic material resides within the phage particle (Benhar, 2001). The technology has been successful in isolation of antibodies, peptide ligands for numerous protein targets, enzyme inhibitors, and mapping of functional protein epitopes and even engineering of the binding specificity and affinity of domains (Sidhu, 2000).

Combining high throughput genome sequencing and bioinformatics tools have allowed identification of a number of genes with potential use in biotechnology. They include sequences that carry conserved regions of genes associated with antibiotic biosynthesis and lysis genes. Promoter sequences and DNA polymerases have also been isolated from phages and they facilitate expression and DNA synthesis, respectively (Studier and Moffatt, 1968). In addition, the ability of phage integrases to specifically and efficiently recombine DNA sequences makes them potentially useful in a variety of genetic engineering applications. Phage integrases are now being used in the *in vitro* GATEWAY™ cloning method developed by Life Technologies (Invitrogen Corporation, Carlsbad, CA).

1.6. Enumeration and isolation of phages

Phages have been traditionally enumerated by culture-based method followed by use of TEM to identify the morphology of the isolated phage. Plaque assays are generally the most used methods to quantify phages using the agar layer method. The method is dependent on the successful infection and lysis of the host cell. In this case a first layer of agar inoculated with the bacterium host is poured, and after it has hardened a second layer inoculated with the phage is added on the surface of the hardened agar (Sambrook and Russell, 2001). A plaque is a region of lysed host cells, and formed by the growth of viruses in a thin layer of hardened agar containing evenly distributed host cells. Plaque growth starts when a free virus particle diffuses to a host cell, adsorbs to its surface, replicates within, and finally lyses it, releasing a new generation of infective viruses, which in turn diffuse to neighboring hosts and repeat the progress (You and Yin, 1999). Theoretically, each plaque is formed by one virus and the number of plaques multiplied by the dilution factor is equal to the total number of viruses in a test suspension. Plaque assays are very specific and only detect the infectious phages for a particular host. A limitation to the technique is that it only selects for the most virulent particles in a heterogeneous phage population, thereby masking the detection of temperate phages and

those with a small burst size (Goyal, 1987). Though plaque assays provide useful information, they are not suitable for the study of phages from the environment where the microbial community from the desired environment will have to be identified first. In addition only a small subset of the microbial community has been successfully grown using traditional culture techniques (Cowan *et al.*, 2005). Hence the use of metagenomic or culture-independent approaches is necessary for the detailed study of environmental phage communities.

1.7. Metagenomics: Definition

Metagenomics is the culture-independent genomic analysis of microbial communities from the environment. The technique has been under development since the late 1990s to overcome limitations involved in gene cloning from the environment (Handelsman, 2004, Kimura, 2006). All metagenomic approaches start with direct isolation of total DNA from the environment followed by use of molecular techniques to analyze the microbial communities. They involve the direct cloning of environmental DNA into different vectors creating large clone libraries to facilitate the analysis of genes and the sequences within these libraries. In most cases metagenomic approaches are coupled with phylogenetic studies based on small ribosomal RNA (16S rRNA) analysis to assess microbial diversity and ecology. To date the culture independent techniques have advanced to such a degree that the DNA isolated from the environment does not require cloning and can be directly sequenced using the newly developed sequencing techniques.

1.7.1. *Diversity studies using ribosomal RNA gene sequence*

The 16S rRNA gene sequence (rDNA) is used for deducing the phylogenetic diversity and evolutionary relationship among bacteria and archaea (Weisburg *et al.*, 1991, Ochsenreiter *et al.*, 2003) and the 18S rDNA is used in eukaryotes (Díez *et al.*, 2001). This ribosomal unit is characterized by highly conserved regions separated with hyper variable stretches, and this feature makes it possible for PCR primer design (Garćıa-Martínez *et al.*, 1999). In this approach the 16S rDNA is amplified from the environmental DNA, the PCR amplicons are then cloned into a vector creating a library which can be screened by sequencing. The sequencing can be coupled with restriction fragment length polymorphism analysis (RFLP) or denaturing gradient gel electrophoresis (DGGE). RFLP analysis involves the digestion of clones with different sets of restriction endonucleases which create different profiles of individual 16S rDNA sequences when

separated on agarose gel. DGGE is a sequence-specific separation of 16S rDNA amplicons of the same size to facilitate profiling of microbial communities. During gel electrophoresis, short 16S rDNA amplicons migrate toward increasing denaturant concentrations, leading to a partial melting of the DNA helix and to a decrease and subsequent ending of electrophoretic migration. As a consequence, a band pattern is produced in which each band represents a bacterial taxon (Schabereiter-Gurtner *et al.*, 2001). In most cases the above mentioned techniques are coupled to sequencing of the clones or sequencing of the excised DGGE bands. The results are then compared to the available 16S rDNA sequences, both unculturable and culturable ones, which in turn provide measures of richness and relative abundance for operational taxomic units (OTUs) in microbial communities (Kemp and Aller, 2004, Hughes *et al.*, 2001).

1.7.2. *Shotgun library constructions*

Library construction is main the metagenomic technique which was initially designed to combat the limitations associated with culturing microorganisms from the environment. The technique has been under development since early 1990s following the success of the use of 16S rDNA as index of diversity which revealed microbial diversity from the environment. The method starts with direct isolation and purification of DNA followed by cloning into a suitable vector and transformation of a host strain. The classical approach involves cloning of small inserts and the use of *E. coli* as the host (Henne *et al.*, 1999). The use of standard sequencing vectors however does not allow cloning of large DNA fragments (> 10kb) and these necessitated cloning into BAC (Rondon *et al.*, 2000) vectors or Fosmids (Lee *et al.*, 2004). Once the libraries have been constructed they can be analysed using two strategies, the sequence-based or the functional driven approach (Handelsman, 2004). The latter depends on the successful expression of target gene(s) in the metagenomic host and clones that express function for desired traits are screened for. The method also depends on the availability of the assay for the target gene; hence proteins with convenient phenotypic characteristics are usually selected. They include the following genes amylases, lipolytic enzymes or antibiotic resistance (Gillespie *et al.*, 2002). The drawback with this approach is the fact that most genes cannot be heterologously expressed in *E. coli* which is mainly used as the host, furthermore other genes function in an operon (Schloss and Handelsman, 2003). However the approach can still be used for the identification of novel and known proteins for applications in biotechnology.

The sequenced-based approach depends on the conserved regions that can be detected using hybridization or PCR amplification. Due to the increasing development the analysis of clones

has now shifted to direct sequencing. The library clones are sequenced using universal primers on the vectors and then compared to the GenBank database in order to identify genes carried within these environments. Direct sequencing of the metagenomic libraries generates vast amounts of data and can be used to deduce metabolic pathways and population structure of the microorganisms. The Global Ocean Sequencing (GOS) is largest shotgun sequencing project with more than 6.12 million proteins predicted from this project (Yooseph *et al.*, 2008). The dataset covers all known prokaryotes and approximately 6000 ORFans that lacked similarity to known proteins have matches to the GOS dataset. In addition 57% of unassembled data was unique. The following closely related organisms were also detected in abundance, *Prochlorococcus, Synechococcus*, *Pelagibacter, Shewanella*, and *Burkholderia (*Rusch *et al.*, 2007*)*.

Shotgun libraries have also been applied in viral metagenomes, and in this case one of the crucial steps is the isolation of phage DNA. The presence of cellular DNA, which is 50 times bigger than the average viral DNA, may overpopulate the viral signal. In addition very few techniques are available for studying phages in environmental samples because of the limitations posed by high dilution in aquatic systems and adsorption to other materials in terrestrial and coarse ecosystems (Benyahya *et al.*, 2001). Therefore a combination of differential filtration, DNAse and RNAse treatment, density centrifugation in cesium chloride is used to separate intact phage particles from bacteria and free DNA. Isolation of Phi29 polymerase has also made it possible to amplify environmental DNA thereby increasing the initial concentrations of DNA obtained. The polymerase has the ability to efficiently displace an annealed DNA strand in front of its advancing 3' end coupled with its very long processivity resulting in multiple displacement amplification reactions (Lovmar and Syvänen, 2006). At this stage the amplified environmental DNA is ready for sub-cloning. However viral genomes contain genes that cannot be directly cloned into the cloning host (e.g. *E. coli*). These gene or gene products include holins and lysozymes and they must be disrupted before cloning, making it difficult to construct a representative of a viral library from the environment. The introduction of linker amplified shotgun libraries (LASL) has made cloning of viral DNA possible. Furthermore the use of vectors that has modification sequences such terminators are now being used to prevent transcription of inserted DNA (Breitbart *et al.*, 2002). Phage libraries are screened by sequencing (Breitbart *et al.*, 2003) as the encoded proteins cannot be heterologously expressed in metagenomic hosts.

1.7.3. *Direct sequencing*

Recent advances in DNA sequencing technologies have accelerated the detailed analysis of genomes from many organisms. The Sanger sequencing method has been used to obtain sequences from clones, but the cloning and sub-cloning into respective vectors is necessary prior to sequencing (Sanger, 1977). DNA isolated from the environment can now be sequenced directly without being cloned in a vector, using pyrosequencing which is a new sequencing technique. The technology is based on sequencing-by-synthesis principle. It is built on a 4-enzyme (Klenow fragment of DNA polymerase I, ATP sulfurase, luciferase and apyrase) real-time monitoring of DNA synthesis by bioluminescence (Ahmadian, *et al.*, 2006). To date the technique has advanced and it takes advantage of DNA capture beads that can contain on average one single-stranded template. Fragmented DNA is attached to the beads by adapters which are also used for amplification of the template into millions of copies in an oil emulsion PCR (emPCR). The beads are then distributed on a solid-phase sequencing substrate (a PicoTiterPlate™) with more than million wells. The wells contain the bead and the following additional reagents, the polymerase, luciferase, and ATP sulfurylase (Margulies *et al.*, 2005). Each fragment is then amplified in its own well and microfluidics cycles each of the four nucleotide triphosphates over the PicoTiterPlate™. The DNA polymerase catalyzes incorporation of complementary dNTP into the template strand. The nucleotide incorporation is followed by release of inorganic pyrophosphate (PPi) in a quantity proportional to the amount of incorporated nucleotide. ATP sulfurylase quantitatively converts PPi to ATP in the presence of adenosine 5' phosphosulfate (APS). The generated ATP drives the luciferase-mediated conversion of luciferin to oxyluciferin, producing visible light in amounts that are proportional to the amount of ATP and can be detected by a charge coupled device (CCD) camera (Figure 1.1). The generated light is observed as a peak signal in the pyrogram. Each signal peak is proportional to the number of nucleotides incorporated (Huse *et al.*, 2007).

Pyrosequencing is now being utilized to obtain genomic information from cultured organisms as well as metagenomes. Different environmental samples have been studied using pyrosequencing and novel sequences identified from these environments (Roesch *et al.*, 2007; Yooseph *et al.*, 2008).

Viral genomes contain modified nucleotides that cannot be directly cloned into the cloning host and must be disrupted before cloning. Pyrosequencing can be used for characterizing unculturable phage communities (Edwards *et al.*, 2006).The technique has been applied in phage environmental genomics, and Angly *et al.*, (2006), could assemble a partial genome of a

single-stranded DNA phage, chp1-Like microphage from the Sargasso Sea. Other marine phage genomes have also been identified using pyrosequencing and the overall results shows that the majority of phage proteins are not similar to the ones in public databases (Dinsdale *et al.*, 2008; Williamson *et al.*, 2008).

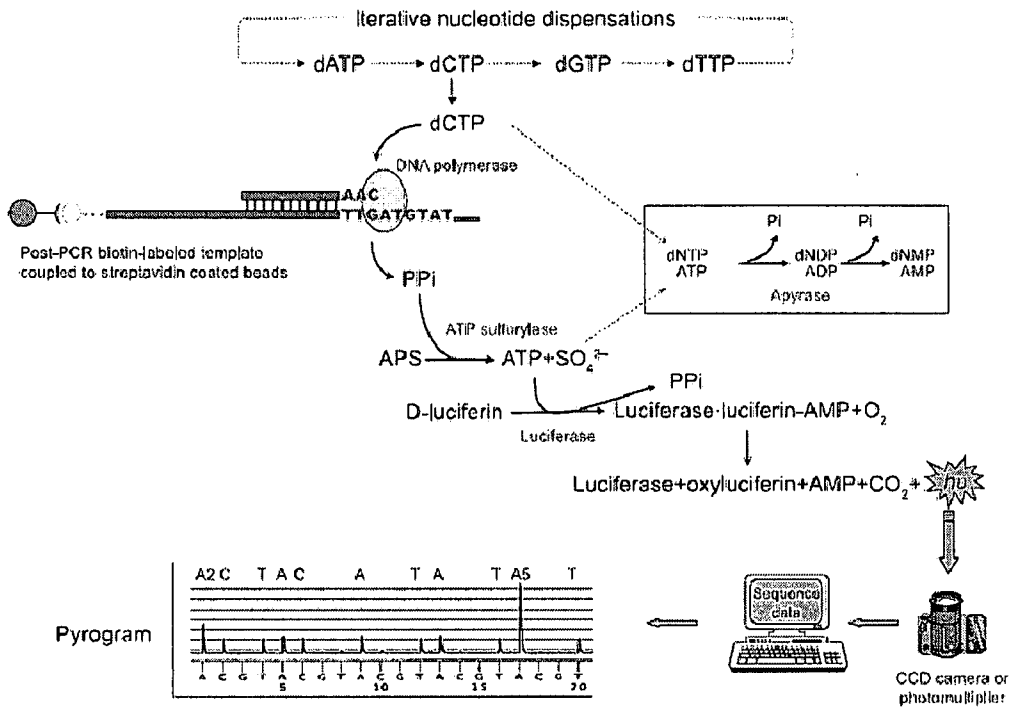


Figure 1.1: Schematic representation of the pyrosequencing technique (Taken from Baback *et al.*, 2006)

1.8. Phage metagenomics

Metagenomic studies have been applied to prokaryotes, and the technique has generated considerable advances in understanding microbial communities from diverse environments. Identifying and studying the diversity of viruses in the environment has always been limited by the following: firstly the classical approach using plaque assays is restricted by unavailability of the cultured of microbial hosts (Riesenfeld *et al.*, 2004, Hambly and Suttle, 2005). Furthermore, there is no single gene that is common to all viral genomes; therefore total

uncultured viral diversity cannot be monitored using approaches analogous to ribosomal DNA profiling. In addition isolated viral DNA from the environment is often overpopulated by free nucleic acids from prokaryotes.

Introduction of phage metagenomics circumvent these limitations and provide insight on viral composition and structure from different environments. Studies of viruses from the environment started in early 2000's with the analysis of marine viral communities (Steward, 2000) and have always focused on aquatic environments. To date the technology has progressed and viral metagenomic libraries has been constructed from different environment, including marine (Breitbart *et al.*, 2002); human fecal samples (Breitbart *et al.*, 2003) and human infant gut (Breitbart *et al.*, 2008). Recent advances include analysis of RNA viral communities by construction of cDNA libraries (Zhang *et al.*, 2006). The following culture independent and metagenomic techniques are now being used to study the diversity and the genomic composition of the viral communities from different environments.

1.8.1. *Ultra-centrifugation and ultra-filtration*

The key to culture independent viral discovery is increasing the levels of viral nucleic acids while reducing background prokaryotic and eukaryotic nucleic acids. Viruses are highly diluted in natural samples; hence concentration of samples is necessary prior to further analysis. For environmental samples that are available in large volumes such as seawater, the concentration of viral like particles and reduction of background nucleic acids from prokaryotic and eukaryotic cells can be performed. This can be archived by using ultra-filtration and ultra-centrifugation. The latter mostly involves the use of gradients (e.g. cesium chloride gradients) (Sambrook and Russell, 2001). Ultra-filtration is based on size exclusion whereby the separation of viruses from other contaminating materials can be obtained by filtering through small pore size membrane (e.g. 0.22 µm and 100 kDa) (Paul *et al.*, 1991). Tangential flow filtration is now being used and the technology has the ability to filter and concentrate water samples (Casas and Rohwer, 2007), however after filtration the samples still have to be treated with DNAse and RNAse to remove free nucleic acids from the environment. Though the above methods have been useful in viral metagenomics their main limitation is the sampling bias, which is due to loss of large viruses during filtering. In addition, the cesium chloride gradients only recover known phage groups, meaning phages with the known genome size.

1.8.2. Microscopic techniques

Different techniques have been designed to estimate viral concentrations in different ecosystems and they include epifluorescence microscopy, transmission electron microscopy and quantitative flow cytometry. Each methodological approach has its advantages and disadvantages, and to a certain extent making viral abundances determined by these methods inconsistent.

1.8.2.1. *Transmission electron microscopy (TEM)*

Until recently, TEM has always been used to study the morphology of phages from infected hosts. The technique can be used to estimate total viral counts from different environments (Demuth *et al.*, 1993). High numbers of phage particles were revealed using direct TEM on aquatic samples (Proctor and Furhman, 1990) and viral counts from other environments are now being estimated with this method. Bacteriophages are first adsorbed on carbon coated film, then stained with heavy metals (e.g. using uranyl acetate or tungstic acid) (Gentile and Gelderblom, 2005). The samples can be positive or negative stained, the latter is conveniently used with phages. Usually before total viral counts can be obtained they are concentrated by ultra-filtration or ultra-centrifugation.

1.8.2.2. *Epifluorescence microscopy (EFM) and flow cytometry (FCM)*

The use of high fluorescence yield nucleic acid dyes such as SYBR Green and SYBR Gold or YOPRO-1 in combination with epifluorescence microscopy has facilitated the quantification of the viruses from the environment (Marie *et al.*, 1996). EFM and FCM are now used to estimate the viral count from metagenomic samples (Weinbauer, 2004). EFM is the preferred method for counting viruses because of its higher accuracy and precision although FCM shows promise as a high-throughput method. The use of flow cytometry has been successful for rapid and accurate counting of free viral particles (Chen *et al.*, 2001, Brussaard, 2004). Individual viral families differ in fine structure, hence the above methods provides direct insight into the morphological variability of phage populations without being dependent on the isolation of suitable host strains (Wen *et al.*, 2004).

Table 1.2: Advantages and disadvantages of some methods used to enumerate viruses (taken from Weinbauer, 2004)

Technique	Advantages	Disadvantages
Transmission electron microscopy	Total counts Rough morphological characterization and sizing	Slight underestimation of total abundance Need for expensive equipment No information on infectivity
Epifluorescence microscopy	Total counts No detection limit on environmental samples	No information on infectivity and morphology No distinction between viruses and DNA bound colloids Distinction between phages and bacteria based on size and staining intensity
Flow cytometry	total counts environmental samples	No information on infectivity and morphology No distinction between viruses and DNA bound colloids staining intensity

1.8.3. *PCR detection*

The study of phage diversity using sequence-based approaches has always been hindered by lack of niversally distributed genes or gene products, in contrast to prokaryotes in which sequences such as 16S rDNA can be used for phylogenetic comparisons (Maniloff, 1995). Development of molecular biology methods for detection of phages from the environment was delayed due to this sequence barrier. Rowher and Edwards in 2002 therefore developed a phage proteomic tree based on 105 available complete phage genomes. The method highlights genes or sequence fragments that are conserved in specific clades of phages to enable the comparative sequence analysis and identification of phages from the environment. Different groups of phages can now be detected using PCR and degenerate primers specific for the gene of interest. With the increasing interest in phages and increasing sequencing of phages, an updated phage proteomic tree was constructed with the total of 510 genomes including those from marine samples (Angly *et al.*, 2006).

All of the T4-type genomes analyzed to date contain a large block of DNA homologous to the T4 sequences that specify virion morphology. On the basis of the comparisons of the sequence of the three major virion structural proteins (gp18; gp19 and gp23), the T4-type phages can be further divided into four subgroups with increasing divergence from T4: the T-evens, the pseudo-T-evens, the exo-T-evens and the schizo-T-evens. The major phage protein, g23 can be used to detect the first three subgroups from the environment, and g20 is used to detect cynophages (Zhong *et al.*, 2002, Dorigo *et al.*, 2004, Filée *et al.*, 2005). Other genes that are being used include DNA polymerase fragment for the identification of unculturable T7-like podoviruses (Breitbart *et al.* 2004) and the intergrase gene for temperate phages (Balding *et al.*, 2005).
Direct sequencing of the viral communities can be performed using the shotgun sequencing or pryrosequencing; detatils disscused on sections 1.7.2 and 1.7.3, repsectively.

1.9. Bioinformatics in phage metagenomes

Sequencing projects using either shotgun sequencing or pyrosequencing produces a large amount of data. After sequencing the assembly is necessary to put together created fragments of DNA. Very often complete assembly is required with genomes and they can be assembled using the following programs; Celera Assembler (Goldberg *et al.*, 2006) for whole genomes, or using Phrap (Wicker *e t al.*, 2006) for cloned targets. Prosequencing data is assembled with Newbler which is sold with this 454 technology.
It is however impossible to reliably assemble metagenomic sequencing reads into longer contigs regardless of the read length, because diversity in metagenomic samples is often too large to provide a high sequencing coverage of single species. Hence only a taxonomic characterization of DNA fragments or contigs is performed for a deeper understanding of metagenomic communities. Metagenomic Rapid Annotation Subsystem Technology (MG-RAST) (Meyer *et al.*, 2008) is a web server that provides annotation, phylogenetic as well as functional classification through the use of a subsystem-based annotation approach. The approach has the ability to compare metagenomic samples to see both shared and unique genes/subsystems. Another widely used approach for processing and exploring metagenomic data is the MEGAN program (Huson *et al.*, 2007). The above techniques can also be used to analyze viral metagenomes. Bioinformatics analysis of viral metagenomes is still at early stages hence, most viral metagenomes studied are compared against the Gen-Bank using BLAST searches. Other programs that are used to characterize viral metagenomes include, PHACCS, PHiGO and Prophage finder. PHACCS (PHAge Communities from Contig Spectra) is a web tool

used for the estimation of diversity and structure of uncultured viral communities, utilizing modified Lander-Waterman algorithm (Angly *et al.*, 2005). Due to the availability of the large sequencing data the tool has been upgraded to PHACCSIII to accommodate large data sets (Domes *et al.*, 2007). Both the Prophage finder (Lima-Mendez *et al.*, 2008) and PHiGO (Toussaint *et al.*, 2007) uses ACLAME database to identify prophages.

1.10. Phage diversity

Bacteriophages are found in all habitats in the world where bacteria proliferate, they are estimated to be the most widely distributed biological entities in the biosphere, with the estimated viral population of 10^{30} in the oceans alone (Hendix, 2003). The diversity of phages is reflected by their genome size which ranges from few kilobases (kb) to several hundred thousand kb. In addition lack of universally conserved genes or sequences is also evidence that phages are highly diverse. The dsDNA tailed phages or *Caudovirales* amount up to 96% of all phages reported thus far and possibly make the up the majority of phages on the planet.

Genomic analyses of cultured and uncultured phages also show that most of the Open Reading Frames (ORFs) are novel (Cann *et al.*, 2005) and only about 10% of the sequences from environmental microbial metagenomes and cultured microbial genomes are novel when analysed in similar ways. Together, these observations indicate that much of the global microbial metagenome has been sampled, whereas the global viral metagenome is still relatively uncharacterized. Furthermore in ACLAME, which is a database for sequenced phage genomes (Leplae *et al.*, 2004) fifty-two percent of the families contain 3 or more members and about one-third of the proteins analyzed are singletons (~36%), or in two-member families (~ 11%). These percentages remain almost constant even if proteins from newly sequenced phage genomes are included in the data set (Lima-Mendez *et al.*, 2007).

1.11. Conclusions

Phages are diverse and largely unexplored component of the microbial community in different environments. Along with their hosts, phages make up the largest biomass on earth, residing mostly in aquatic habitats (Angly *et al.*, 2006). It is clear that though phages are known to be highly specific for their host, there are some that infect a broad range of bacterial species (Jensen *et al.*, 1998). They are currently classified on the basis of their genome (either RNA or

DNA, with ss and ds genomes) and their morphology into seventeen phage families and floating group (Ackermann, 2006). Tailed bacterial phages (*Caudovirales*) amount up to 96% of the total viruses and they have been assigned into three families on the basis of tail morphology.

Phages are not homogenous and their diversity is reflected by the diversity of genome sizes, which ranges from barely 4 kb to up to 600 kb. In addition there is no universal gene that can be used to make inferences regarding the diversity in natural phage community (Maniloff, 1995). However, several genes can be targeted that are associated with specific subsets of phage communities (Rohwer and Edwards, 2002).

Traditionally the discovery of viruses was dependent on culturing the virus in the host cells in order to propagate and isolate enough pure virions for characterization. However the majority of viruses cannot be cultured using standard techniques. Hence, culture-based methods are insufficient for large-scale characterization of the viral community. Recent advances in genomics of viruses and cellular life forms have therefore greatly stimulated interest in the origins and evolution of viruses. Use of metagenomic techniques such as library construction (Breitbart *et al.*, 2002, 2003), has described the population structure and genome size distribution of phage communities. In addition the genomes of phages from the environmental samples can now be sequenced directly without cloning (Williamson *et al.*, 2008). The results from both approaches have indicated that viral genes show no detectable homologs to other species and they have therefore become of increasing interest as environmental sampling suggests that there are many more such novel genes in yet-to-be-cultured phages (Hsiao *et al.*, 2005).

Phages play a critical role in the mortality of aquatic bacteria (Fuhrman. 1999), thereby affecting the microbial food web and biogeochemical processes, as well as affecting bacterial diversity by restructuring the microbial community. Phages also affect their hosts via lysogeny and transduction and by mutually providing genes that enhance host survival (Canchaya *et al.*, 2003, Mann *et al.*, 2003). Other functions are biotechnological as well as therapeutic roles that phages play in molecular biology and modern medicine.

Chapter 2

Uncultured phages from Loch Logan pond, Bloemfontein, South Africa: Optimization of phage isolation and detection

Summary

Uncultured phages are present in high quantities from aquatic environments; hence research on phage diversity has focused on phages from these environments. In this study phages from Loch Logan pond were identified using different microscopic techniques and sequenced-based approaches. All three major *Caudovirales* families were identified when using TEM and high viral counts were also observed with EFM. Both the T7-like podoviruses and T4-like phages were amplified by PCR from water and sediments. Sequencing of DNA polymerase fragment from the T7-like phages revealed that this fragment is highly conserved. In contrast the T4-type major capsid protein was highly variable as expected due to the plasticity of the T4 phages. T4-type and T7-type phages are classified as *Myoviridae* and *Podoviridae*, respectively therefore the results obtained with sequencing and microscopic techniques corresponded. The results further demonstrate that both the techniques can be efficiently used to detect uncultured able phages from water and sediment/soil.

2.1. Introduction

The observation that viruses are the most abundant biological entities has generated a major interest in the characterization and distribution of viral diversity from different ecosystems. There are estimated 10^{31} viruses on earth, most of which are phages and must be cultured on microbial hosts (Rohwer, 2003). However this standard technique is limited by the fact that only a small fraction of environmental microbes are readily cultured and the uncultured fraction includes diverse organisms that are only distantly related to the cultured ones (Riesenfeld, *et al.*, 2004).

Different microscopic techniques have therefore been designed to estimate viral concentrations from different ecosystems and they include epifluorescence microscopy (EFM), transmission electron microscopy (TEM), and flow cytometry (FCM). Enumeration of viruses using the latter two techniques is based on the use of highly fluorescent nucleic acid dyes such as SYBR Green and SYBR Gold (Marie *et al.*, 1999; Wen *et al.*, 2004). Though TEM can be used to estimate total viral counts (Demuth *et al.*, 1993), it is more suitable for morphology characterization, as the method under estimates the amount of viral particles available in the samples. When using TEM, phages adsorbed on a carbon coated film can be negatively or positively stained with heavy metals (e.g. uranyl acetate or tungstic acid) (Gentile and Gelderblom, 2005). EFM and FCM are used to estimate the viral count from metagenomic samples (Weinbauer, 2004). EFM is the preferred method for counting viruses because of its higher accuracy and precision, although FCM shows promise as a high-throughput method. The use of flow cytometry has been successful for rapid and accurate counting of free viral particles (Chen *et al.*, 2001, Brussaard, 2004).

Prior to any microscopic techniques and any other further analysis, viral particles from the environment have to be concentrated. Use of ultra-filtration and ultra-centrifugation has been very effective in this area (Sambrook and Russell, 2001). Ultra-centrifugation involves the use of gradients (e.g. cesium chloride and sucrose gradients) to extract viral particles of known molecular weight. Ultra-filtration is based on size exclusion whereby the separation of viruses from other contaminating materials can be obtained by filtering through small pore size membrane (e.g. 0.22 µm and 100 kDa) (Paul *et al.*, 1991). Tangential flow filtration (TFF) is now being used and the technology has the ability to filter and concentrate large water samples (Casas and Rohwer, 2007).

The use of PCR to study the diversity of uncultured viral communities has always been limited by lack of universal gene markers (e.g. 16S rRNA in bacteria) in phage genomes (Maniloff, 1995). In the early 2000s the phage proteomic tree was developed based on the overall sequence similarity of predicted protein sequences of completely sequenced phages (Rohwer and Edwards, 2002). This genome-based taxonomy classifies phages according to their sequences; consequently identifying conserved genes or sequences in phage genomes. The detection and study of different groups of phages can therefore be performed using gene sequences that are specific to a viral group. Denaturing gradient gel electrophoresis (DGGE) can be coupled with the above PCR-based method depending on the specific study. The method is used within the same viral groups (e.g. same gene) and a band on the gel can correspond to a different cluster of the same phage group (Dorigo *et al.*, 2004).

In this study the viral population from Loch Logan was studied using microscopic and sequenced based techniques. The efficiency of TEM to identify different phage morphotypes from environmental samples was evaluated using samples from water and sediments. Primers specific to different groups of phages were also used in a PCR for the detection of these phages. The study was aimed at using Loch Logan as model sample before all these techniques can be applied to the mine samples where microbial population is not expected to be high. The presence of *E. coli* and other coliforms in high numbers from this pond increases the chances of phages being available high in quantities, making this pond an ideal source of material for optimization of techniques.

Methods for isolation of viral-like particles, viral DNA extraction, microscopic techniques and PCR amplifications from water and sediments, were developed for fissure water and biofilms, respectively. However this study will serve as the basis and if further optimization is required when working with mine samples it will still be done, as lower concentrations for both viral particles and DNA are expected.

2.2. Material and Methods

2.2.1. Microbial strains and growth conditions

E. coli Top 10 strain and pGem-T-Easy vector (both from Promega) were used as the cloning host and vector, respectively. All the cultures were harvested in LB medium (10 g NaCl, 10 g Tryptone and 5 g Yeast extract per liter), and when solid media was required, 15 g/L of agar was added. LB media for blue/white selection contained Isopropyl β-D-1-thiogalactopyranoside (IPTG) and 5-Bromo-4-chloro-3-indolyl-β-D-galactopyranoside (X-gal) to final concentrations of 0.2 mM and 40 μg/ml respectively. The final concentration of ampicillin used was 100 μg/ml.

2.2.2. General recombinant DNA techniques

2.2.2.1. Plasmid DNA isolation

Manual plasmid DNA isolation was done using the lysozyme boiling method as follows; 5ml LB containing the appropriate concentration of an antibiotic was inoculated with a single colony and incubated at 37°C overnight. The cells were resuspended in 400 μl STET buffer (8 % w/v sucrose, 0.1% v/v Triton X-100, 50 mM EDTA and 50 mM Tris-HCl, pH 8) after centrifugation at 14 000 RPM using the HE111 rotor (Eppendorf 5810 centrifuge). To lyse the cells, 10 μl of lysozyme (10 mg/ml) stock was added and the cell suspension and incubated at 37°C for 20 minutes, then placed in a boiling water bath for 60 seconds. The lysate was cooled on ice for 10 minutes then centrifuged as described to separate the cell debris. The supernatant was transferred to a new microcentrifuge tube; 400 μl of cold isopropanol was added then mixed by inversion then centrifuged for 10 minutes. The pellet was washed by adding 1 ml of 70 % ethanol and then centrifuged for another 5 minutes. The supernatant was discarded and the remaining ethanol was evaporated in a vacuum dryer. The plasmid pellet DNA was resuspended in 100 μl of 50 mM Tris-HCl buffer, pH 8.5 containing RNAse.

Plasmid DNA was purified using the Bioflux DNA/RNA extraction/purification for Biospin plasmid DNA extraction kit (Separation Scientific) when pure DNA was required.

2.2.2.2. *PCR reactions and conditions*

PCR amplifications were done using *Taq* DNA polymerase with thermostable buffer (New England Biolabs) unless stated otherwise; all PCR products were excised from the agarose gel and DNA bands were purified using Biospin Gel Extraction kit (Separation scientific) according to the manufacturer's instructions.

2.2.2.3. *DNA manipulations*

DNA digestions and ligation reactions were done as described by Sambrook and Russell (2001) using restriction endonucleases and T4 DNA ligase from Fermentas Life Sciences. The DNA sequencing was done at Inqaba Biotech (South Africa) or at the University of the Free State, Department of Microbial, Biochemical and Food Biotechnology. The sequencing PCR was performed with Big Dye® terminator v3.1 Cycle sequencing kit (Applied Biosystems) the samples were run and analyzed using 3130x / Genetic Analyzer (Applied Biosystems).

2.2.2.4. *Agarose gel electrophoresis*

DNA fragments were separated on 1 % agarose gel dissolved in 1x TAE buffer (242 g Tris, 50 mM EDTA and 57.1 ml glacial acetic acid, per liter for a 50X buffer). Ethidium bromide was added to a final concentration of 1 μg/ml and prior to electrophoresis 10x bromophenol dye was added to the samples. Mass ruler (lower range) from Fermentas Life Sciences was used as the DNA ladder for all the agarose gels.

2.2.2.5. *Bacterial transformation*

E. coli Top 10 competent cells were prepared using the rubidium chloride method (http://micro.nwfsc.noaa.gov/protocols/rbc.html). A single colony was used to inoculate 100 ml Psi media (5 g yeast extract, 20 g tryptone and 5 g magnesium sulfate per litre, pH 7.6) followed by incubation at 37°C until OD_{550nm} of 0.48. The cells were cooled on ice for 15 min before centrifugation at 5000 x *g* for 5 min (Beckman J2-MC, JA 14 rotor) and then washed with 40 ml Tfb I buffer (potassium acetate 30 mM, rubidium chloride 100 mM, calcium chloride 10 mM, manganese chloride 50 mM and glycerol 15 % v/v, pH 5.8). The pelleted cells were resuspended in 4 ml Tfb II buffer (MOPS 10 mM, rubidium chloride 10 mM, calcium chloride 75 mM, manganese chloride 50 mM and glycerol 15 % v/v, pH 6.5), and 50 μl aliquots were stored

at -80°C. For transformation, competent *E. coli* cells (100 µl) were mixed with the appropriate amount of the plasmid or ligation mixture then incubated on ice for 20 min. The transformation mixture was heat-shocked at 42°C for 40 sec, followed by cold shock on ice for 5 min. LB broth (600 µl) was immediately added and the mixture was incubated for 1 hour with shaking at 37°C. The transformants were selected on LB agar medium containing appropriate antibiotic concentration.

2.2.3. Sampling

Sediments and water samples were collected from Loch Logan pond on 19th of March 2007; sediments were scooped into a sterile 50 ml falcon tube. Autoclaved 2L shot bottles were used to collect water samples. Upon arrival pH values of the water samples were measured. All the samples were processed immediately or stored at 4°C until processed. However all samples were still processed within 14 days after sampling date. Samples were labeled LLS and LLW for (S) sediments and (W) water, respectively.

2.2.4. Isolation of phage particles from sediments

Viral particles were isolated from soil/sediments using a combination of methods by Sambrook and Russell (2001) and Williamson, *et al.* (2005). Phosphate buffered saline, (PBS) (pH 7.4) containing 0.5 % w/v sodium dodecyl sulphate (SDS) was added to the soil sample (10 % w/v). NaCl was also added to the mixture to a final concentration of 1 M, followed by brief vortexing. The mixture was sonicated (Bandelin Sonopuls, 50 % power) at 4°C for 5 min (with each minute interrupted by manual shaking for 30 s). The debris was removed by centrifugation at 11 000 X g for 10 min (Beckman J2-MC, JA 14 rotor) at 4°C. The supernatant was transferred to a clean flask followed by precipitation of the viral particles by addition of 10 % (w/v) PEG 6000. The viral suspension was incubated at 4°C for >12 hrs, and viral particles were pelleted by centrifugation at 13 000 X g for 30 min (Beckman J2-MC using JA 14 rotor). The pellet was dried and resuspended in minimal volume of milli Q water that was filtered through a 0.02 µm membrane filter.

2.2.5. Concentration and purification of phage particles from water

Water samples were pre-filtered through glass fiber (100 µm) and 0.45 µm cellulose acetate filters (Millipore) to remove large particles. Viral particles suspended in water samples were concentrated and purified using a tangential flow filtration (TFF) system. Prior to concentration new filters were washed according to the manufacturer's instructions. Water samples was first filtered through 0.2 µm (Amersham Biosciences, CFP-2-E-4A) and then through 100 kDa (Amersham Biosciences, UFP-100-C-4A) TFF cartridges. The first cartridge removes bacteria, whereas the second cartridge concentrates viruses in the retentate. The final retentate volume was approximately 50 mL. After filtration the aliquots from 100 kDa retentates were applied to a stepwise cesium chloride gradient consisting of 2 ml of 1.50, 1.25 and 1.15 g/ml and ultra-centrifuged at 22 000 rpm using the Beckman SW 32.1 Ti rotor in clear Quick Seal tubes at 4°C for 5 hrs. After centrifugation about 2 ml of the phage suspension was recovered between the 1.25 and 1.5 g/ml layers and then dialyzed overnight against PBS. The buffer was changed twice.

2.2.6. Enumeration of viral-like particles

2.2.6.1. *Epifluorescence microscopy (EFM)*

Viral like particles stained with SYBR Gold were quantified using epifluorescence microscopy (Noble and Furhman, 1998). Aliquots (500 µl) from 100 µm, 0.2 µm and 100 kDa concentrates were fixed using 2% formaldehyde (filtered through 0.02 µm Anodisc filter, Millipore) at 4°C overnight. Anodisc filters (0.02 µm) were used to prepare the samples for staining; filters were placed on a pre-wetted 0.8 µm filter (Millipore) on a 25mm glass filter unit. Fixed aliquots were then filtered until dry. The 0.02 µm Anodisc filters were removed and placed on a 100 ul drop of 5% SYBR Gold (5 µl diluted stock + 95 µl 0.02 µm filtered water, made fresh from 1:10 stock that can be frozen for a week) in a petri dish with the sample side facing up. Controls consisted of a filter with no treatment and a filter through which water and gluteraldehyde (both filtered through 0.02 µm) had been filtered. Viral suspensions from sediments and water samples were first fixed with 2% gluteraldehyde for three hours at 4°C, followed by staining for 15 min in the dark. After staining, filters were dried then placed on a 30 µl drop of anti-fading solution, Citiflour mountant media (TED PELLA, INC.) on a slide cover slip, followed by visualization under blue excitation light and 100X objectives using epifluorescence

microscope. The pictures were taken with the digital camera connected to the microscope (Kodak digital camera, 4 megapixels).

2.2.6.2. *Transmission Electron microscopy (TEM)*

Aliquots of viral suspensions from sediments and water samples were first fixed with 2 % gluteraldehyde for three hours at 4°C and 10 µl of the phage suspension was then overlaid on formvar, carbon coated grid (400 mesh). The suspension was allowed to dry on the grid, which was then negatively stained with 3% uranyl acetate. The excess stain was removed using filter paper cut in to small pieces and the grid was allowed to air-dry prior to examination using Philips (FEI) CM100 TEM.

2.2.7. Isolation of DNA from viral particles

DNA was extracted from the isolated viral-like particles using a modified formamide method as described by Sambrook and Russell (2001). The volume of phage suspension was measured and 0.1 volume of 2 M Tris (pH 8.5), 0.5 volume of 0.5 M EDTA (pH 8.0), and 1 volume of deionized formamide were added. The mixture was incubated at 65 °C for 30 min, followed by precipitation with 6 volumes of ice cold ethanol for 30 min at -80 °C or overnight at -20 °C. The sample was centrifuged at 10000 x g for 10 min and the pellet was re-dissolved in 300 µl of TE buffer (pH 8). DNA was re-precipitated by addition of 6 µl of 5M NaCl and 750 µl absolute ethanol, followed by incubation at -80 °C for 30 min. The sample was centrifuged as described and the pellet was allowed to air-dry at room temperature. Viral DNA was then resuspended in 50 µl TE buffer.

2.2.8. Detection of different groups phages by PCR

Primer sets specific for different groups of phages (Table 1.1) were used to detect respective phages from the environment. All PCR amplifications were done using the isolated viral DNA as the template. Primers MZIA1bis and MZIA6 were used for the detection of g23 (a major capsid protein) from T4-type phages (Filée *et al.*, 2005) using the following PCR conditions: initial denaturation for 2 min at 94°C, and 30 cycles of the following steps denaturation at 94°C for 45 sec, annealing at 50°C for 1 min and elongation at 72°C for 45 sec. The final elongation was done at 72°C for 10 min.

The gene encoding major capsid protein from Cynophages was detected using the following primer combinations, CPS3 and either CPS4 or CPS8 (Zong *et al.*, 2002). The PCR parameters were as follows; initial denaturation step of 94°C for 2 min, followed by 35 cycles of denaturation at 94°C for 30 sec, annealing at 36°C for 30 sec, ramping at 0.3°C/s, and elongation at 72°C for 1 min, with a final elongation step of 72°C for 10 min. To detect the DNA polymerase from Uncultured Podophages (PUP) clade (Breitbart *et al.*, 2004), the primer set, HectorPol29F and HectorPol711R was initially used. An additional reverse; primer HectorPol500R was also designed. A temperature gradient PCR was performed using the following parameters for both primer sets: 94°C for 5 min, followed by 35 cycles at 94°C for 1 min, annealing at 61°C -0.5°C/ cycle for 1 min, extension at 72°C for 1 min and final extension for 10 min at 72°C. All the PCR products were cloned into pGem-T-Easy, recombinant clones were then sequenced using T7 promoter or SP6 universal sequencing primers.

Table 2.1: Oligonucleotides used

Primer name	**Sequence from 5' to 3'**
HECTORPol29F	GCA AGC AAC TTT ACT GTG G
HECTORPol711R	TTC GTT GGT GTA TCT CTC G
HECTORPol500R	GAA TGA TCT ACA CTC TTT GCC ATA CGG TG
MZIA1bis	GAT ATT TGI GGI GTT CAG CCI ATG A
MZIA6	CGC GGT TGA TTT CCA GCA TGA TTT C
DGGT4F	CGCCCGCCGCGCGCGGCGGGCGGGGCGGGGGCACGGGGGG-TTC AGC CGA TGA CTG GYC CAA C
CPS3	TGG TAYGTY GAT GGM AGA
CPS4	CAT WTC WTC CCA HTC TTC
CPS8	AAA TAY TTD CCA ACA WAT GGA

2.2.9. Primer design

The additional reverse primer for the PUP clade was designed using the available partial DNA polymerase sequences from uncultured T7-like podoviruses. The new reverse primer, HECTORPol500R is internal to the previously constructed primer, HECTORPol711R (Breitbart *et al.*, 2004) and was designed using the more conserved region of the DNA polymerase (Figure 2.1).

Figure 2.1: Multiple alignment of the DNA polymerase fragment, the region that was used for primer design is highlighted in light blue and only the conserved part of the sequence is shown.

2.2.10. T4-type phage diversity using Denaturing Gradient Gel Electrophoresis (DGGE)

Denaturing gradient gel electrophoresis was performed with the T4-type major capsid protein. The forward primer containing a 40 bp GC-clamp was designed and the PCR products obtained with the primers lacking the clamp were used as the template. The same PCR parameters for the g23 amplification were used with increased annealing at 55°C. T4 PCR products were excised from the gel and purified prior to the DGGE analysis. The bands were then electrophoresed on an 8% acrylamide/bis gel in 1X TAE buffer using the DCode™ universal mutation detection system (Bio-Rad Laboratories). The urea-formamide gradient between 20% and 80% (100% is defined as 7M urea and 40% v/v acrylamide/bis) was used to separate the bands. The products were then electrophoresed at 60°C at a constant voltage of 200 V, after electrophoresis the gel was stained with EtBr (10 µg/ml) on a gyratory shaker for 10 min.

2.2.11. T4-like phages and phylogenetic analysis

Sequences obtained with the g23 clones from Loch Logan were used to construct a neighbor joining tree using MEGA4 (Tamura *et al.*, 2007). The evolutionary distances were computed using the Maximum Composite Likelihood method (Tamura *et al.*, 2004) and are in the units of the number of base substitutions per site. All positions containing gaps and missing data were eliminated from the dataset. There were a total of 388 positions in the final dataset. Sequences from the primers were included in the analysis.

2.3. Results and Discussions

2.3.1. Sampling site

Loch Logan pond is an impoundment that was created near the city centre of Bloemfontein (grid reference: 29°06'845" S and 26°12'505" E), and collects runoff water collected from urban areas through one of the Bloemspruit canals. The water has a slightly alkaline pH value of 7.4 at the time of sampling, and the temperature values cannot be reported as they were not recorded on site.

2.3.2. Enumeration of viral-like particles

Viral like particles (VLPs) were enumerated using EFM and TEM, the latter technique was mainly used to determine the morphology of phages. Phage particles were observed from water and sediments, including the concentrates that were obtained by CsCl gradients. Different families of dsDNA tailed phages were observed including *Myoviridae* with a short contractile tail. The capsid from this family has icosahedral symmetry and it is separated by a neck-like structure from its complex tail (Figure 2.2 A). Other morphotypes corresponded to that of *Siphoviridae* with long tails (Fig. 2.2 B and C) and *Podoviridae,* a family with short tails (Figure 2.2 D). Both families have non-contractile tails and isometric heads, and some of the *Siphophages* obtained were characterized by spikes (Figure 2.2 B). Isolated phages were of different sizes with the heads ranging from 80 nm in diameter for the *Siphoviridae* to 140 nm for *Myoviridae*. The tails were between 100 nm and 150 nm in length. The results show that tailed phages are the most abundant as described in literature and as obtained with many environmental samples where uncultured phages were being investigated. Polyhedral,

filamentous, and pleomorphic phages were not observed with TEM and this might be due to their presence in low concentrations.

Though TEM is very useful when determining mophorlogy of viruses, the method is however not effective for quantitative estimation of viral-like particles from the environment. Hence EFM was used to have an idea on the total number of the phage particles in the samples. After staining, fluorescence obtained with the negative controls indicated that there were no particles or background fluorescence on the filters (Figure 2.3 A & B). Large particles (e.g. prokaryotes) were observed with both 100 µm filtrate and 0.2 µm retentate, and viral-like particles with 100kDa retentate (Figure 2.3 D & E). High viral concentrations were observed with 100 kDa concentrate as expected; however the actual counting of the viral-like particles was not done as a grid was not available.

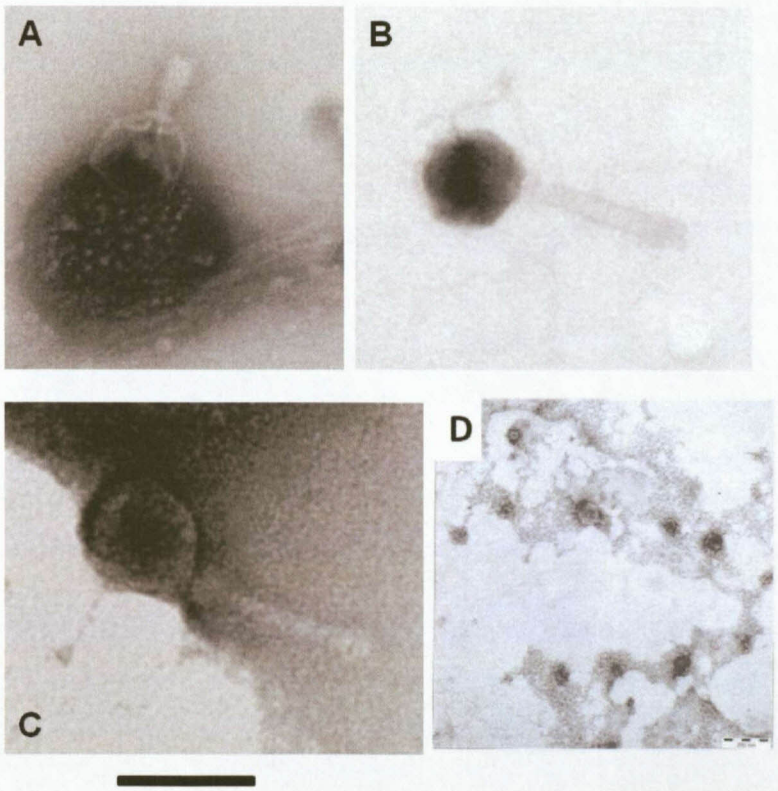


Figure 2.2: TEM pictures obtained with sediments and water samples. A = phage particles isolated from soil, B, C and D represent the 100 kDa retentate. The bar corresponds to 100 nm for pictures A, B and C.

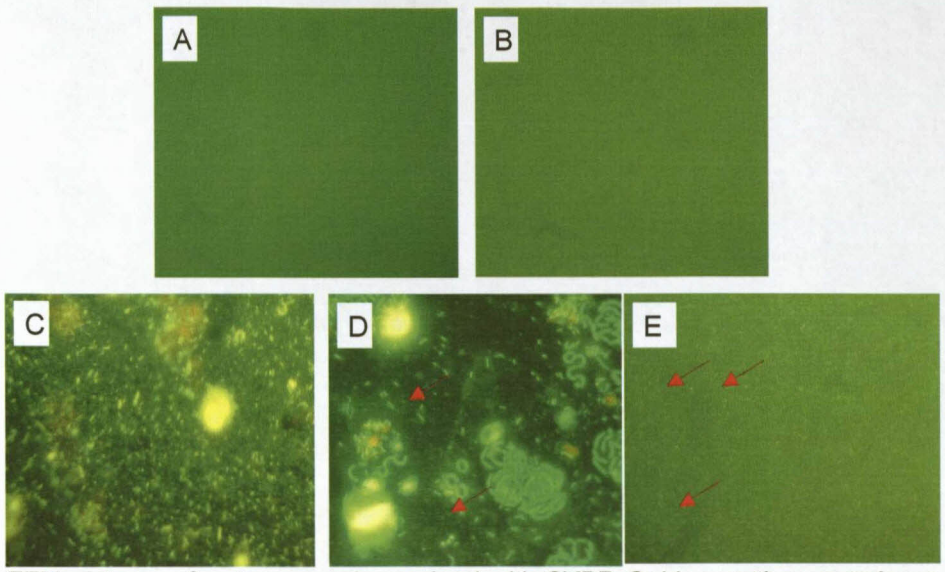


Figure 2.3: EFM pictures of waters samples stained with SYBR Gold, negative controls are designated on A and B. C, D and E represent 100 µm, 0.2 µm and 100 kDa retentates, respectively. Phage particles are indicated with arrows.

2.3.3. PCR detection of uncultured phage groups

Though viruses do not have universally conserved gene that can be used to study their diversity, different gene sequences have been identified which can be used to detect specific groups of phages from the environment (Rohwer and Edwards, 2002). In this study primers specific for g20, g23 and DNA polymerase fragment were used to detect different groups of phages. The first two genes encode major capsid proteins from Cynobacteriophage and T4-like phages, respectively. The last gene is conserved in uncultured T7-like phages. Amplification of g23 and DNA polymerase genes resulted in products with PCR indicating the presence of these phages as represented by these proteins from both water and sediments. Sequencing confirmed the results as the T4-clones hit with the major capsid protein from uncultured *Myoviradae* (Table 2.2). The primer set used for g23 detection can amplify this segment in all subgroups of T4-type phages, resulting in amplicons of different sizes. A PCR product of about 450 bp was obtained with both water and sediment (Figure 2.4 A), including the viral suspension obtained with CsCl gradient (Figure 2.4 B). The band indicates the presence of ExoT-even phages from this environment. However, the 600- and 640 bp DNA fragments corresponding to T-even and Schizo T-even respectively could not be detected by PCR. Though no amplifications were obtained with these sub-groups it does not mean that they are

not present in the viral suspensions. Low concentrations of viral DNA could be responsible for the lack of PCR products. In addition though phages are most abundant from the environment, DNA isolations from the environments have proved difficult resulting in low concentrations of viral DNA after extractions.

Several attempts to amplify the gene encoding g20 capsid protein from cyanophages failed. This could be due to the fact that the Loch Logan pond experiences periodic agal blooms which deplete nutrients and oxygen followed by limited or no growth of algae for certain periods; hence cynopahges could not be detected by PCR.

Initial detection of DNA polymerase from uncultured podophages provided no results after PCR when using the first set of primers (reverse HECTRPol711R). This necessitated the design of a new reverse primer based on the region of the polymerase gene that is highly conserved (Figure 1.1). The amplification with the new combination resulted in a PCR product of about 500 bp (Figure 2.5), consequently showing that the T7-type phages also form part of the viral population from this pond. In contrast to the T4-type phages the DNA polymerase fragment from the uncultured podoviruses was highly conserved with high sequence similarity among each other, even when compared to the clones isolated from marine environments (Figure 2.6) (Breitbart *et al.*, 2004). The T4-type and T7-type phages are classified as *Myoviridae* and *Podoviridae*, respectively the results obtained with sequencing are very significant as they correspond to those obtained with TEM; both families were observed when using this microscopic technique.

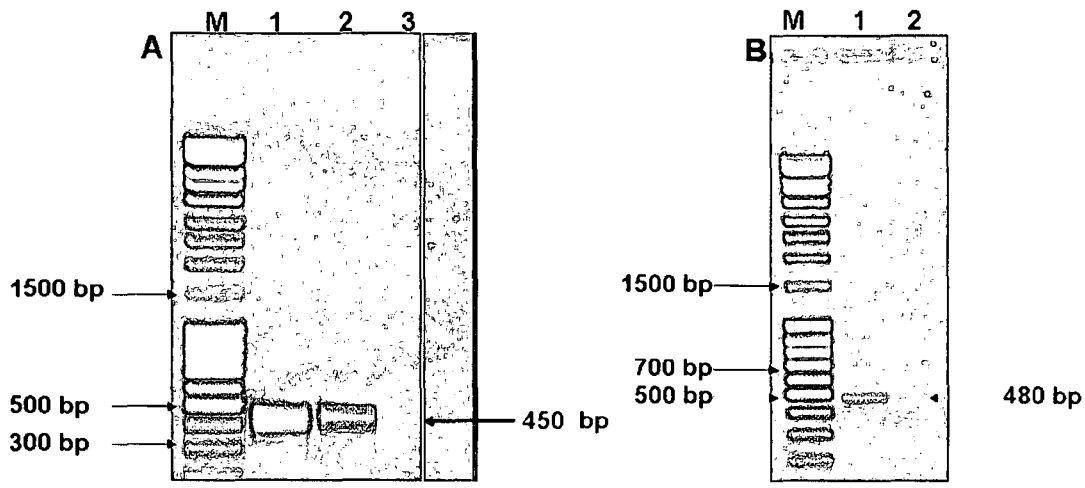


Figure 2.4: PCR amplification of the T4-type phages from Loch Logan, A) products obtained with water (100 kDa retentates) on lanes 1 and sediment on lane 2. The negative control is indicated on lane 3 and the DNA ladder with lane M. B) products from 100 kDa concentrate after CsCl gradient. The negative control is on lane 2 and the DNA ladder on lane M.

Table 2.2: T4 phage g23 protein hits from Loch Logan

Clone	Hit accession number	e-value
T4LL8R	DQ105935	$4e^{-46}$
T4LL19	AB284345	$6e^{-40}$
T4LL1R	GQ283502	$9e^{-61}$
T4LL6	AB505035	$2e^{-31}$
T4LL10	AB365588	$3e^{-99}$
T4LL3	DQ105894	$5e^{-45}$
T4LL5R	AB365587	$7e^{-94}$
T4LL4R	DQ105870	$2e^{-43}$
T4LL9R	DQ105932	$2e^{-30}$
T4LL2R	AB373675	$7e^{-32}$

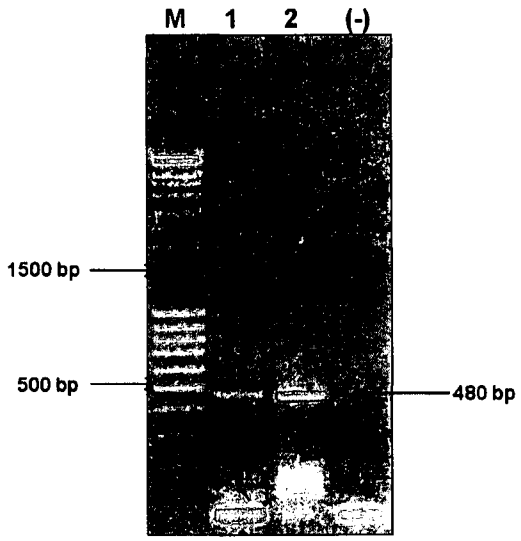


Figure 2.5: Detection of T7-like podoviruses from Loch Logan, PCR amplification of DNA polymerase using viral DNA from water (lane 1) and sediments (lane 2). The DNA ladder used and negative control are on lanes M and (-), respectively.

```
600058          ----------------------CAGGACATAACACACAAGATGCGACAGATAACGGAG 36
LLT76002        GATTGCAACGCAACTTTACTGTGGCAGGACATAACACACAAGATGCGACAGATAACGGAG 60
600054          ----------------------CAGGACATAACACACAAGATGCGACAGGTAACGGAG 36
599955          ----------------------CAGGACATAACACACAAGATGCGACAGATAACGGAG 36
LLT75002        GATTGCAAGCAACTTTT-CTGTGGCAGGACATAACACACAAGATGCGACAGATAACGGAG 59
LLT75001        GATTGCAAGCAACTTTTACTGTGGCAGGACATAACACACAAGATGCGACAGATAACGGAG 60
                                 *************************** ********

600058          CAACTACAGAAAGTGTTTCCACCAATAGTGGAGGAACGCTGGAGCGAGAAGACAGGGAAG 96
LLT76002        CAACTACAGAAAGTGTTTCCACCAATAGTGGAGGAACGCTGGAGCGAGAAGACAGGGAAG 120
600054          CAACTACAGAAAGTGTTTCCACCAATAGTGGAGGAACGCTGGAGCGAGAAGACAGGGAAG 96
599955          CAACTACAGAAAGTGTTTCCACCAATAGTGGAGGAACGCTGGAGCGAGAAGACAGGGAAG 96
LLT75002        CAACTACAGAAAGTGTTTCCACCAATAGTGGAGGAACACTGGAGCGAGAAGACAGGGAAG 119
LLT75001        CAACTACAGAAAGTGTTTCCACCAATAGTGGAGGAACGCTGGAGCGAGAAGACAGGGAAG 120
                ********************************** ********************

600058          CGACTGAAGGACAAGGTGACAGAGTTTAACGTAGGCTCTCGTAAGCAGATTGCAGAGAGG 156
LLT76002        CGACTGAAGGACAAGGTGACAGAGTTTAACGTAGGCTCTCGTAAGCAGATTGCAGAGAGG 180
600054          CGACTGAAGGACAAGGTGACAGAGTTTAACGTAGGCTCTCGTAAGCAGATTGCAGAGAGG 156
599955          CGACTGAAGGACAAGGTGACAGAGTTTAACGTAGGCTCTCGTAAGCAGATTGCAGAGAGG 156
LLT75002        CGACTGAAGGACAAGGTGACAGAGTTTAACGTAGGCTCTCGTAAGCAGATTGCAGAGAGG 179
LLT75001        CGACTGAAGGACAAGGTGACAGAGTTTAACGTAGGCTCTCGTAAGCAGATTGCAGAGAGG 180
                ************************************************************

600058          CTGGTAGGTGTAGGCGTTAAGTTCAAACAGAAAACTGAGAAGGGCGCTGTCATTGTCAAC 216
LLT76002        CTGGTAGGTGTAGGCGTTAAGTTCAAACAGAAAACTGAGAAGGGCGCTGTCATTGTCAAC 240
600054          CTGGTAGGTGTAGGCGTTAAGTTCAAACAGAAAACTGAGAAGGGCGCTGTCATTGTCAAC 216
599955          CTGGTAGGTGTAGGCGTTAAGTTCAAACAGAAAACTGAGAAGGGCGCTGTCATTGTCAAC 216
LLT75002        CTGGTAGGTGTAGGCGTTAAGTTCAAACAGAAAACTGAGAAGGGCGCTGTCATTGTCAAC 239
LLT75001        CTGGTAGGTGTAGGCGTTAAGTTCAAACAGAAAACTGAGAAGGGCGCTGTCATTGTCAAC 240
                ************************************************************
```

```
600058          GAGAAGGTGTTAGAAGGCATTGACATACCAGAGGCTAAGACGATATACGAGTACCTGATG 276
LLT76002        GAGAAGGTGTTAGAAGGCATTGACATACCAGAGGCTAAGACGATATACGAGTACCTGATG 300
600054          GAGAAGGTGTTAGAAGGCATTGACATACCAGAGGCTAAGACGATATACGAGTACCTGATG 276
599955          GAGAAGGTGTTAGAAGGCATTGACATACCAGAGGCTAAGACGATATACGAGTACCTGATG 276
LLT75002        GAGAAGGTGTTAGAAGGCATTGACAT-CCAGAGGCTAAGACGATATACGAGT-CCTGATG 297
LLT75001        GAGAAGGTGTTAGAAGGCATTGACATACCAGAGGCTAAGACGATAT-CGAGT-CCTGATG 298
                ************************** ****************** ***** *******

600058          TTGCAGAAGAGAGCAGCACAGATTGACTCTTGGCTAACCCACGAGAAAGATGGCAGGGTA 336
LLT76002        TTGCAGAAGAGAGCAGCACAGATTGACTCTTGGCTAACCCACGAGAAAGATGGCAGGGTA 360
600054          TTGCAGAAGAGAGCAGCACAGATTGACTCTTGGCTAACCCACGAGAAAGATGGCAGGGTA 336
599955          TTGCAGAAGAGAGCAGCACAGATTGACTCTTGGCTAACCCACGAGAAAGATGGCAGGGTA 336
LLT75002        TTGCAGAAGAGAGCAGCACAGATTGACTCTTGGCTAACCCACGAGAAAGATGGCAGGGT- 356
LLT75001        TTGCAGAAGAGAGCAGCACAGATTGACTCTTGGCTAACCCACGAGAAAGATGGCAGGGTA 358
                ************************************************************

600058          CACGGCAGAGTTATCACCAACGGTGCTGTAACAGGCCGTATGACTCACCACAGCCCTAAC 396
LLT76002        CACGGCAGAGTTATCACCAACGGTGCTGTAACAGGCCGTATGACTCACCACAGCCCTA-C 419
600054          CACGGCAGAGTTATCACCAACGGTGCTGTAACAGGCCGTATGACTCACCACAGCCCTAAC 396
599955          CACGGCAGAGTTATCACCAACGGTGCTGTAACAGGCCGTATGACTCACCACAGCCCTAAC 396
LLT75002        CACGGCAGAGTTATCACCAACGGTGCTGTA-CAGGCCGTATGACTCACCACAGCCCTAAC 415
LLT75001        CACGGCAGAGTTATCACCAACGGTGCTGTAACAGGCCGTATGACTCACCACAGCCCTA-C 417
                ***************************** ************************** *

600058          CTAGCACAAGTGCCGTCAGTATCTG---------------------------------- 421
LLT76002        CTAGCACAAGTGCCGTCAGTATCTGCACCCGTATGGCAAAAAAGTGTAGATCATTCTGGT 479
600054          CTAGCACAAGTGCCGTCAGTATCTG---------------------------------- 421
599955          CTAGCACAAGTGCCGTCAGTATCTG---------------------------------- 421
LLT75002        CTAGCACAAGTGCCGTCAGTATCTGCACC------------------------------ 444
LLT75001        CTAGCACAAGTGCCGTCAGTATCTGCACCGTATGGCAAAGAGTGTAGATCATTCAATCAC 477
                *************************
```

Figure 2.6: Nucleotide sequence alignment of the DNA polymerase fragment clones from Loch Logan and marine clones, accession numbers are used for the marine clones

2.3.4. Abundance of T4-type phages

Sequence analysis of the clones indicated that the fragments have conserved N- and C-terminal regions, with the variable part located in the middle of the protein (Figure 2.7). These variable regions demonstrate that the diversity of T4-type phages could be high. Approximately 5 clusters were obtained when doing phylogenetic analysis, and these were obtained with only 10 sequenced clones (Figure 2.8). Hence, DGGE analysis was done on both samples to estimate the diversity of these phages from this pond. Amplification with GC-clamped primers resulted in a band of approximately 490 bp from both water and sediments (Figure 2.9A). The DGGE profile revealed seven and five bands patterns for sediments and water, respectively (Figure 2.9B). The T4-type diversity from the sediments appeared greater using DGGE as it contained two bands more than water. The phylogenetic results and the DGGE bands could represent a genotype of the T4-like phages, though the bands were not sequenced. Since samples from this location were used for standardization of the methods only, very few clones were sequenced.

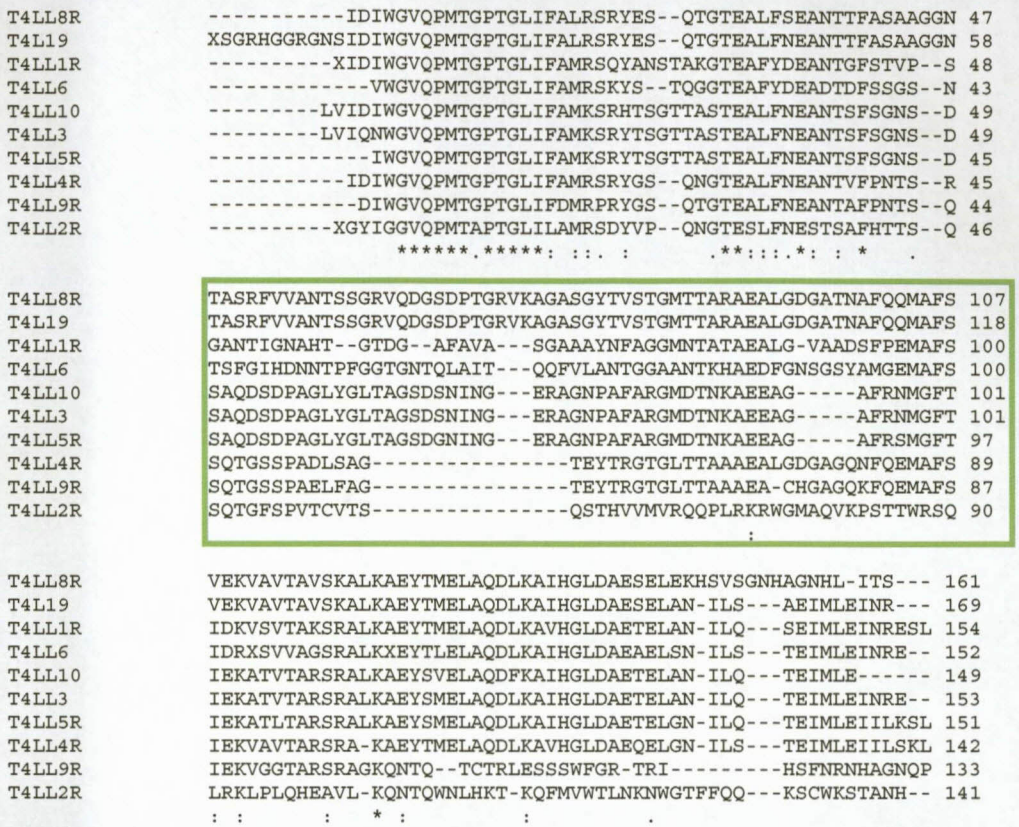


Figure 2.7: Amino acid sequence alignment of T4 phage g23 protein obtained with Loch Logan clones. The variable region is inside the green block

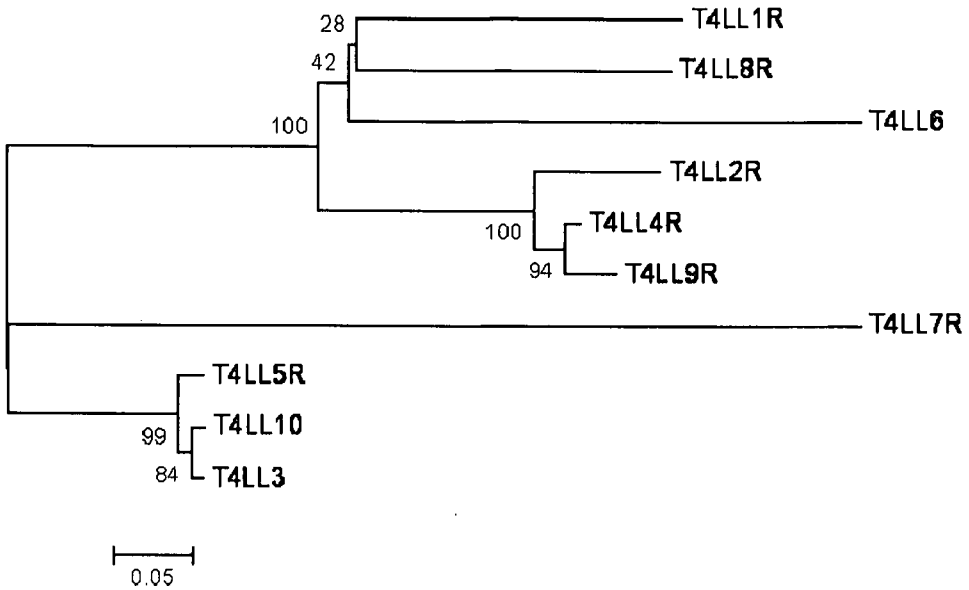


Figure 2.8: Evolutionary relationships of T4 clones from Loch Logan. The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (5000 replicates) is shown next to the branches. The tree is drawn to scale, with branch lengths in the same units as those of the evolutionary distances used to infer the phylogenetic tree.

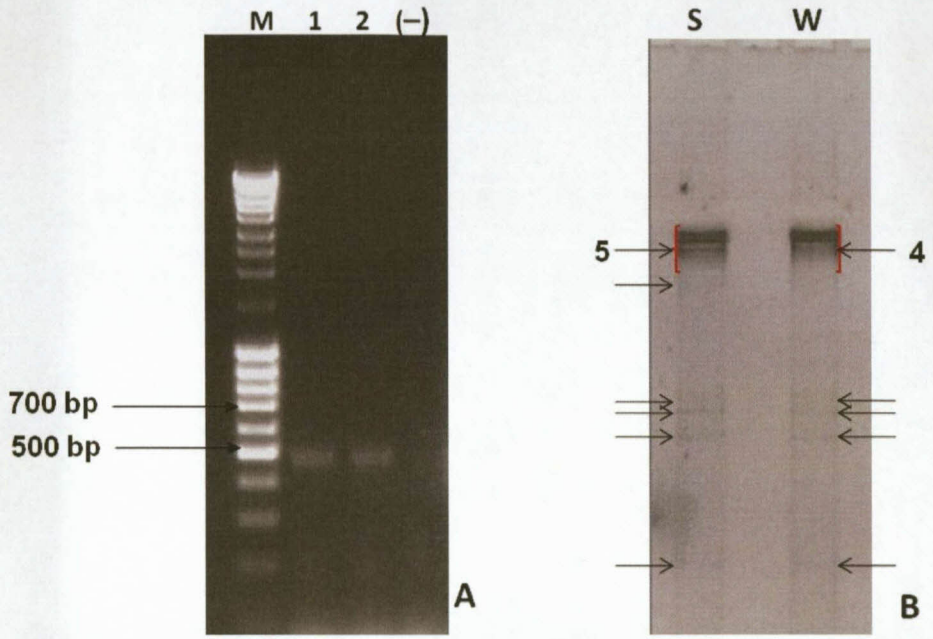


Figure 2.9: DGGE analysis of the g23 gene product from Loch Logan pond, (A) PCR products obtained with GC-clamped primers, lanes 1 and 2 are sediments and water, respectively. The negative control is represented with lane (-) and the DNA ladder in lane M. (B) The DGGE of the products amplified from sediments (S) and water (W). The arrows indicate the bands obtained with each sample and the numbers next to the arrows are the number of bands within the brackets.

2.4. Conclusions

Culture-independent techniques have been useful in identifying phages from the environment, and to date microscopic techniques and sequence based methods are mostly used for this purpose. Different researchers have therefore used these techniques to independently identify uncultured phages from various environments, especially the *Caudovirale*. In this study EFM and TEM were used to successfully observe and classify phages from Loch Logan pond. TEM results revealed the presence of tailed phages with various morphotypes, representative of all three major double-stranded DNA phage families (*Myoviridae*, *Siphoviridae* and *Podoviridae*). The results indicate that tailed phages are most abundant, as it has been reported in literature. Viral-like particles were also observed when using EFM, further confirming the high numbers of viruses from the environment. Though microbial diversity was not performed with our study, Vos and Roos (2005) reported high number of coliforms and *E.coli* counts during summer periods from this pond, hence high

numbers of phages could be expected in this aquatic environment. The results also show that TEM and EFM can be used effectively for morphology identification and counting of viruses from the environment, respectively.

Detection of different phages using PCR has proven to be useful in identifying different phage groups from the environment. Major capsid proteins from both T4-type phages and Cyanophages were detected by PCR. The major capsid protein is encoded by a homologue of the T4 gene 23 in all characterized T4-type phages, the primers used amplify a homologous segment of the g23 sequence in all subgroups of T4-type phages. The DNA fragment representing the presence of ExoT-even phages was obtained after PCR also with the CsCl gradient suspensions, showing that the isolation procedure used was efficient and could be applied in mining environments. The other two sub-groups were not detected in all aliquots. Failure to amplify the other sub-groups can be attributed to factors which could not be addressed in this part of the study. They include low concentrations of viral DNA and specificity of primers due to the plasticity of the T4-type phages. Sequencing results obtained with the few g23 clones confirmed the presence of T4-type phages from Loch Logan water and sediment. The pond contained a total of 7 genotypes as estimated by the DGGE analysis and phylogenetic studies showed that the T4-type phages are diverse.
The detection of the g20 from cynophages however resulted in no PCR products indicating that these phages are not present in this environment or they are available in concentrations that cannot be detected by PCR.

Uncultured T7-like phages were detected using a newly designed reverse primer as the previously designed primer resulted in no amplification, which was due to the fact that this oligonucleotide was designed using the region of the DNA polymerase fragment that is not highly conserved. Hence the second primer was designed (internal to the HECTORPol711R), based on the more conserved region of the DNA polymerase fragment. The PCR detection with this primer was more effective as products were detected with samples from both water and sediments. In addition the results demonstrate that this pair of primers would be suitable for the detection of phages from environments with low concentrations of phages due to its specificity.
The overall results show that culture-independent techniques are effective and can be used for the analysis of uncultured phages from the environment. Consequently these methods could be applied in the detection and morphology studies of phages from mine fissure water and biofilm samples.

Chapter 3

Uncultured T4-like and T7-like phages from four South African deep mines

Summary

Intensive studies done on phages from marine, sediments and oceans resulted insight on the diversity as well as the distribution of these phages. However, the presence of phages from deep mines has not been reported before. In this study we show the presence of uncultured T7-like and T4-like phages from fissure water collected from different South African mines. The major capsid protein (g23) was cloned to detect the latter phage group and the DNA polymerase fragment for the T7-like podoviruses. Very few g23 clones from the mines showed similarity to the known capsid proteins, indicating that the T4-like phages are highly mosaic. In contrast to the T4-like phages the DNA polymerase fragment had high similarity to the previously reported sequences which ranges from marine, freshwater, sediments, terrestrial extreme and metazoan environments, indicating that the gene is highly conserved. TEM studies also revealed the presence of tailed phages those showing morphology similar to *Myoviridae*.

3.1. Introduction

The South African mines represent the deepest excavations on earth and they offer unique opportunities to sample biofilm and water at great depths. Most of the gold mines in South Africa are in the Witwatersrand Basin. The basin comprises the following principal formations, the 2.9 Ga quartzites of the Witwatersrand Supergroup, the 2.7 Ga metamorphosed basalt, basaltic andesite of the Ventersdorp Supergroup and the sediments and volcanic strata of the 2.45 Ga Transvaal Supergroup (Bau *et al.,* 1999). Detailed geochemistry of the basin is described elsewhere (Onstott *et al.*, 2006).

Microorganisms have been found to inhabit every environmental niche examined, including deep mine environments. Microbial studies on deep mines around the world using a combination of 16S rDNA sequencing and DGGE or T-RFLP techniques have yielded valuable information about the diversity of the microbes (Nicomrat *et al.*, 2006, Sahl *et al.*, 2008). Microbial diversity from deep South African mines has previously been reported (Takai *et al.*, 2001, Onstott *et al.*, 2003, Pfiffner *et al.*, 2006). However, there are no reports on phage diversity from the deep mine sub-surfaces. Studies on phages have concentrated on marine environments as well as the oceans where phages are most abundant (Williamson *et al.*, 2008, Rohwer, 2003, Hendrix *et al.*, 1999). Other recent studies have begun to target non-culturable phages from environments such as soil (Williamson *et al.*, 2005); coral associated communities (Wegley *et al.*, 2007) and human infant gut (Breitbart *et al.*, 2008). In spite of this, phages are still an unexplored component of the microbial community. Current sequencing projects on phages shows that most of the phage proteins have no detectable homology to existing proteins. Furthermore, the sequencing project on Global Ocean Sampling (GOS) contained a higher proportion of viral sequences, indicating that to date phage sampling is still insufficient (Yooseph *et al.*, 2008).

Phage studies are limited by the following; firstly, phages require a specific host for propagation and 99% of microbes in the environment have not been cultured (Riesenfeld *et al.*, 2004, Hambly and Suttle, 2005). Secondly, sequence-based approaches are limited by lack of universally distributed genes and gene products (such as 16S rDNA, in bacteria and archaea) in phages (Maniloff, 1995). All these present a problem for the development of molecular methods for detection of phages from the environment in general. To enable comparative sequence analysis and identification, the phage proteomic tree was constructed (Rowher and Edwards, 2002, Angly *et al.*, 2006). The method highlights genes that are conserved in specific clades of phages that can be used for the detection of specific viral clades. Breitbart *et al.* (2004) reported

the detection of uncultured T7-like podoviruses using a DNA polymerase fragment. The other genes that are being used include g20 and g23 capsid proteins for the detection of Cynophages and T4-like phages, respectively (Zhong *et al.*, 2002, Dorigo *et al.*, 2004, Filée *et al.*, 2005). The intergrase gene is also being used as the as the molecular maker to analyze temperate phages from the environment (Balding *et al.*, 2005).

Tailed phages are the most abundant, accounting to about 96% of the total phages. In this chapter PCR was used to detect the presence of uculturable T7-like podophages and T4-like phages from deep mine fissure water collected from four different South African mines. All these mines are part of the Witwatersrand Basin, three mine gold and one Kimberlite intrusions. Microscopic techniques were also be used for morphology analysis and estimation of the viral counts.

3.2. Materials and Methods

3.2.1. Sites and sampling

Fissure water was collected from already existing boreholes from Beatrix (BE), Masimong (MM) (both Harmony Gold mining Company), Star Diamonds (SD) (Petra Diamonds) and Tau Tona (TT) (Anglo Gold Ashanti) mines in South Africa. The holes were sealed using steel valves and were sampled within a few weeks of striking the water, except for DPH3886 which was part of a water draining project which had continued for several months and LIC118 which had slowly been draining for about a year. The mines studied, except for Star Diamonds which mines Kimberlite intrusions in the Witwatersrand Supergroup, are gold mines which exploit conglomerate or carbon leader reefs. Aseptic sampling was performed as described by Pfiffner *et al.*, (2006). Samples from BM mine were collected in August 2007, in July and August 2007, respectively for DPH3886 and LIC118, and water samples from MM and SD were collected in February 2008 on different days. In line filters (Supor membrane, 0.1/0.1 um, Cole palmer) were also connected to the valve of DPH3886 to trap any viral particles. Upon arrival at the laboratory, water samples were immediately filtered and concentrated by tangential flow filtration (TFF) and half of the in line filters were also processed immediately, the remaining filters were stored at -80°C until they were required.

3.2.2. Processing of the water samples

Suspended viral particles in the water samples collected from different mines were aseptically concentrated using a TFF system as follows: samples were first filtered through 0.2 µm filter (Amersham Biosciences, CFP-2-E-4A) and then concentrated with 100 kDa NMWCU filter (Amersham Biosciences, UFP-100-C-4A). The TFF systems were washed as follows before filtering: washing solution (10 g/L NaOH and 200 µ/L of bleach) was passed through the filter followed by washing with 5 L of autoclaved distilled water, finally with 5 L milli-Q water (0.2 µm filtered) (Rohwer, 2005). The filters were washed before and after use, and each sample was processed with a clean filter to prevent any contamination of the mine water, and cross contamination from mine to mine.

3.2.3. Processing of in line filters

Operations were performed in the laminar flow hood with sterile equipment, half of the filter was used and the remaining was stored at -80°C until required. The filter was cut into small pieces and then immersed in PBS buffer (filtered through 0.02 µm filter) for an hour at 4°C, followed by vortexing to remove attached phage particles. The mixture was centrifuged at 5000 x *g* for 10 minutes (Beckman J2-MC, JA 14 rotor) and the supernatant was first filtered through a 0.45 µm filter (Millipore) to remove the particles from the filter and then filtered through a 0.2 µm filter (Millipore) to remove prokaryotes.

3.2.4. Transmission electron microscopy (TEM)

Prior to staining, all the viral suspensions were fixed with 2% gluteraldehyde for three hours at 4°C. Initially 10 µl of fixed aliquots were used for TEM analysis, and then an increased volume of the viral suspensions was used with modified ultracentrifuge tubes. Tubes were prepared using two component epoxy, which was placed in the bottom of a centrifuge tube then centrifuged as described by BØrsheim *et al.*, (1990) to create a flat surface. Concentrated viral suspensions (10 ml) were ultracentrifuged at 30 000 RPMs using the SW 32 Ti rotor (Beckman, Optima™ L-100XP BioSafe). Prior to centrifugation the carbon coated, Formvar film copper grids (200 mesh, Agar Scientific) were placed at the bottom of the tube. Particles attached on the grids were negatively stained with 3% uranyl acetate and then visualized using Philips (FEI) CM100 transmission electron microscope.

3.2.5. Phage DNA isolation

The concentrated phage suspensions were treated with DNAse I (5 U/ml) and RNAse A (50 µg/ml) (Both from Fermentas Life Sciences) prior to isolation of the phage DNA. Phage suspension containing both enzymes was incubated at 37°C for 30 minutes in the following buffer reaction (10 mM Tris-HCl (pH 7.5), 5 mM $MgCl_2$ and 0.1 mM $CaCl_2$). Both enzymes are inactivated by SDS with the DNAse I also being inactivated by EDTA, these chemicals are used in viral DNA isolation, hence viral DNA was isolated from concentrates using the formamide method as described in section 2.2.7 (Sambrook and Russell, 2001). Due to low recovery of the viral DNA, proteinase K was added to a final concentration of 20 µg/ml to enhance the removal of phage protein coats.

3.2.6. PCR detection of uncultured phages

Primers HectorPol29F (Breitbart *et al.*, 2004) and HectorPol500R (Table 3.1) were used to detect the olymerase from the PUP clade. PCR parameters were used as described in section 2.2.8. and the products were electrophoresed on a 0.1% (w/v) agarose gel.

The major capsid protein which is encoded by gene 23 in all characterized T4-type phages was detected using the primers, MZIA1bis and MZIA6 (Table 3.1). This pair of primer can amplify a homologous segment of the g23 sequence in all subgroups of T4-type phages (Filée *et al.*, 2005). Additional oligonucleotide primers, g23F and g23R2 specific for the same protein were designed as indicated in the section 3.2.7. The new reverse primer was also paired with MZIA1bis. The following the parameters were used for PCR: initial denaturation for 2 min at 94°C, and 30 of the following steps denaturation at 94°C for 45 sec, annealing at temperatures between 50°C and 58°C for 1 min and elongation at 72°C for 45 sec. The final elongation was done at 72°C for 10 min. The intergrase gene was used as conserved locus for the detection of lysogenic phages, using the following specific primers, 1F, 1R, 8F and 8R (Table 3.1). Similar PCR parameters as above were used with annealing at 40°C.

All PCR reactions were performed using New England Biolabs Taq polymerase and negative controls were run with all PCR reactions to make sure that the amplifications are from the mine water samples. Viral DNA isolated from different mines was used as the template for all of the above reactions. The products were purified from the gel using a Biospin Gel Extraction kit (Separation scientific) according to manufacturer's instructions.

Table 3.1: Oligonucleotides used

Primer name	Sequence from 5' to 3'
HECTORPol29F	GCA AGC AAC TTT ACT GTG G
HECTORPol500R	GAA TGA TCT ACA CTC TTT GCC ATA CGG TG
MZIA1bis	GAT ATT TGI GGI GTT CAG CCI ATG A
MZIA6	CGC GGT TGA TTT CCA GCA TGA TTT C
CPS3	TGG TAYGTY GAT GGM AGA
CPS4	CAT WTC WTC CCA HTC TTC
CPS8	AAA TAY TTD CCA ACA WAT GGA
1F	GTT ACM GGG CAR MGA GTH GG
1R	ATG CCC GAG AAG AYG TTG AGC
8F	TGC TTA TAA CAC CCT GTT ACG TAT
8R	CAG CCA CCA GCT TGC ATG ATC
T4g23F	CTGATCGCCTTCGATATTGCTGGTGTTCAG
T4g23R2	GGGTTAARACCGATYSCGTAACGAG

3.2.7. New T4-like primers

An additional set of primer specific for the g23 protein was designed using the available T4-like genome sequences, the g23 protein from these genomes were aligned. Conserved parts of this gene as indicated by the blocks A, B and C (Figure 3.1) were used for primer design. The primer, MZIA1bis (Filée *et al.*, 2005) aligned with the region in Block A indicating that this part of the sequence was used. Hence the additional primers were designed based on the regions that have not been used for primer construction. The sequence in block B was used for the forward primer; T4g23F and the block C for the reverse primer; T4g23R2. The primer sequences are indicated on Table 3.1.

```
                                                                                A
NC008515        AQAFG --------SFLTEAE IGGDHGYNATNIAAG QTSGAVTQIGPAVMG MVRRAIPNLI 107
NC000866        AQAFG --------SFLTEAE IGGDHGYNATNIAAG QTSGAVTQIGPAVMG MVRRAIPNLI 107
NC004928        AQAFG --------SFLTEAE IGGDHGYNAQNIAAG QTSGAVTQIGPAVMG MVRRAIPNLI 108
NC010105        SEAFG --------SFLTEAE IGGDHGYDATNIAAG QTSGAVTQIGPAVMG MVRRAIPHLI 105
NC007022        IEAFG --------QSLMEAE VAGDHGYDPTNIAAG QSSGAITNIGPAVIG MVRRAIPSLI 107
NC005135        IEAFG --------QSLMEAE VAGDHGYDPTNIAAG QSSGAITNIGPAVIG MVRRAIPSLI 107
NC008208        IEAFG --------QSLMEAE VAGDHGYDPTNIAAG QSSGAITNIGPAVIG MVRRAIPSLI 107
NC009821        VEAFG --------GFIAEAE VAGDHGYDASQIAAG QTTGAITNVGPAVIG MVRRAIPNLI 106
NC007023        VESFG --------GFLAEAE IAGDHNYDQTNIASG KSSGAITNIGPAVIG MVRRAIPNLI 107
NC005260        VNSMV DVKGRIEEARLAEAN IGGDHGYDATKIASG ETSGSITNVGPAVMG LVRRAIPQLI 115
                  :::             : **::.***.*:  :**:*:::*::*::****: *:****** **
```

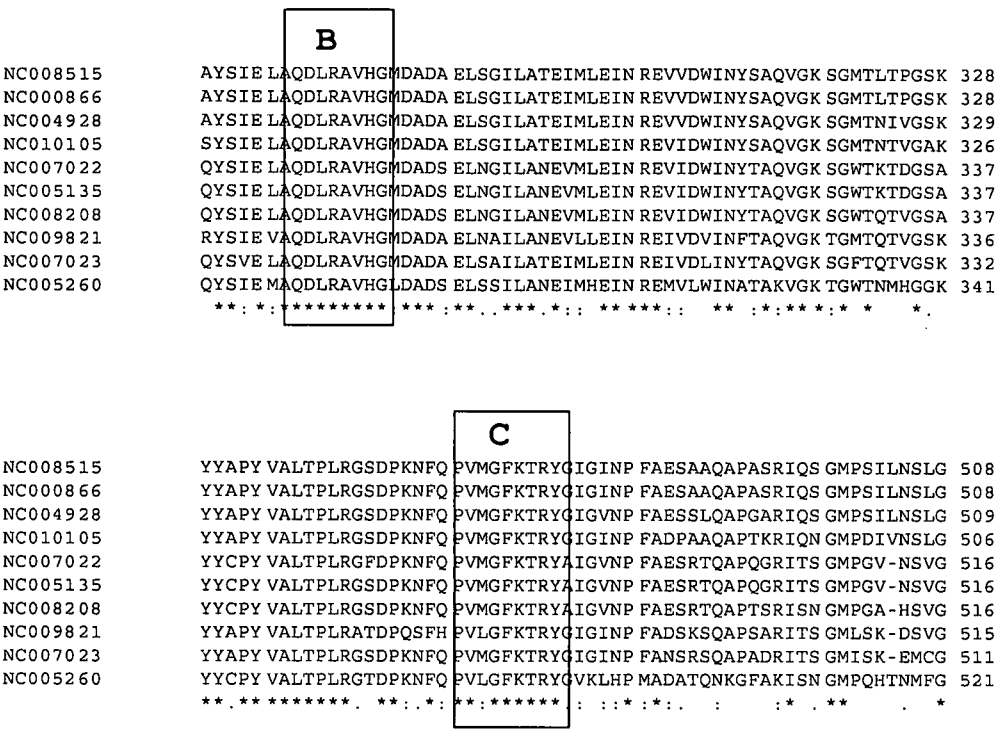


Figure 3.1: Amino acid sequence alignment of the g23 protein from T4-like phage genomes. Sequences used for primer designs are in block A, B and C; and only the regions that were used are indicated and not the complete gene. Accession numbers are used to identify different genomes where g23 sequences were obtained. Primers in blocks B and C were designed in this study, and the block A was used by Filée *et al.*, (2005) for the forward primer design.

3.2.8. Sequencing

Purified PCR products were sub-cloned into pGem-T-Easy (Promega), and ligation mixtures were used to transform *Escherichia coli* (TOP10) competent cells. Plasmid DNA was purified using the Bioflux DNA/RNA extraction/purification for Biospin plasmid DNA extraction kit (Separation Scientific) and clones were sequenced using T7 promoter or SP6 universal sequencing primers. The DNA sequencing was done at Inqaba Biotech (South Africa) or at the University of the Free State, Department of Microbial, Biochemical and Food Biotechnology. The sequencing PCR was performed with Big Dye® terminator v3.1 Cycle sequencing kit (Applied Biosystems) the samples were run and analyzed using 3130x / Genetic Analyzer (Applied Biosystems).

3.2.9. Phylogenetic analyses

A neighbor joining phylogenetic tree of the HectorPol DNA polymerase of T7-like podophage clones was constructed using MEGA 4 (Tamura *et al.*, 2007), excluding the primer sequences, for the tree construction. The bootstrap consensus tree inferred from 5000 replicates was taken to represent the evolutionary history of the taxa analyzed. Branches corresponding to partitions reproduced in less than 61% bootstrap replicates were collapsed. The tree is drawn to scale, with branch lengths in the same units as those of the evolutionary distances used to infer the phylogenetic tree. The evolutionary distances were computed using the Maximum Composite Likelihood method and are in the units of the number of base substitutions per site. All positions containing gaps and missing data were eliminated from the dataset.

3.3. Results and Discussions

3.3.1. Description of Sites

Basic chemistry of the water was performed on site and the information is supplied in Table 3.2. Beatrix Mine (BE) is situated at latitude 28°15'S and longitude 26°47'E near the towns of Welkom and Virginia about 240 km southwest of Johannesburg, in the Free State Province and it consists of three shafts. Masimong Mine (MM) is located approximately 280km south-west of Johannesburg, and 10 km from Virginia. The complex comprises two shafts: Masimong 4 and 5 Shafts. The reefs mined at Masimomg are the Basal Reef, B Reef and A Reef. The mine is situated at latitude 28 °00' S and longitude 26°30' E on the northern limit of the Welkom Goldfields in the Free State Province. The Star diamond (SD) mine property covers approximately 1,034 hectares in the Theunissen district in the Free State at approximately 28°19' S 27 °47' E. The mine exploits a series of kimberlite fissure segments over an east-west trending strike length of 4.5 kilometres which is part of a more extensive 15 kilometres long series of fissures. The individual fissures range in width between 5 cm and 80 cm. TauTona (TT), the deepest mining operation in the world, is located on the West Wits Line near the town of Carletonville, approximately 70 km west of Johannesburg in the Gauteng Province at 26°10' S and 27°26' E. Two holes were sampled in Tau Tona mine. DPH3886 (TT100) was drilled through sedimentary rocks of Central Rand Group of the Witwatersrand Supergroup. The hole intersects the Carbon Leader Reef and associated Marker Beds at the

base of the Central Rand Group. LIC118 (TT118) was drilled through the West Rand Group of the Witwatersrand Supergroup. The hole intersects the Carbon Leader Reef and associated Marker Beds at the base of the Central Rand Group. The holes are 2032 meters apart horizontally and 488.16 m apart vertically. More detail is supplied in Table 3.2.

Table 3.2: Sampling site information

Mine sample	T (°C)	Vol. (L)	Depth (kmbs)	Co-ordinates of hole collar
Beatrix	33.6	36	1.176	X:24536 Y:21837 Z:193.6*
Masimong	42	50	1.73	Not available
Star diamonds	32	10	0.560	Y: 20805 X:133430 Z: 909.358*
Tau Tona DPH3886	44.5	21	3.1	X:46031.585 Y:26840.814 Z:3013.840 Final depth 309 m
Tau Tona LIC118	52	20	3.6	X:44774 Y:28332 Z:3502 Final depth 869 m
Tau Tona DPH3886 (0.1 μm)	44.5	820	3.1	
Tau Tona DPH3886 (0.1 μm)	44.5	2 670	3.1	

* Z coordinate of these mines (BE and SD) is height above mean sea level

3.3.2. TEM

Initial attempts to detect the presence of phages from the mines when using the method described in chapter 2, section 2.2.6.2 were not successful. The method was therefore modified to accommodate low phage counts from the mines, by using modified ultracentrifuge tubes. However when using this method the film on the grids was sometimes damaged leaving most of the grid with almost no film for visualization. After several attempts viral-like particles were obtained with a sample from Beatrix mine, the other mines resulted in no viral like particles for TEM. Phage morphology obtained with water samples from Beatrix resembled that of *Myoviridae* (Figure 3.2) indicating the presence of this family of phages from the mine. Viral like particle counts from the mines were very low as compared to the high quantities obtained with the Loch Logan samples and counts could not be obtained using EFM. Hence, phages could not be observed with most of the mines, however these results do not necessarily mean viral like particles are not present from these environments. Therefore the use of PCR was important as it can detect fragments low concentrations of DNA.

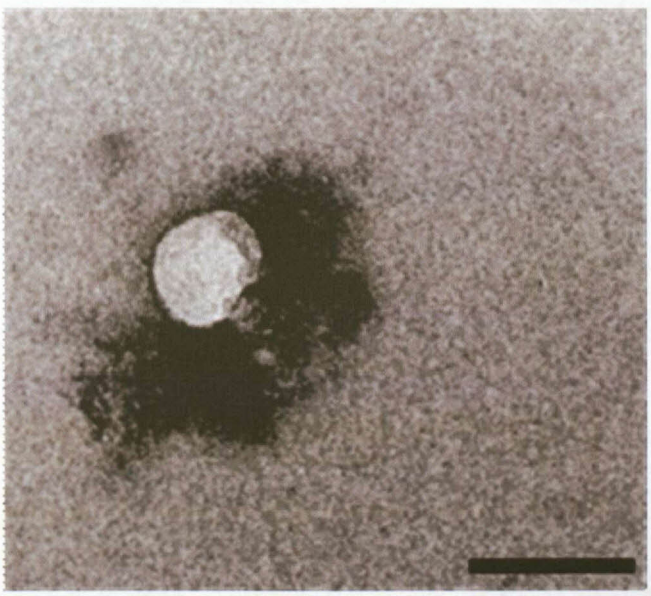


Figure 3.2: TEM picture obtained with Beatrix mine fissure water sample, the bar corresponds to 200 nm.

3.3.3. Abundance of uncultured T4-like phages

PCR amplification of T4-like g23 fragment resulted in a band of an approximately 490 bp (Figure 3.3) corresponding to ExoT-even phages, with all mine water samples. The 600 and 640

bp DNA fragments corresponding to T-even and Schizo T-even respectively could not be detected by PCR. A total of 120 clones were sequenced and when performing NCBI BlastX six of the clones from TTLIC118 and two from BM mine hit with the known uncultured *Myoviridae* g23 capsid protein (Table 3.3). The following accession numbers, GQ283496 to GQ283503 correspond to the T4-like g23 capsid protein fragment sequences which were submitted to the GenBank database. The rest of the clones showed no similarity to the available capsid proteins in databases, suggesting that the T4-like phage community in the deep mine sub-surfaces is different from the marine one. The clones were however showing similarity to the adenylyltransferase gene from a sulphate reducing bacteria. Hence the same primers were used in a PCR reaction using the genomic DNA from the sulphate reducing bacteria as the template. These sulphate reducing bacterial strains were cultured from the same mine samples using the retentate from the 0.2 µm TFF system. No PCR products were obtained after several attempts indicating the amplification is not from the sulphate reducing bacteria but from phage origin. The fragment could have been introduced by horizontal gene transfer (Comeau *et al.*, 2007, Williamson *et al.*, 2008).

Furthermore the results suggest that the major capsid protein from the mines is different from the marine communities and that the T4-like phages are highly diverse as very few clones hit with the available sequences. Similar results were obtained when detecting T4-like phages from the rice field using the same gene (Jia *et al.*, 2007). Only 8% of their clones from soil showed similarity to the available g23 capsid protein and 5.3% from rice straw. In addition, environmental major capsid proteins have been shown to be more divergent (deeper branching) than the vast majority of the major capsid protein sequences coming from the other sources (Comeau and Krisch, 2008).

The tyrosine recombinase gene was used as the conserved locus for the detection of lysogenic phages, however several PCR attempts to amplify this fragment from the mine samples were not successful (Table 3.3). Different PCR parameters and reactions resulted in no amplification.

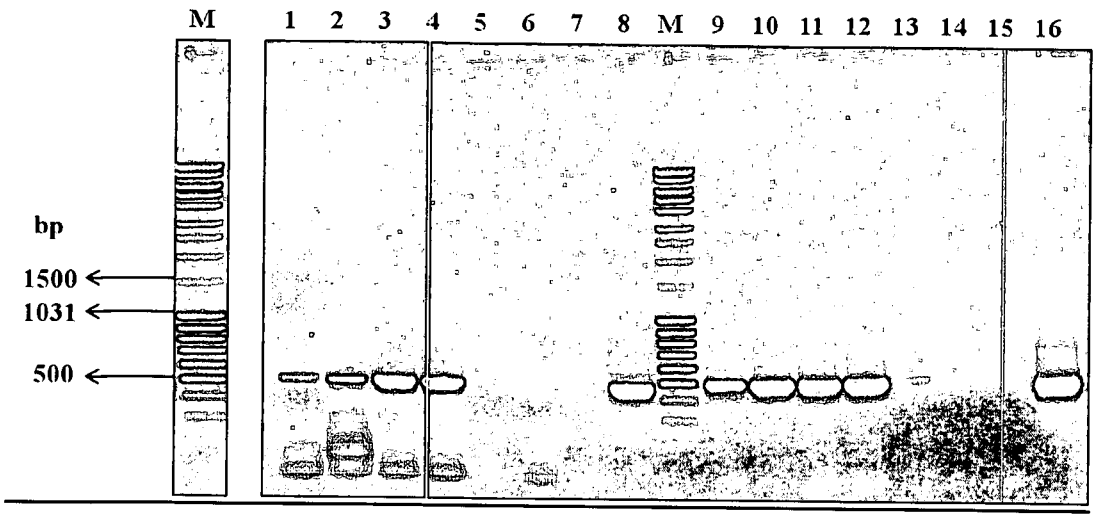


Figure 3.3: PCR amplification of T7-like and T4-type phages, the DNA ladder (Fermentas Mass ruler) used is designated as lane M; T7-like phages are represented on lanes 1-6 and T4-like phages on 9-14. Negative controls are on lanes 7 and 15, and positive controls on lanes 8 and 16. Viral DNA from the following mines was used as the template, MM (1 & 9), SD (2 & 10), BM (3 & 11), TTDPH3886 (4 &12) and TTLIC118 (5, 6, 13 & 14). The numbers in the brackets are the lanes that correspond to the specified samples.

Table 3.3: phage population of South African mines, the presence of a specific phage group is indicated by √ and the groups that were not detected are indicated by X.

		Mine samples		
	TT[1]	**BM**[2]	**MM**[3]	**SD**[4]
Lysogenic phages	X	X	X	X
T4-like phages	√	√	X	X
T7-like podoviruses	√	√	√	√

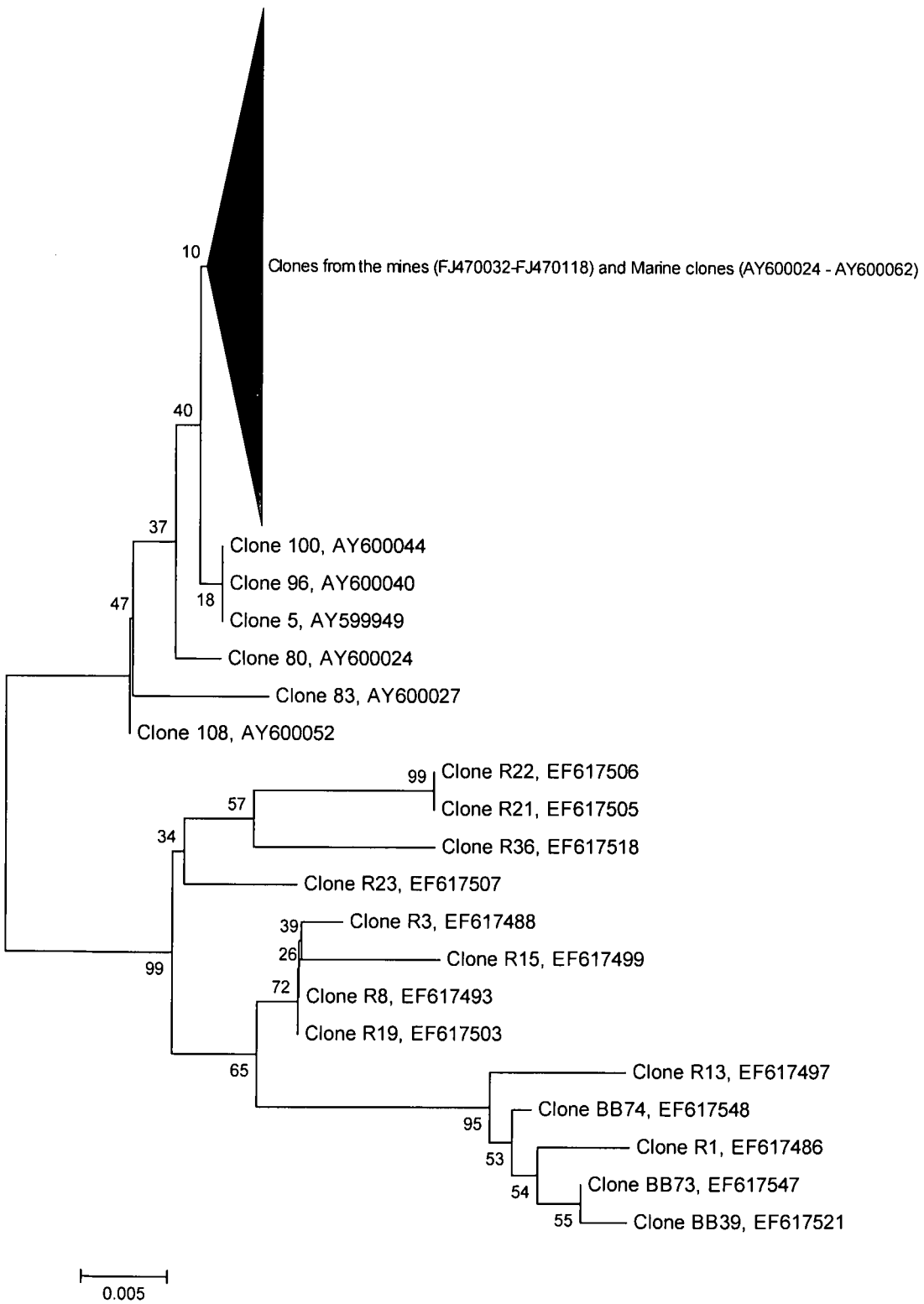
1 – Tau Tona, 2 – Beatrix mine, 3 – Masimong and 4 – Stardiamonds

3.3.4. T7-like Phylogenetic analyses

Primers specific for a DNA polymerase gene fragment from T7-like podoviruses were used with the viral DNA isolated from different mines as the template. A PCR fragment of 500 bp was amplified (Figure. 3.3) from all mine samples, indicating the presence of T7-like podoviruses in deep mine fissure water. A total of 145 T7-like clones were sequenced from Beatrix, Tau Tona (DPH3886 and LIC118), Masimong and Star Diamonds mines. Similarity searches done using BlastX (Altschul *et al.*, 1990) indicated that all T7-like clones contained the expected DNA polymerase fragment. The T7-like DNA polymerase fragment sequences have been submitted to the GenBank database under accession numbers FJ470032 to FJ470118. These results indicate the presence of uncultured T7-like podoviruses from the deep mine fissure water and that they are ubiquitous.

A neighbor joining tree (Figure 3.4) constructed with the sequences indicated that our clones had high similarity to the available T7-like sequences, which originated from different sources, ranging from marine, freshwater, sediments, terrestrial extreme and metazoan environments.

Figure 3.4 (next page): Phylogenetic tree of DNA polymerase using clones from the following mines, Beatrix, Star diamonds, Tau Tona (levels DPH3886 and LIC118) and Masimong. The compressed part consists of the clones from above mentioned mines and marine clones. The accession numbers are indicated in the brackets. Accession numbers for other sequences obtained from the database are also indicated on the tree. The tree is drawn to scale, with branch lengths in the same units as those of the evolutionary distances used. Bootstrap consensus tree inferred from 5000 replicates was taken to represent the evolutionary history of the taxa analyzed. The percentage of replicate trees in which the associated taxa clustered together in the bootstrap is shown next to the branches.


10
Clones from the mines (FJ470032-FJ470118) and Marine clones (AY600024 - AY600062)
40
37
Clone 100, AY600044
Clone 96, AY600040
18
47
Clone 5, AY599949
Clone 80, AY600024
Clone 83, AY600027
Clone 108, AY600052
99
Clone R22, EF617506
57
Clone R21, EF617505
34
Clone R36, EF617518
Clone R23, EF617507
39
Clone R3, EF617488
26
Clone R15, EF617499
99
72
Clone R8, EF617493
Clone R19, EF617503
65
Clone R13, EF617497
95
Clone BB74, EF617548
53
Clone R1, EF617486
54
Clone BB73, EF617547
55
Clone BB39, EF617521
0.005

3.4. Conclusions

The South African mines provide a unique opportunity for direct exploration of the deep sub-surfaces. Our findings show that phages are present in these mines but in low counts as no viral like particles were observed from three mines with TEM analysis or EFM. Some phages with morphology corresponding to that of *Myoviridae* were obtained with concentrates from Beatrix mine, indicating the presence of this family in the mine sample. The results indicate that this technique requires high concentrations of viral particles making it inefficient for determining phage diversity from low sources.

The sequence-based approaches were therefore utilized and PCR was used to detect different families of phages from the mine water. Tailed phages amount to ~96% of the total phages; hence T4-like phages and T7-like podoviruses were detected. Amplification of the g23 gene from T4-like phages resulted in the product corresponding to one sub-family (ExoT-even) with all the mine samples. However, after sequencing few clones from Beatrix and Tau Tona LIC118 hit with the correct g23 product. The results indicate that the community of uncultured T4-type phages from the mines is different from the marine one. Similar results were obtained by Jia *et al.*, (2007) further confirming that the T4-type phages are highly diverse. In addition distinct groups of phages dominate different environments. Hence, the current set of primers might not be efficient for the detection of these phages from other environments, suggesting the need for new sets of primers that can cover all the diversity of this group of phages. This set of primers is however still effective in other environments, examples include the Loch Logan pond in South Africa (Chapter 2), soil (Wang *et al.*, 2009a) and rice field (Wang *et al.*, 2009b), indicating the detection is based on the viral community of different environments.

The analysis of the DNA polymerase gene from uncultured T7- like podophages produced different results to the T4-like phages. The DNA polymerase fragment was detected with all South African mines, and sequencing of the clones showed that the fragment is highly conserved. A neighbor joining tree constructed with the sequences indicated that our clones had high identity to the available T7-like sequences, which originated from different sources, ranging from marine, freshwater, sediments, terrestrial extreme and metazoan environments. These sequence similarities confirm the movement of phages between biomes as it was previously proposed by Sano *et al.*, (2004). In contrast to the T4-like phages which are highly diverse the T7-like podoviruses seems to be evolving slowly.

The overall results show that tailed phages are the most abundant as other forms of phages could not be detected with microscopy. The results further confirm the hypothesis that phages

are everywhere even in the mines. Furthermore the results demonstrate that though PCR approaches are useful for the detection of uncultured phages from the environmental samples, these methods can only be used for preliminary purposes and not for a complete diversity study of phage communities from the environment. The study clearly illustrates that not all families can be detected with this strategy. Furthermore; continuous evolution of phages makes it difficult with some viral assembledges as the conserved regions are diverse that they cannot be detected by PCR. This is demonstrated by the results obtained with the T4-phages. Therefore the use of random unbiased approaches is more efficient when studying diversity of phages from the environment. Hence, the next chapter uses a shotgun sequence based methods to study pages communities from the mine samples, by sequencing a viral metagenome from the mines.

Chapter 4

Sequencing of viral communities from South African deep gold mines

Summary

Shotgun libraries and pyrosequencing are now the two approaches that are used to asses the diversity of microbial communities from different environments. In this study shotgun sequencing revealed the untapped viral communities from four deep South African mines. Majority of the clones from the four mines shared no similarity to the known proteins. Pyrosequencing and annotation of Beatrix mine showed that more than 75% of the proteins had no similarity to the known proteins indicating that the viral diversity is not known. About 40% of the proteins assigned to a specific function from this mine were from phage origin. These proteins included DNA and RNA replication, structural and modifying enzymes mainly from three dsDNA tailed phage families. Most protein hits were from *Enterobacteria* phages and different *Bacillus* species. Proteins *Acanthamoeba polyphaga mimivirus* were also observed. Seven prophage regions size bigger than 5 kb were identified using prophage finder program, and these regions contained the high number of hypothetical proteins During ORF correction a lot of the ORFs were overlapping which is a common phenomenon with viral proteins. In addition novel viral proteins with biotechnological functions were identified.

4.1. Introduction

Whole-genome shotgun sequencing has been used to characterize a number of microbial community projects (Venter *et al.*, 2004; Rusch *et al.*, 2007), and the technique has managed to identify new and novel proteins from different environments, also resulting in near-complete genomes with some of the environments (Tyson *et al.*, 2004). To date genetic diversity studies of the entire viral assemblages are possible using sequence analysis of shotgun libraries constructed from the total viral DNA. Viral libraries have been constructed from different environments, including marine (Breitbart *et al.*, 2002); human fecal samples (Breitbart *et al.*, 2003) and human infant gut (Breitbart *et al.*, 2008). A single-stranded DNA virus was discovered from a library constructed from the fibropapilloma of a Florida green sea turtle (Ng *et al.*, 2009). Recent advances on the technique include analysis of RNA viral communities by construction of cDNA libraries (Zhang *et al.*, 2006). A diverse assemblage of RNA viruses, including a broad group of marine picorna-like viruses, and distant relatives of viruses infecting arthropods and higher plants were identified from reverse transcribed whole-genome shotgun sequencing (Culley *et al.*, 2006).

Sequencing of the libraries is usually done using the conventional Sanger sequencing technology (Sanger, 1977) with the primers that are available for the vector used. However, novel high throughput sequencing technologies have emerged, and they include Solexa (Bennett, 2004) and 454 pyrosequencing technology (Ahmadian, *et al.*, 2006; Baback *et al.*, 2007). Both methods do not require cloning and the latter can now produce reads up 400 bp with the new Titanium series (Roche Applied Science, 2008). Solexa sequencing produces shorter reads ~25 bp hence the technique is mainly used for closing the gaps in genome sequencing projects or resequencing and not for initial sequencing of genomes and metagenomics where novel proteins need to be identified (Warren *et al.*, 2008).

Pyrosequencing is therefore the most used sequencing technology for various genomes (Iacono *et al.*, 2008) as well as for environmental genome sequencing projects (Edwards *et al.*, 2006). The technology is based on sequencing-by-synthesis principle (Ronaghi, 2001) and employs a series of enzymes to detect incorporation of each nucleotide by measuring the release of inorganic pyrophosphate. To date the technique has advanced and it takes advantage of DNA capture beads that can contain on average one single-stranded template. Fragmented DNA is attached to the beads by adapters which are also used for amplification of the template into millions of copies in an oil emulsion PCR (emPCR). The beads are then distributed on a solid-phase sequencing substrate (a PicoTiterPlate™) with more than million wells. The wells contain the bead and the following additional reagents, the polymerase, luciferase, and ATP sulfurylase

(Margulies *et al.*, 2005). Each fragment is then sequenced in its own well and microfluidics cycles each of the four nucleotide triphosphates over the PicoTiterPlate™. The DNA polymerase catalyzes incorporation of complementary dNTP into the template strand. The nucleotide incorporation is followed by release of inorganic pyrophosphate (PPi) in a quantity proportional to the amount of incorporated nucleotide. ATP sulfurylase quantitatively converts PPi to ATP in the presence of adenosine 5' phosphosulfate (APS). The generated ATP drives the luciferase-mediated conversion of luciferin to oxyluciferin, producing visible light in amounts that are proportional to the amount of ATP. Generated light is captured by a charge-coupled device camera and converted into a digital signal (Huse *et al.*, 2007).

Various viral communities have been characterized using pyrosequecing leading to identification of novel viral assemblages. Cyanophages and a clade of newly discovered single-stranded DNA phages dominated the Sargasso Sea sample (Angly *et al.*, 2006). Different unculturable phage communities have been characterized using this technique. Furthermore, the sequencing project on Global Ocean Sampling (GOS) contained higher proportion of viral sequences when using pyrosequencing, indicating that to date phage sampling is still poor (Yooseph *et al.*, 2008). Pyrosequencing analysis showed that most of the genes identified are novel, as up to 75% of the proteins showed no similarity to the known sequences. In addition environmental viral genome sequencing studies also result in high numbers of proteins that do not show any detectable homology to existing ones.

In chapter 3 the presence or diversity of some known phage groups was determined. The main objective of this study is however to investigate the viral communities from the South African deep mines, the methods used in chapter 2 and 3 only identifies few families of phages. In addition, bacteriophages lacking conserved loci could be missed. Hence, for this chapter random sequencing of the viral communities was performed. Libraries were constructed with water from a deep mine and a number of clones sequenced and pyrosequencing was also done on phage DNA extracted from a biofilm collected in Beatrix mine. Novel genes have been identified using this technique.

4.2. Materials and Methods

4.2.1. PCR parameters and sequencing

Standard PCR reactions and sequencing were done as described in sections 2.2.2.2. and 2.2.2.3. Oligonucleotides primers were designed with reference to the relevant literature as indicated on table 4.1. Other primers were designed for sequencing the libraries as explained in section 4.2.3.

Table 4.1: Oligonucleotide primers used

Primer name	Sequence 5' end to 3' end	References
27F	AGA GTT TGA TCM TGG CTC AG	Lane, 1991
1492R	GGT TAC CTT GTT ACG ACT T	Lane, 1991
20bF	YTC CSG TTG ATC CYG CSR GA	Rincon *et al.*, 2006
1040R	GGC GAT GCA CCW CCT CTC	Reysenbach and Pace, 1995
BM11F	GTGCGGCAGGTACGGCTTCTCCG	
MM1F	GCCAATCAGCGGATTCATGCC	
MM3F	AAGTGGCCTATAACTATATC	
MM4F	GCGATACTGGATGGCGCAGAGC	
SD7F	GGTTGAGCATGATCCGAATGAG	
SD9F	ACAGCCACTTATCGGCAGCG	
BM12F	TGACCGGCGTGACCTGTCTGC	
BM14F	CTGCCGGTGATCTAGTTTGTCC	
TT19F	TGCCGTCTGGCAGCTTGGCCGAC	
TT20F	TGCCAATGTACCGGATATTGGAG	
HECTORPol29F	GCA AGC AAC TTT ACT GTG G	Breitbart *et al.*, 2004
HECTORPol500R	GAA TGA TCT ACA CTC TTT GCC ATA CGG TG	
MZIA1bis	GAT ATT TGI GGI GTT CAG CCI ATG A	Filée *et al.*, 2005
MZIA6	CGC GGT TGA TTT CCA GCA TGA TTT C	Filée *et al.*, 2005

Oligos without references were designed as part of this study

4.2.2. Library construction

Shotgun libraries were constructed using samples from the following mines; Tau Tona, Masimong, Beatrix and Star diamonds. Details on sampling and description of the samples are

provided in chapter 3, section 3.2.1 and 3.3.1, respectively. Viral-like particles and DNA were isolated from fissure water as described in chapter 3 (section 3.2.2 and 3.2.5). Due to low recovery of viral genomic DNA from the environment use of whole genome amplification was necessary to increase the concentration of viral DNA isolated from different mines. The Illustra genomiphi V2 DNA amplification kit was used following manufacturer's instructions (GE Health Care). The method uses Phi29 DNA polymerase and random hexamers, and amplifies genomic DNA through strand displacement. Extracted viral DNA from the four mines (20 ng from each mine) was amplified at 30 °C for 18 hours before it could be used for the library construction. The amplified DNA was digested with *EcoRV* and the fragments between 1 and 3 kb were excised from the gel and purified using the Bioflux, Biospin Gel Extraction kit (Separation scientific). Blunt-end, fragmented DNA was then ligated into linearized pcrSMART-HC Kan (Lucigen Corporation) vector following manufacturer's instructions. Recombinant clones were selected on kanamycin (50 µg/ml).

4.2.3. Library screening

Plasmid DNA was isolated using the Bioflux DNA/RNA extraction/purification for Biospin plasmid DNA extraction kit (Separation Scientific). Clones containing inserts were selected after restriction digestion. The library was screened using the sequence-based approach as viruses express toxic proteins that might be lethal to the host. The clones were therefore sent for sequencing at Inqaba Biotechnologies. A combined number of about 80 clones in pSMART from different mines were sequenced from both directions using SL1 and SR2 primers on the pSmart vector. Internal primers were designed using the partial sequences obtained with either SL1 or SR2 when necessary for large fragments in order to get the complete sequence. The primer name (Table 4.1) corresponds to the clone that required internal sequencing. Similarity searches were done using BlastX (Altschul *et al.*, 1990) with the cutoff e-value of 0.01 to identify the proteins.

4.2.4. New sampling

Biofilm and water samples were collected from Beatrix mine on the 19th of September 2008. Biofilm samples were collected using sterile 250 ml centrifuge tubes, and a portion from each of the samples was fixed on site with 3% gluteraldehyde (filtered through 0.02 µl) in the 15 ml sterile falcon tubes for TEM studies. Four different biofilm samples were collected; labeled

Snow White, Site 1, Site 2 and Black beauty; depending on the appearance or the site they were collected from. Upon arrival 100 ml PBS (pH 7.4, 0.02 µm filtered) containing 0.5% w/v and 1M NaCl was added to 20 g of the biofilm and viral-like particles were isolated using the soil method as described in section 2.2.4. The mixture was sonicated (Bandelin sonopuls, 50% power) at 4°C for 5 min (with each minute interrupted by manual shaking for 30s). The debris was removed by centrifugation at 11 000 X g for 10 min (Beckman J2-MC, JA 14 rotor) at 4°C. The viral-like particles were precipitated with 10% w/v PEG 6000 by incubating at 4°C for >12 hrs. Phages were then pelleted by centrifugation at 13 000 X g for 30 min (Beckman J2-MC using JA 14 rotor). The pellet was dried and resuspended in minimal volume of Milli Q water that was filtered through a 0.02 µm membrane filter.

The water sample was concentrated with TFF system as described in section 3.2.2 and concentrates from both biofilm and fissure water were used for TEM and viral DNA isolation (section 3.3.5).

4.2.5. Check points before pyrosequencing

Prior to pyrosequencing the following preliminary experiments were done to detect phages and also as a way of selecting the best sample for sequencing. The presence of viral-like particles from each sample was investigated using electron microscopy and PCR. Aliquots from concentrated viral suspensions were used for TEM analysis. Tailed phages, T4-like and T7-like, were detected by PCR amplification. The above techniques were performed as described in the previous chapters.

4.2.6. Sample selection for pyrosequencing

Further analysis was done on the two candidate pyrosequencing samples; Snow White and Black Beauty. Total genomic DNA was isolated from the two samples using Fast DNA spin for soil kit (MP Biomedicals). Bacterial 16 rDNA was amplified using 27F and 1492R primers (Table 4.1). The following PCR parameters were used; initial denaturation for 2 min at 94°C, and 30 cycles of the following steps, starting with denaturation at 94°C for 30 sec, annealing was done at 55°C for 45 sec and elongation at 72°C for 90 sec. The final elongation was done at 72°C for 10 minutes. The sequences were identified using BlastN (Altschul *et al.*, 1990).

4.2.7. Pyrosequencing

The presence of prokaryotes (bacteria and archaea) was detected in the samples prior whole genome amplification. Bacterial 16s rDNA was amplified as described on section 4.2.4 and similar parameters were used for archaea with annealing at 60°C and elongation at 72°C for 1 minute. Viral DNA was amplified to increase the concentration prior to pyrosequencing. Whole viral genome amplification was performed using Illustra genomiphi V2 DNA amplification kit (GE Health care) at 30°C for 18 hours. Amplified viral DNA from black beauty (20 ng) was used as the template. Two quarter runs of the 454 pyrosequencing with the GSFLX were done at Inqaba Biotechnologies.

To determine if there was a significant presence prokaryotic and eukaryotic domains after amplification, the reads of one quarter plate (96000 reads) were subjected to blastn analysis. The output file was used as input for MEGAN (Huson *et al*, 2007) which was used to compute the taxonomical content of the reads. MEGAN uses a lowest common ancestor algorithm to assign reads to a taxon so that the taxonomical level of the assigned taxon reflects the level of conservation of the sequence.

4.2.8. Assembly and finishing

Raw sequence data was assembled using Roche Newbler software (Version 2.0.00). The assembled data was converted to Gap4 format using the conversion pipeline of Bernd Senf (http://genome.imb-jena.de/software/roche454ace2caf/) before editing and joining of contigs using Gap4.

4.2.9. Automatic annotation

All contigs bigger than 500 bp were submitted to the JCVI Annotation Service for automatic annotation. The contigs were first joined by inserting the following sequence in between; NNNNNCACACACTTAATTAATTAAGTGTGTGNNNNN creating a pseudo genome. The sequence creates stop codons in all six reading frames and the genome was then submitted for annotation. The annotation pipeline entails the following, gene finding with Glimmer, Blast-extend-repraze (BER) searches, HMM searches, TMHMM searches, SignalP predictions, and automatic annotations from AutoAnnotate. The first step involves the use of Glimmer system (Delcher *et al.*, 1999) to identify genes in bacterial, archaeal, or viral genomes.

Once the candidate genes have been identified, each protein is searched against an internal non-identical amino acid database (niaa) using BLAST-Extend-Repraze (BER). The database is made up of all proteins available from GenBank (http://www.ncbi.nlm.nih.gov), PIR (http://pir.georgetown.edu), SWISS-PROT (http://www.expasy.ch/sprot) and JCVI's CMR database, the Omniome (http://www.jcvi.org/cms/research/projects/cmr). First a BLAST search (Altschul, *et al.*, 1990) is performed and all significant matches are stored. Then a modified Smith-Waterman alignment (Smith and Waterman, 1981) is performed to identify potential frame shifts or point mutations in the sequence, this done by extending the gene 300 nucleotides upstream and downstream of the predicted coding region.

The proteins are also searched against hidden Markov models (HMMs) using the HMMER package (Eddy, 1998) which consists of the Pfam HMMs (Bateman, *et al.*, 2000), and TIGRFAMs (Haft, *et al.*, 2001, 2003). The manual annotation tool Manatee was also downloaded from SourceForge (manatee.sourceforge.net) and used to manually review the output from the prokaryotic pipeline of the JCVI Annotation Service.

4.2.10. Correction of the ORFs using Artemis

Artemis is a sequence viewer and annotation tool that allows editing of annotated genomes (Carver *et al.*, 2008). Automatic annotated results from TIGR were loaded into Artemis for manual checking and correction of open reading frames (ORFs). The ORFs were checked for the correct start and the stop codons. Blast searches were also done in some cases to confirm the start codons of genes of interest.

4.2.11. Evidence of phage proteins or genomes

To identify phage proteins, annotated data was viewed with Manetee which is a genome editing and manual annotation package from JCVI (manatee.sourceforge.net). All the ORFs in different role categories were manually checked. The origin of each protein was determined by viewing the BER results. All the proteins from viral origin were counted and the Blast results of proteins with assigned function were only used for this analysis. Conserved hypothetical proteins were excluded.

The presence of probable phage genomes or phage protein clusters were also determined using Prophage Finder (Lima-Mendez *et al.*, 2008). The program is written in Perl and uses the

ACLAME database (Leplae *et al.*, 2004) as source of phage data for similarity searches. Phage-like coding sequences are identified by gapped BLASTP search (Altschul *et al.*, 1997) of all the translated CDSs from the input genome against all phage proteins in ACLAME. All contigs bigger than 2 000 bp were used in this analysis and submitted online (http://bioinformatics.uwp.edu/~phage/ProphageFinder.php).

4.3. Results and discussions

4.3.1. Library screening

Amplified viral DNA (Figure 4.1) from the four mines was partially digested and used in library construction. The constructed viral library was screened by directly sequencing recombinant clones containing inserts (Figure 4.2). The average insert size was 2000 bp and full length inserts were sequenced. Similarity searches done on the sequences revealed that only two known proteins were of viral origin viz. phage SPO1 DNA polymerase-related protein from Beatrix mine clone 3 and site-specific tyrosine recombinase / phage integrase family protein from Masimong clone 9. The majority of the clones were hypothetical proteins and other clones (indicated unknown) showed no similarity to any available protein sequences (Table 4.2). The results indicate that the diversity of viral communities from the South African mines is highly uncharacterized which is characteristic of phage metagenomes. Random sequencing studies performed on different viral communities to date show a similar pattern of results where more than 65% of the proteins are unknown (Breitbart *et al.*, 2004; 2008). Other sequenced fragments showed similarity to bacterial genes (such as RNA polymerases, histidine kinases, intergrases etc.) when compared against Gene Bank non-redundant database (Table 4.2). These proteins could have been introduced by horizontal gene transfer (Comeau et al., 2007; Williamson *et al.*, 2008). The results show that the shotgun libraries are useful for the analysis and identification of the new and unknown viral proteins, however sequencing of high number of clones is required for this technology to be efficient in identifying these novel proteins. At this point the Sanger sequencing of the clones was discontinued and pyrosequencing costs were now competitive in terms of the information obtained.

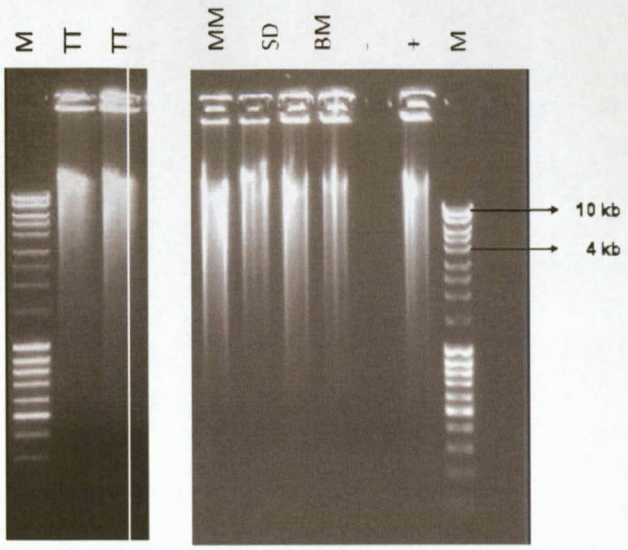


Figure 4.1: Whole genome amplification of viral DNA from water sampled from selected mines, the DNA ladder is represented with lane M. Lanes TT, MM, SD and BM are Tau Tona, Masimong, Star diamonds and Beatrix mine, respectively. The negative and positive controls are on lanes (-) and (+), respectively.

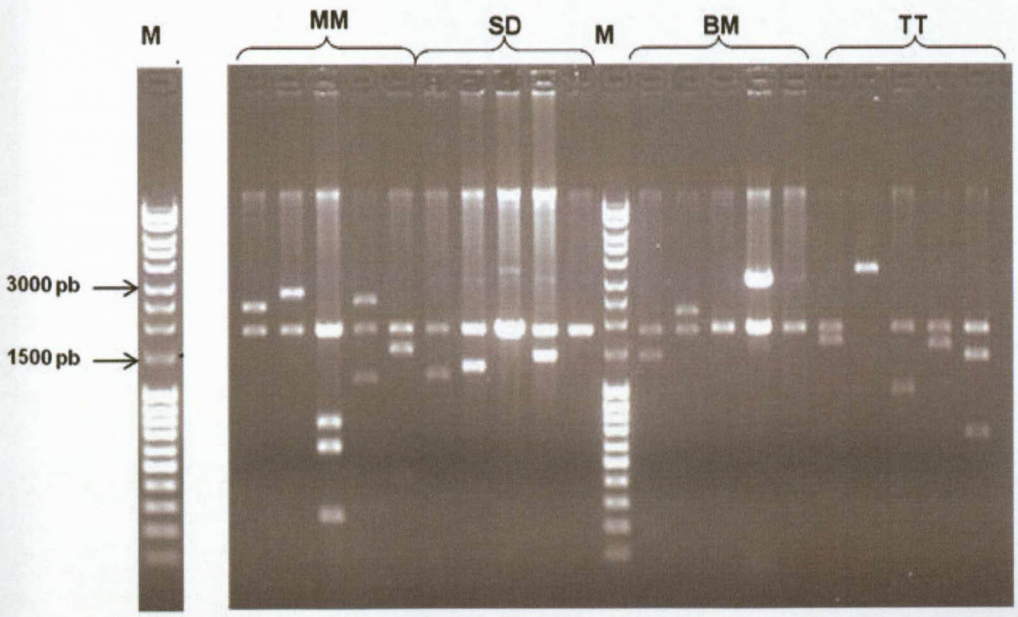


Figure 4.2: *EcoR*I restriction digests of pcrSMART clones. Fermentas Mass Ruler mix is indicated on with lanes M. Recombinant clones from different mines are on lanes MM, SD, BM and TT; the lanes also correspond to the mine where clones were obtained.

Table 4.2: Proteins obtained with the library BlastX results; annotation based on the GeneBank

	Site			
	Number of hits from each mine			
Protein name	**MM**	**SD**	**BM**	**TT100**
Hypothetical protein	9	14	8	13
Resolvase/intergrases	3	1		
Proteases			1	
Nucleases			1	
DNA polymerases			1	
RNA polymerases		1	1	
Amino acid synthesis proteins	1		1	
Membrane proteins	2			
Antimicrobial proteins		1		
Kinases	2	1		2
Phosphatase	1			
Unknown	2		6	5
Multidrug resistance proteins		1		
Transcriptional regulators		1		

4.3.2. New sampling

Biofilm samples (Figure 4.3) were collected from Beatrix mine, level 26 (1.176 km below the surface) to identify the viral community structure from this mine. Basic information about the site is supplied in chapter 3 (section 3.3.1). The water sample (22 L) was concentrated to 50 ml using the TFF system as previously described (section 3.2.2).

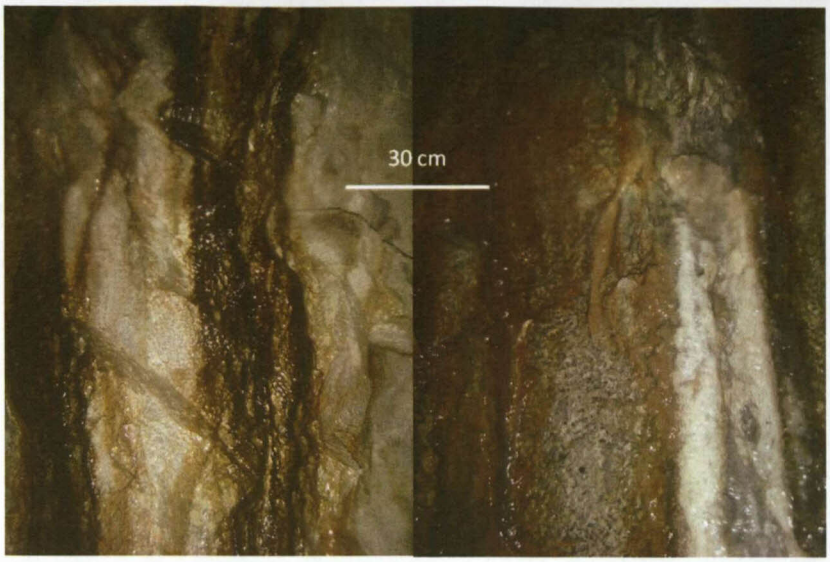


Figure 4.3: Site at Beatrix level 26 where biofilm samples were collected.

4.3.3. TEM

Viral-like particles were obtained with concentrates from the filter and biofilm samples. The particles obtained with viral suspensions from the filters appeared to be tailed, however the micrographs obtained were not of good quality and the viral counts were extremely low so it was difficult to distinguish if observed particles were not artifacts of the TEM (Figure 4.4). Phage particles were not observed with water concentrates, hence further analysis of both water and filter samples was discontinued.

Phage particles obtained with biofilm samples were tailed with icosahedral heads (Figure 4.5). Particles resembled *Siphoviridae* morphology with their long non-contractile tails and an average length of 200 nm. Some of the phage particles had morphology resembling that of *Myovividae* with short tails and base plates. All these particles were obtained with Black beauty biofilm viral concentrates, indicating the abundance of tailed phages from this sample.

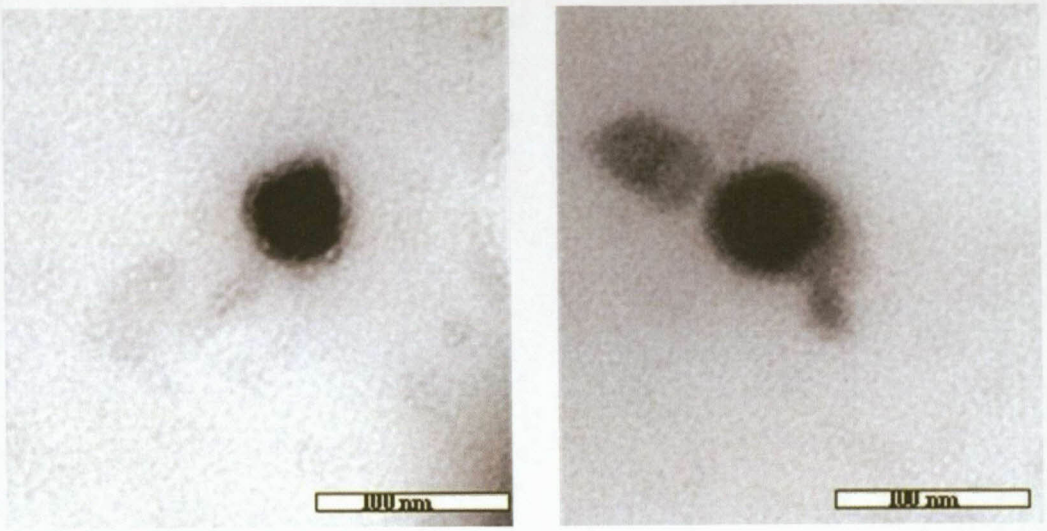


Figure 4.4: Viral-like particles obtained with filter samples.

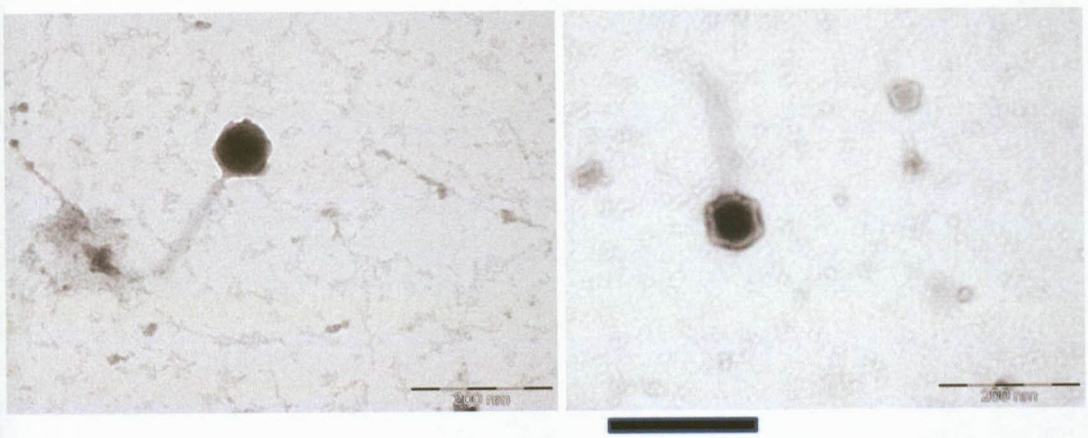


Figure 4.5: TEM micrograph of phage particles obtained with Black Beauty biofilm sample. Scale bar is 200 nm.

4.3.4. Biofilm sample processing

PCR products were obtained for T4-like phages (Figure 4.6 A) and T7-like podoviruses (Figure 4.6 B) indicating the presence of these families in Beatrix biofilm samples. The fragments were however not sequenced.

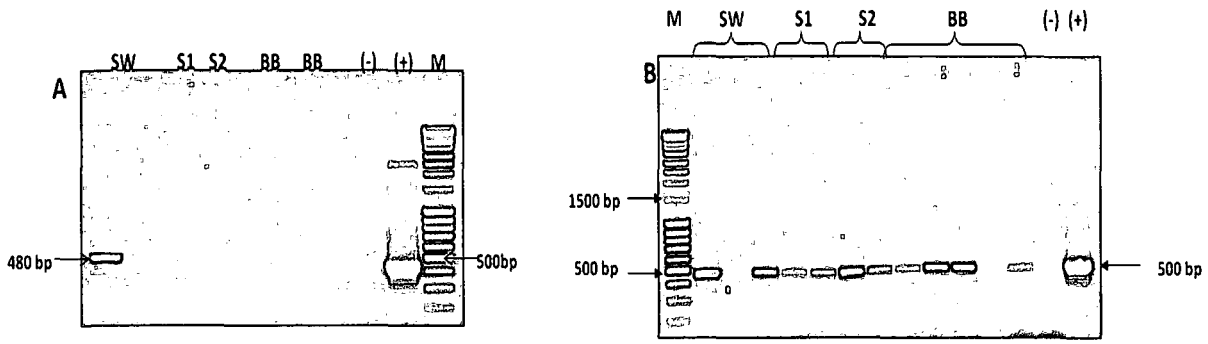


Figure 4.6: PCR detection of T4-like, A and T7-like phages, B. Fermentas Mass ruler is on lanes M; negative and positive controls lanes (-) and (+), respectively. Different biofilm samples are represented on lanes SW, S1, S2 and BB.

4.3.5. Sample Selection for pyrosequencing

Higher viral DNA yields were obtained with biofilm samples from Snow White and Black Beauty, therefore these samples were selected for further analysis. The 16S rDNA PCR (Figure 4.7) was performed as a way of selecting the best sample for pyrosequencing. The best sample in this project was regarded as the one with highest microbial diversity. After sequencing of 10 clones from each sample, the clones obtained with Snow White showed similarity to *Thiofaba tepidiphila*, whereas clones from Black Beauty hit with 16S rDNA sequences from different uncultured bacterial species. These results indicated that Black Beauty had higher novel diversity than snow white. In addition, the unculturable clones detected in Black Beauty, held promise that novel phages would likely be found in this biofilm. PCR detection indicated the presence of the T4-like phages and T7-like podoviruses and when doing TEM, phage particles were obtained with concentrates from this sample further confirming the abundance of phages from the respective sample. Black Beauty was therefore selected as the sample for sequencing.

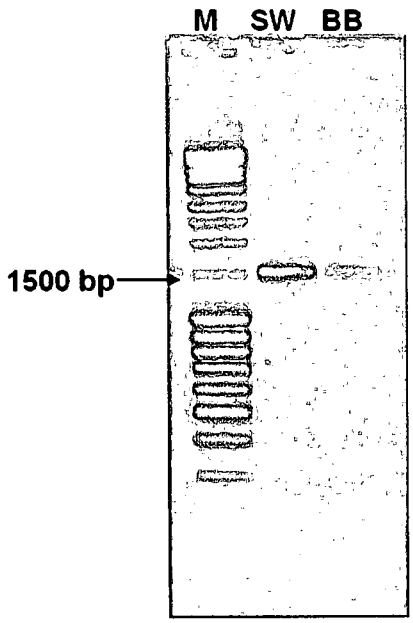
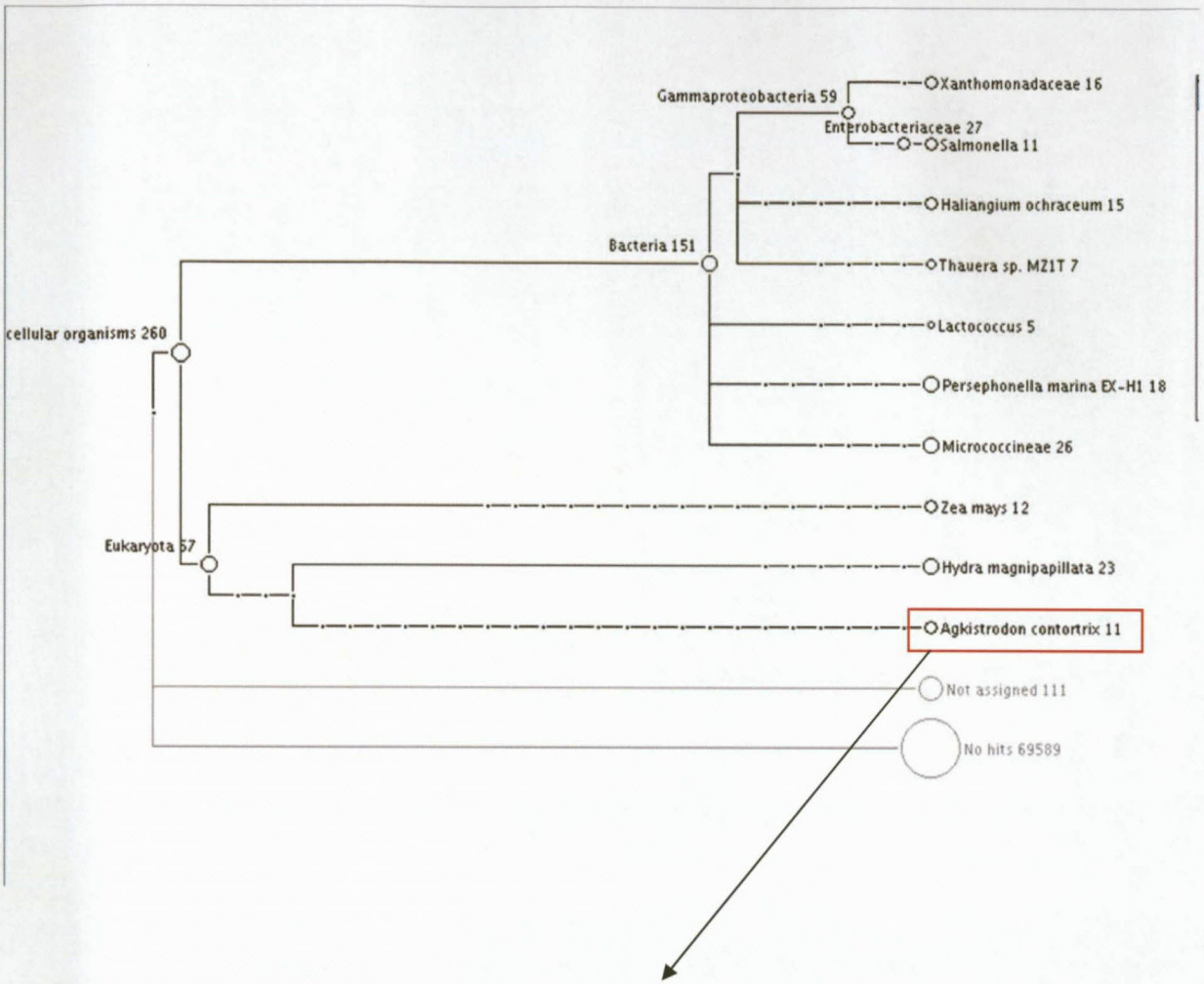


Figure 4.7: 16S rDNA amplification, lane M is the DNA ladder used; lanes SW and BB are products obtained with biofilm genomic DNA snow white and black beauty, respectively.

4.3.6. Pyrosequencing

Approximately 1ug of DNA was obtained with whole genome amplification of Black Beauty genomic DNA. Archaeal 16S rDNA PCR resulted in no products indicating the absence of this domain in the amplified sample. An amplification product was however obtained with bacterial 16S rDNA. In addition MEGAN analysis of 96 000 reads also showed only 0.53% of the reads were from both prokaryotes and eukaryotes (Figure 4.8) indicating that a phage metagenome has been susscessfully sequenced. Furthermore only short fragments were showing similarity to these prokaryotic and eukaryotic sequences.



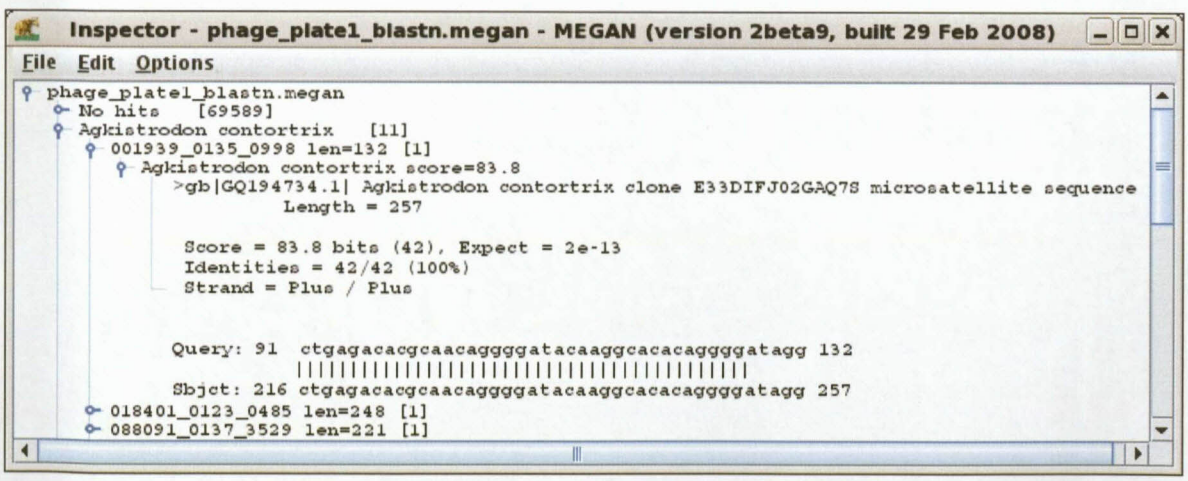


Figure 4.8: Megan analysis of the quarter plate of phage pyrosequencing data, the arrow points towards the window showing the blast hit of the sequence in red rectangle.

The relative size distribution of fragments generated with nebulization was between 300 and 900 bp when it was run on an Agilent BioAnalyzer DNA 1000 LabChip (Figure 4.9). The single stranded library ranged between 300 and 800 bp when analyzed on the Agilent RNA Pico 6000 LabChip, and the yield was low (Figure 4.9), however enough for a pyrosequencing run. The two quarter runs resulted in 43.7 mb of sequence data.

The results of the Newbler assembly are given in Table 4.3 and more than 94% of the bases called had a score of 40 and above indicating the pyrosequencing run was successful. After assembly a total of 560 contigs were obtained. A high number of singletons were obtained illustrating this point and suggesting that the viral diversity from Beatrix mine is high. Furthermore the two quarter GS FLX runs performed were not enough to cover the diversity from this metagenome. The contigs bigger than 500 bp were surprisingly low in number considering that this was a metagenome sample. This could possibly reflect the successful separation of phages from bacteria and bacterial DNA in the sample preparation phase. The largest contig was 141813 bp which is the size range of the T4-like phages (Filée *et al.*, 2006).

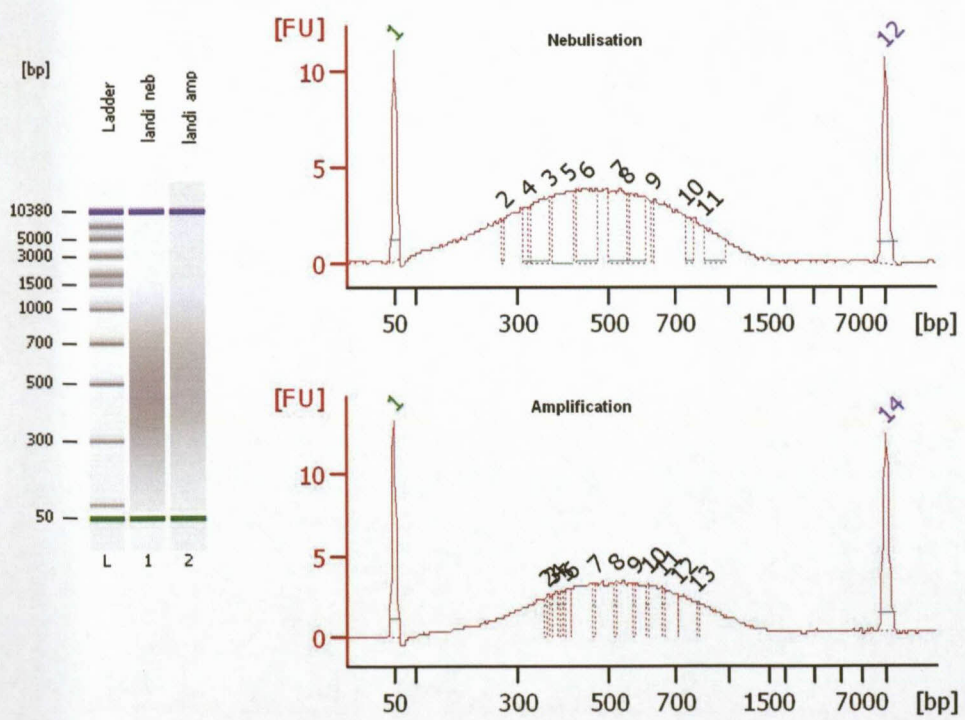


Figure 4.9: Size distribution of the double stranded DNA fragments after nebulisation and the single stranded library before the sequencing run.

Table 4.3: Pyrosequencing run results

Total Number Of Reads	179040
Total Number Of Bases	43751280
ReadStatus	
Number assembled	121456
Number partial	9487
Number singleton	6537
Number repeat	711
Number duplicate	38003
Number outlier	2846
LargeContigMetrics	
Number of contigs	560
Number of bases	1410427
Average contig Size	2518
N50 contig sze	7343
Largest contig size	141813
Q40 plus bases	1329821, 94.28%
Q39 minus bases	80606, 5.72%

4.3.7. Finishing

The Newbler assembled pyrosequencing data was edited, base calling was checked and the sequence was polished as far as possible using Gap4 of the STADEN package (Bonfield *et al.*, 1995). Because Newbler was found to be conservative in the construction of contigs, in our lab, additional joining of contigs was done where possible using the contig comparator of Gap4, which compares all contigs against each other to find possible overlaps which could be sufficient to join contigs. Potential candidates are indicated as dots on the comparator plot (Figure 4.10). After manual inspection using a cutoff of 5% mismatch for overlaps, 36 contigs were joined bringing down the total number of contings to 524 with the largest contig at 141813 bp. Being able to assemble such large contigs from phage metagenome data is exceptional, considering the relatively low number of bases sequenced. These could have been due to the whole genome amplification which could have introduced a bias during amplification, as the distribution of the reads was not even (Figure 4.11). The coverage on the large contigs was higher than that of the small contigs.

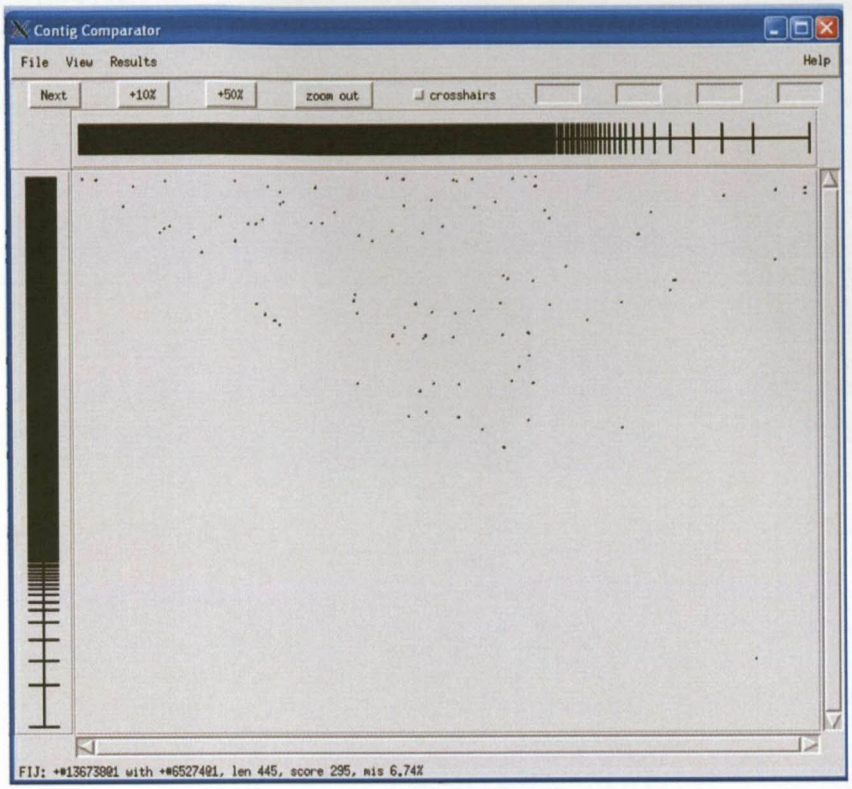


Figure 4.10: Contig comparator window showing all the contigs compared against each other, the bars on the horizontal and the vertical are contigs in ascending order of size. The dots represent the contigs that can possibly be joined.

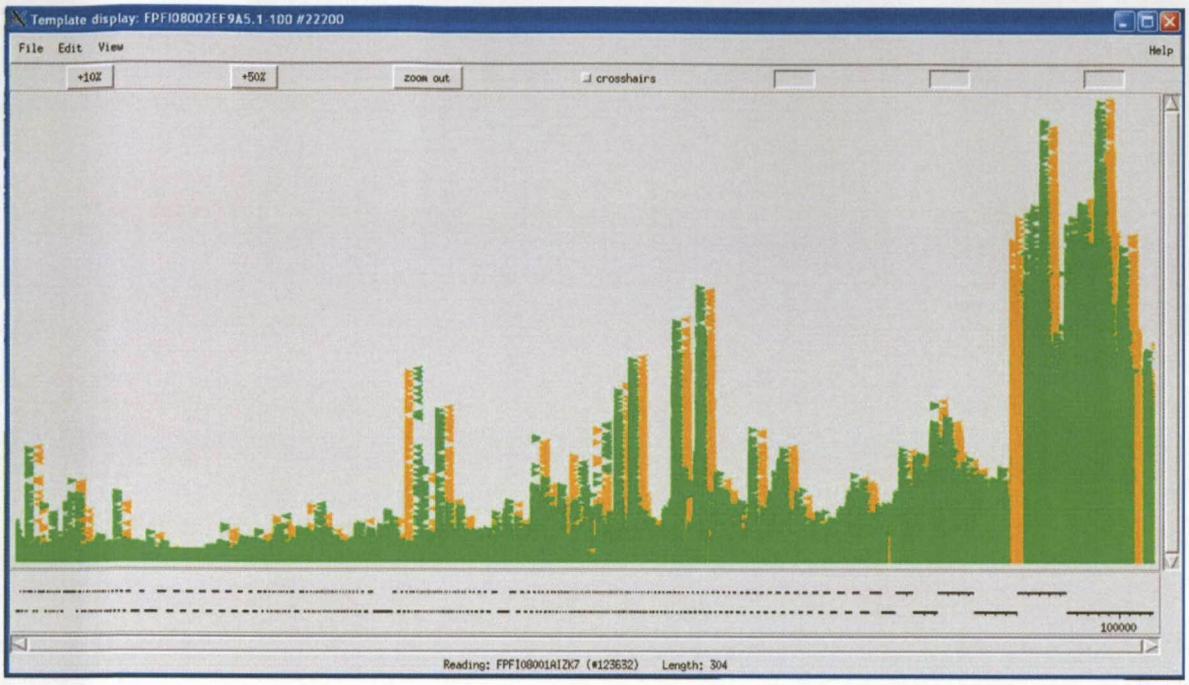


Figure 4.11: STADEN template display window showing the distribution of contigs in increasing size, yellow and green arrows are contigs in opposite directions.

4.3.8. Annotation

After annotation a total of 2913 ORF's were obtained; 75.7% of these were hypothetical proteins (Table 4.4), these proteins do not show similarity any other proteins in public databases. Conserved hypothetical proteins, which are proteins that showed similarity to other hypothetical proteins in public databases, contributed 6.87% and 3.1 % were proteins with unknown function. This is typical of phage metagenome sequencing projects, indicating that our understanding of the physiological role of phage proteins is lacking or that our algorithm for the prediction of genes in phages is inadequate. In addition the annotation engine is optimized for prokaryotes and not for viruses. To date very few viral communities have been analyzed using pyrosequencing, similar results were obtained with the analysis of four oceanic regions where more than 85% of the sequences were unique (Angly *et al.*, 2006). Shotgun sequencing of most viral communities shows that approximately 60% of the viral proteins are unique (Chen and Pachter, 2005), suggesting this approach underestimate the environmental viral diversity.

Other categories contributed differently as indicated (Table 4.4). Proteins that have been assigned to a function were further classified into different TIGR role categories (Table 4.5).

These categories describe the biological roles of proteins. Most phage proteins were classified as mobile and extracellular elements and this category mainly comprised of intergrase genes and resolvases which are sequences responsible for intergration and rearrangement. Again phage proteins contributed greatly as unclassified and this was expected as majority of phage ORFs have not been previously characterized. Other proteins (ORFs) were classified as unknown. The ORFs in this category have no characterized matches, but has an above-trusted cutoff match to a HMM domain of unknown function. Phage capsid proteins were classified as cellular envelope and most DNA modifying enzymes and replication were categorized as DNA metabolism.

During annotation tRNA cluster regions were also identified. In bacterial genomes tRNA regions are favorite sites for insertion of introns (Edgell *et al.*, 2000). A single cluster of the tRNA region of bacterial origin was observed, and the region was characterized by insertion sequences (Figure 4.12) which is a common phenomenon in bacterial genomes. To date the reason for the presence of tRNAs in phages is still not clear and the tRNAs were also present in the Beatrix viral metagenome. Studies have showed that homing endonucleases in phage genomes are located within a cluster of tRNAs (Akulenko *et al.*, 2004). Some of the T4-related phages contain the tRNA cluster and lack the endonuclease and during co-infection with the T4 phage the endonuclease ORF is therefore transferred to the tRNA region (Brok-Volchanskaya *et al.*, 2008). The HNH homing endonuclease was observed located in the cluster of the tRNA (Figure 4.13) and in some cases the endonuclease ORF was absent.

Table 4.4: Automatic annotation results

Total ORFs:	**2913**	**100 %**
assigned function	396	13.6 %
conserved hypothetical	200	6.87%
unknown function	89	3.1 %
unclassified, no assigned role category	41	1.4 %
hypothetical proteins	2204	75.7 %

Table 4.5: Classification of ORFs into different categories by TIGR automatic annotation

TIGR role categories	Total number of proteins	Number of phage Proteins
Unclassified	41	39
Amino acid biosynthes	4	0
Purines, pyrimidines, nucleosides an nucleotides synthesis	29	8
Fatty acid and phospholipid metabolism	7	0
Biosynthesis of cofactors, prosthetic groups and carriers	9	0
Central intermediary metabolism	8	0
Energy metabolism	20	0
Transport and binding proteins	15	1
DNA metabolism	118	35
Transcription	15	2
Protein synthesis	26	4
Protein fate	35	2
Regulatory functions	13	1
Signal transduction	3	0
Cellular envelope	43	16
Cellular processes	29	4
Mobile and extrachromosomal	75	72
Unknown functions	89	26

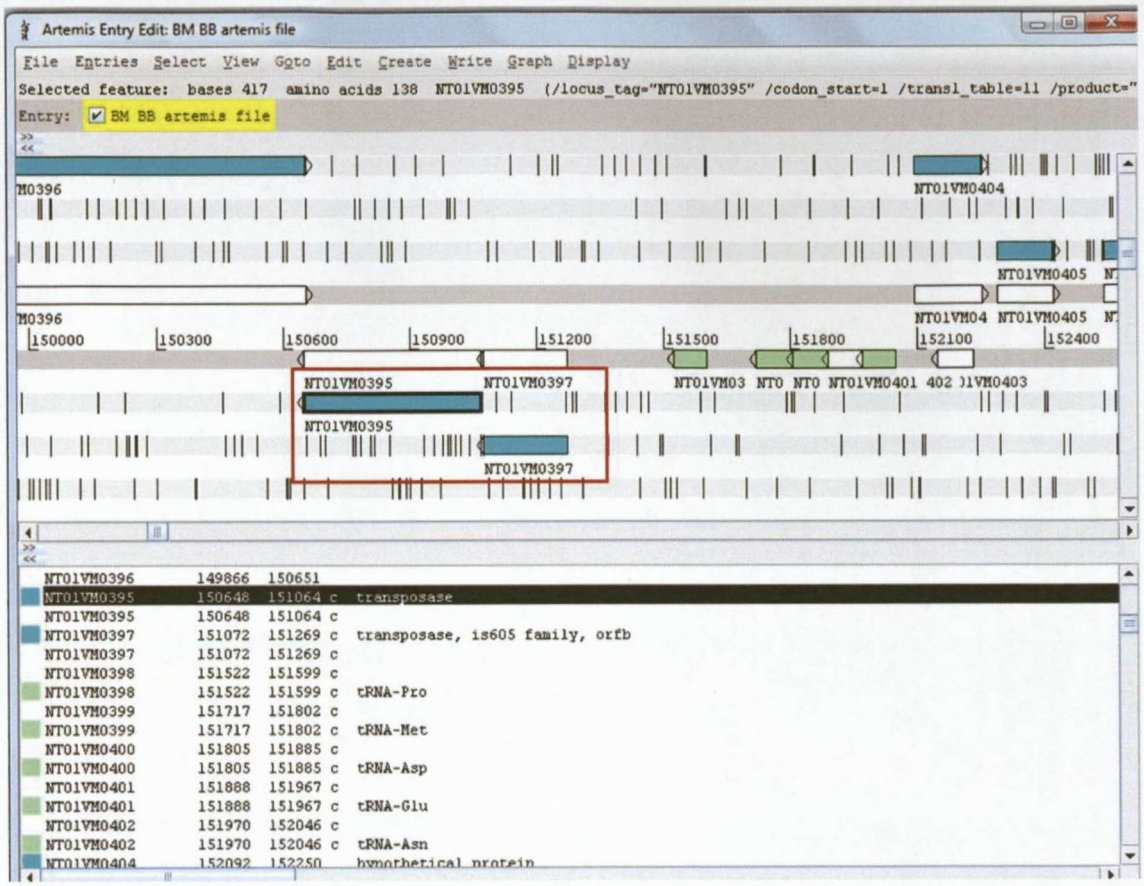


Figure 4.12: Artemis window showing typical bacterial tRNA cluster region, the green rectangles are the tRNAs and the ORFs in the red rectangle are insertion sequences (transposases).

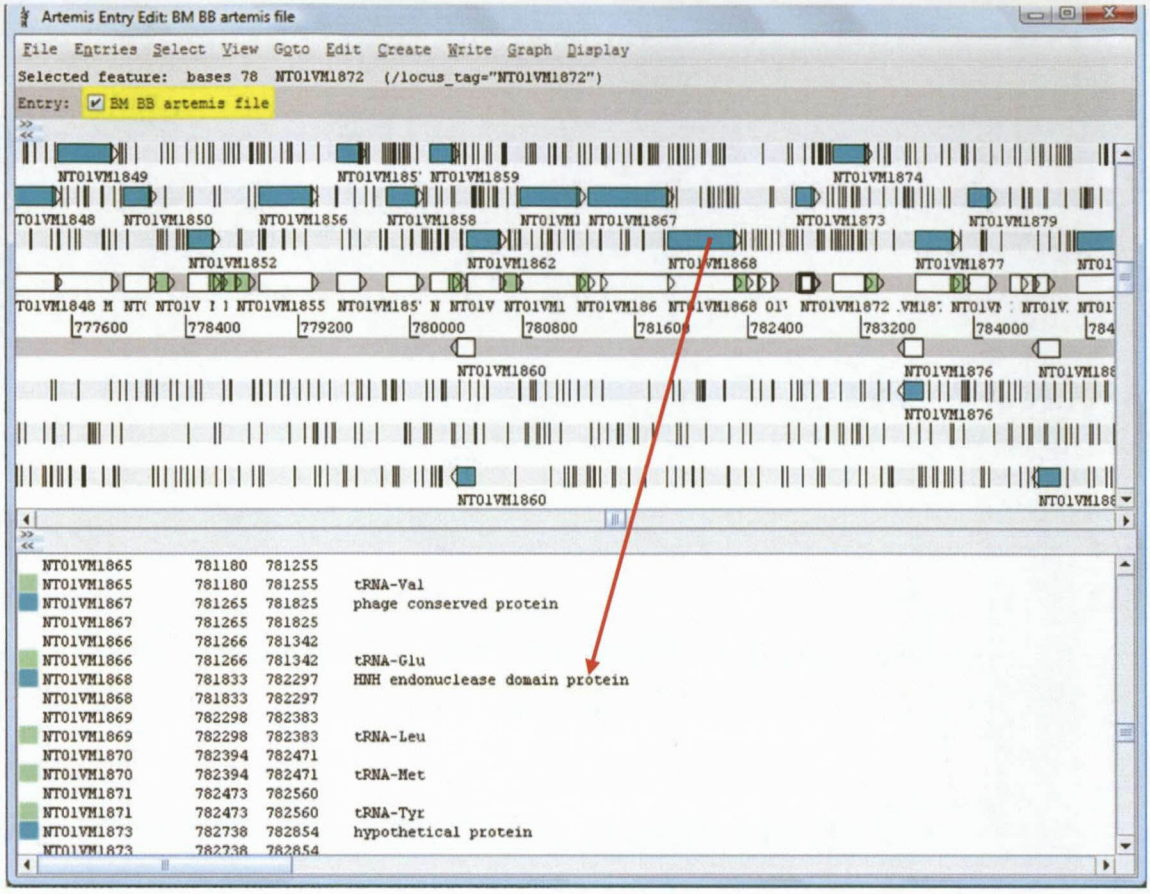


Figure 4.13: Artemis window demonstrating a tRNA cluster region in phages, all the green rectangles are tRNAs. The arrow indicates the HNH endonuclease ORF within the tRNAs.

4.3.9. Evidence of phages from biofilm

The Black beauty viral assemblage resulted in 75% proteins not showing identity to any sequences in public databases. ORFs assigned to a function were checked manually by viewing the BER results in Manatee and 40% of the proteins were from phage/viral origin (Figure 4.14). These excluded hypothetical proteins and conserved hypothetical proteins. The following categories contributed high numbers of phage proteins; unclassified proteins; DNA metabolism; mobile and extra chromosomal elements and proteins with unknown function. Phage structural proteins such as capsid and tail proteins were categorized as cellular envelope (Table 4.5). Despite the high number of ORFs that showed no similarity to the available data, phage proteins contributed greatly as unknown function and unclassified proteins. In addition

BLAST analysis of the conserved hypothetical proteins revealed that these proteins were showing similarity to phage proteins.

All the phage proteins and their respective ORFs are supplied on Appendix B. Majority of these proteins were showed similarity to proteins from the three families of dsDNA tailed phages, dominated by members of *Enterobacteria* phages infecting the following hosts; *E. coli; Burkholderia; Ralstonia; Pseudomonas* and *Vibrio cholerae.* Phages from different *Bacillus* species were also common. The results correspond to the results obtained (Dr A. J. Garcia-Moyano) when investigating the microbial population from this mine. Most of the bacterial population was dominated by *Beta* and *Gamma Proteobacteria* and few *Firmicutes*. Some of the proteins were also from *Acanthamoeba polyphaga mimivirus* which is the biggest eukaryotic virus, the reason for this was unknown, however sequencing studies done on phage genomes (Comeau *et al.*, 2007) and metagenomes (Breitbart *et al.*, 2002, Angly *et al.*, 2006) also revealed the presence of viral proteins from eukaryotes. It is very likely that eukarya could have colonised the biofilm as it was exposed to the air and eukarya have been detected in biofilm as well as water from the deep mines (Personal communication: A.J. Garcia-Moyano)

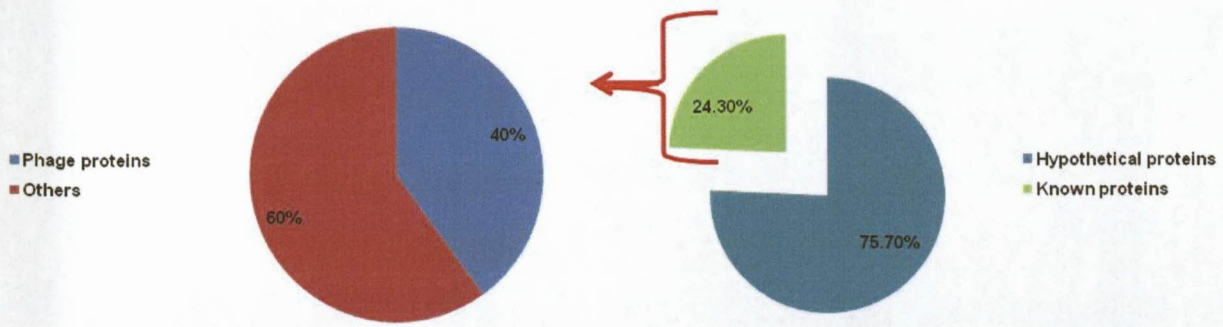


Figure 4.14: Phage protein analysis indicating the percentage of phage proteins to identified ORFs.

4.3.10. Evidence of phage genomes

A total of 12 prophage regions were identified using Prophage finder, which is the prophage prediction method based on the detection of genomic segments statistically enriched

in phage-like genes. The method is mainly used to detect complete prophage regions in prokaryotes. Five of the predicted phage regions were smaller than 4 kb, which is the minimum genome size of a ds DNA phage. Though RNA viruses are smaller, they were not expected because cDNA was not synthesized prior to the whole genome amplification. The seven prophage regions bigger than 5000 bp were located within the largest contigs. Predicted phage regions contained a high number of hypothetical proteins indicating that the phage diversity is not known (Table 4.6). The results explain the high number of hypothetical proteins obtained with automatic annotation. In addition these findings prove that phage genomes comprise of a high number of novel proteins and of unknown function. Phage proteins comprised within the identified prophages included integrases and phage assembly proteins such as terminase, portal protein, coat protein and tail tape measure protein (Appendix A). The presence of these proteins confirms the accuracy and reliability of this programme in identifying the phage genome or prophages with the prokaryotes and also within the metagenomic data. Most of the ORFs contained with the prophage 2 regions are involved in replication and re-arrangement (Figure 4.15).

Table 4.6: Predicted phage genomes

Predicted prophage	Size (bp)	Total No. of proteins	No. of hypothetical proteins	No. of conserved hypothetical proteins
Prophage 1	5645	6	3	1
Prophage 2	12314	15	6	4
Prophage 3	8423	9	6	0
Prophage 4	10606	5	3	0
Prophage 5	9252	13	7	1
Prophage 6	8618	13	7	1
Prophage 7	5182	4	3	0

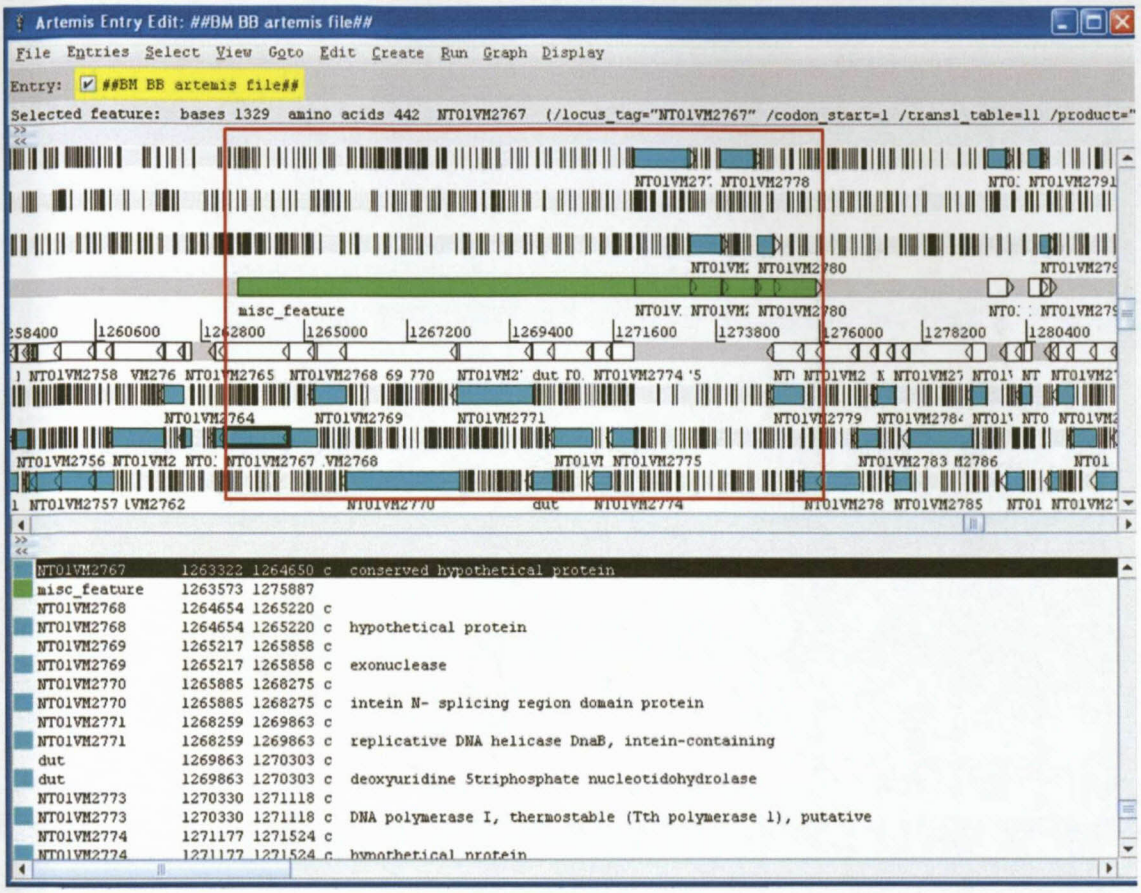


Figure 4.15: Artemis window showing a predicted prophage 2 region (Red rectangle) and the proteins that are comprised within the prophage 2.

4.4. Conclusions

Bacteriophages are the most abundant biological entities and different viral communities analyzed with shotgun sequencing to date show that more than 60% of the phage assemblages have not been previously characterized. The analysis performed using pyrosequencing reveals a high number of uncharacterized ORFs with hypothetical proteins amounting to more than 85%. In this study shotgun sequencing of viral communities from four South African deep mines also revealed high a number proteins not showing a significant hit to other phage proteins. The results suggest that the diversity of phages in mines is high, like other viral communities and they have not been previously characterized. Though shotgun sequencing is a very useful technique, introduction of the new sequencing technology, pyrosequencing is replacing this approach when whole genome sequencing is required. This technology can produce large amount of sequencing data in short time and costs are now comparable to Sanger sequencing in terms of time.

Beatrix mine had highest diversity when compared to the other three mines when using shotgun sequencing; hence this mine was selected for a pyrosequencing run. Two quarter plate runs of the GS FLX were done and after assembly the data was automatically annotated at TIGR. More than 75% of the proteins were hypothetical which indicates that a lot of phage proteins from this environment are unknown and unique. Phage proteins contributed 40% of the proteins with known function. These included proteins such as DNA and RNA polymerases, helicases, exonucleases, integrases, terminases and a variety of structural proteins. The sources of the proteins were mainly dsDNA tailed phages infecting the following host range: *E. coli; Burkholderia; Ralstonia; Pseudomonas; Vibrio cholera* and *Bacillus.* Proteins from *Acanthamoeba polyphaga mimivirus* were also observed. Seven prophages with minimum genome size of 5000 bp were also identified from Beatrix mine. The prophage regions contained a high number of hypothetical proteins. These findings show phages are the source of these proteins.

The overall results indicate that pyrosequencing can successfully be used in analyzing viral metagenomes and that phage communities from the SA mines are highly diverse, furthermore Beatrix mine comprised of all the dsDNA phages in addition to the unknown population.

Apart from identifying the viral assemblages novel proteins with biotechnological importance were also identified with the sequencing and annotation and few proteins were selected for expression studies, details in chapter 5.

Chapter 5

Expression of novel phage proteins from a Beatrix mine phage metagenome

Summary

Whole genome sequencing of viral environments constitutes a valuable resource for obtaining novel biocatalysts. In this study three phage proteins were identified among hundreds of known genes obtained with Beatrix mine sequencing project. Sequence analysis indicated that proteins were novel with similarity ranging between 39% and 46% to known enzymes. The genes encoding these proteins were cloned and expressed in *E. coli*. Two of the proteins; T7 phage DNA ligase and *Seg*B homing endonuclease were expressed functionally. The 41 kDa DNA ligase is ATP-dependent and ligated cohesive and blunt end fragments. The sticky ends were ligated more effectively than the blunt ends. Maximal ligation of the blunt ends was achieved with added polyethelene glycol (10%). The enzyme is active at 4 °C, 16 °C and 22 °C, and has the ability to ligate fragments into a vector. In addition ligation was obtained at temperatures as high as 70 °C.

The *Seg*B homing endonuclease proved to be active when used to cut lambda DNA. Complete digestion was not obtained indicating that the enzyme does not cut DNA randomly, but it recognizes a specific sequence.

Expression of the phosphatase/kinase in *E. coli* was not successful. Assays done with *BamH*I cut pUC19, dephosphorylated and phosphorylated for both domains indicated that the protein is not functional.

5.1. Introduction

Phage genomes have always been useful in molecular biology providing most of the proteins that are used in this field and other biotechnological applications (Cherepanov and de Vries, 2001; Stummeyer *et al.*, 2006). These proteins can be recovered from different sources ranging from phage genomes to metagenomes. With the discovery of novel proteins or gene products being the goal of almost every metagenomic or sequencing project, random genome sequencing technologies have greatly enhanced genomic analysis and discovery of novel viral genes from the environment and other metagenomes (Williamson *et al.*, 2008). Genomic analysis of the entire viral community is now possible using sequencing of shotgun libraries or pyrosequencing of viral communities (Breitbart *et al.*, 2008; Ng *et al.*, 2009; Angly *et al.*, 2006). Gene sequences encoding these novel proteins can therefore be accessed and then cloned using specific primers.

DNA ligases are one of the most frequently used enzymes in cellular processes and molecular biology (Doherty and Suh, 2000) and genes encoding this protein have been identified in eukaryotic, prokaryotic and viral genomes (Martin and McNeill, 2002; Wilkinson *et al.*, 2001; Doherty *et al.*, 1996). They catalyze formation of a phosphodiester bonds at single-stranded or double-stranded breaks between adjacent 3'-hydroxyl and 5'-phosphoryl groups of DNA (Timson *et al.*, 2000; Lehman, 1974). Ligases have been classified into two classes, ATP-dependent ligases and NAD^+-dependent ligases, based on the cofactor required for activity. All DNA ligases catalyze the reaction following the same mechanism which proceeds in the following three steps. First the enzyme is activated through the formation of a covalent phosphoamidate bond between the ε-amino group of the conserved active site lysine and the adenylate group of NAD^+ or ATP. In the second step, AMP is transferred from the ligase to the 5'-phosphoryl group of the nick on the DNA strand. The final step involves closing of the nick coupled the release of AMP from the adenylated DNA intermediate (Jeon and Ishikawa, 2003).

Intron-encoded homing endonucleases are phylogenetically widespread, occurring in archaea, bacteria, eukarya and bacteriophages. They catalyze the lateral transfer of an intron or intein to a homologous allele that lacks the sequence encoding its gene (Chevalier and Stoddard, 2001). They have been grouped into different families based on their conserved nuclease active-site core motifs and catalytic mechanisms; the LAGLIDADG, His–Cys box, GIY–YIG, and H–N–H

families (Kühlmann, *et al.*, 1999, Galburt and Stoddard, 2002). The LAGLIDADG and the His–Cys are highly specific, recognizing DNA target sites ranging from 14 to 40 base pairs (bp) in length (Stoddard, 2006). The last two families are generally encoded in phage introns and they have a broader and relaxed recognition pattern. Bacteriophage T4 have been shown to encode segB-D endonucleases and mobA-E genes which share homology to the GIY-YIG family (Belle *et al.*, 2001, Brok-Volchanskaya *et al.*, 2008).

Polynucleotide kinase/phosphatase (Pnkp) is an example of bifunctional enzymes with two distinct domains of approximately equal size separated by a flexible loop. The most exploited function of PNK is its ability to catalyze the transfer of the phosphate of a nucleoside triphosphate (e.g. ATP) to the 5' hydroxyl terminus of a polynucleotide. The enzyme also catalyzes the hydrolytic removal of 3' phosphate. The peptide comprise of the N-terminal 5' kinase and C-terminal 3' phosphatase activity, respectively (Zhu *et al.*, 2004, 2007). The N-terminus contains the A box which comprises of the two essential amino acid residues (Lys^{15} and Ser^{16}) for kinase activity (Wang and Shuman, 2001). The putative phosphatase motif; FDLDGTL is located at the C-terminus and alanine scanning indicated that Asp^{167} and other residues are important for the phosphatase activity. Amino acid residues responsible for both kinase and phosphatase activity are different, in addition the activities of both domains are independent of each other (Wang and Shuman, 2002).

In this study three novel proteins identified by pyrosequencing of Beatrix mine metagenome were expressed in *E. coli*. Primers specific for respective proteins were designed using the available pyrosequencing data. Genes encoding DNA ligase, polynucleotide kinase and *Seg*B homing endonuclease were PCR amplified from the isolated viral DNA. Protein products were purified and assayed for activity.

5.2. Materials and Methods

5.2.1. Novel viral proteins from the Beatrix mine

Different phage genes were identified by ORF identification and annotation of contigs assembled from pyrosequencing data (Sections 4.3.8 and 4.3.9), and the following proteins were then selected for expression and characterization, DNA ligase, Seg B homing endonuclease and polynucleotide kinase. The proteins show low similarity (indicated in the brackets) to the proteins from the following sources, *E. coli* phage rv5 (46%), *Staphylococcus* phage (45%) and *Enterobacteria* phage RB14 (39%), respectively. Phylogenetic analysis drawn with 100 BLASTX results further showed that the proteins are novel (Appendix C).

5.2.2. Cloning of the selected proteins

Primers specific for the amplification of the three proteins were designed using the available sequences from the whole genome pyrosequencing data. Amplified phage DNA extracted from Beatrix black beauty biofilm was used as the template for all the amplifications using Expand High Fidelity ^{PLUS} PCR System (Roche Applied Science). Primers LligF and LligR (Table 5.1) were used to amplify a DNA ligase gene. The following PCR conditions were used: initial denaturation at 94°C for 2 min, 30 cycles of the following steps were done; denaturation; 94°C for 30 sec, annealing at 55°C for 45 sec and elongation at 72 °C for 90 sec. The final elongation was done at 72°C for 10 min.

Cloning of the endonuclease and the kinase genes was performed using the following primer pairs; LEndoF & LEndoR and LKinF & LKin2R, respectively. The same PCR conditions as above were used to amplify the endonuclease gene with elongation at 60 sec instead of 90 sec. Annealing was done at 58°C for 45 sec and elongation at 72 °C for 1 min to amplify the gene encoding the phosphatase kinase. After PCR all products were excised from the gel and purified using the Biospin Gel Extraction kit (Separation scientific) according to the manufacturer's instructions. After sub-cloning into pGem-T Easy the plasmids were propagated in *E. coli* Top 10 cells. Plasmid DNA was extracted using the lysozyme boiling method as described in chapter 2 section 2.2.2.1. All basic recombinant DNA cloning techniques were done using the methods as described by Sambrook *et al.* (1989).

Table 5.1: Oligonucleotides used

Primers	Sequence in the 5' to 3' direction	Restriction site
LligF	CGCATATGAGCTTCGACACTCTGTACC	*Nde*I
LligR	CTCGAGTTACAGATCCTCTTCCATGCG	*Xho*I
LKinF	GCTAGCATGATTACCTTAATGGTTGGAC	*Nhe*I
LKin2R	CTC GAG TTA ATA ACC CCT TCT TCT CCA	*Xho*I
LEndoF	CGCATATGACTCTTTATAGAATTGGTA	*Nde*I
LEndoR	CTCGAGTTATTCTCCCTTAATTCTTTC	*Xho*I

5.2.3. Expression of phage proteins in *E.coli*

Clones containing inserts of the correct size in pGem-T Easy were used to release the gene fragments using the restriction enzymes incorporated in the primers. The DNA ligase and the endonuclease were cut with *Nde*I and *Xho*I, and *Nde*I and *Nhe*I were used for the kinase gene. The fragments encoding DNA ligase; Seg B homing endonuclease and the polynucleotide kinase were ligated separately into pET28b at 4°C overnight. Expression vectors were first propagated in *E. coli* Top 10 cells and the recombinant clones were selected on Kanamycin (50 µg/ml). Clones containing the correct inserts were then used to transform the expression host, *E. coli* BL21 (DE3) pLys. Expression studies were performed as follows; a single colony was used to inoculate 5 ml LB media (containing Kananamycin 50 µg/m and chloramphenicol 30 µg/ml), followed by incubation at 37°C overnight. The pre-inoculum (2 % (v/v) of the overnight culture) was then used to inoculate 100 ml LB medium, the culture was incubated at 37 °C until the OD_{600nm} was 0.5. The expression was induced by addition of IPTG to a final concentration of 0.5 mM, then incubation at 30°C for 5 hrs. The expression of the endonuclease was expected to be toxic to the cells; hence production of the protein was induced using 0.25 mM IPTG.

Cultures were pelleted and cells lysed using B-PER (in phosphate buffer, 50mM (pH 7.5)) bacterial protein extraction reagent (Pierce Protein Research Products) following the manufacturer's instructions, to release the intracellular proteins. Supernatants were then ultra-centrifuged at 30 000 RPM for 90 minutes using the SW 32 Ti rotor (Beckman, Optima™ L-100XP) to remove membrane proteins.

5.2.4. Functional assays

5.2.4.1. DNA ligase assays

The expressed protein was purified using a 5 ml His-trap column (GE Healthcare) and the Äkta prime plus system (Amersham Biosciences). The system was flushed with binding buffer (50 mM Tris-HCl, pH 7.5 containing 5 mM DTT, 20 mM imidazole and 0.5 M NaCl), the column was then connected and the protein was loaded. The column was washed with 10X volume of the column with binding buffer and 5 mL fractions were collected with the same buffer containing imidazole to a final concentration of 250 mM. Fractions containing purified protein were confirmed with 12.5% SDS-PAGE. Activity of the ligase was determined using lambda DNA cut with different enzymes. To define a unit; different amounts of DNA ligase were used to ligate 200 ng of λ DNA cut with *EcoR*I and *Hind*III. A unit of DNA ligase is defined as the amount of enzyme required to ligate back 50% of cut Lambda DNA in 20 min. The reaction mixture was incubated at 22 °C for 20 minutes. The enzyme concentration was determined with BCA protein kit following manufacturer's instructions (Pierce Protein Research Products). Lambda DNA, 200 ng cut with *EcoR*I and *Hind*III was used to assay the expressed ligase for ligation of the sticky ends; and *Hae*III cut DNA was used for the blunt ends. The enzyme was assayed at 4 °C, 16 °C and 22 °C. The reaction mixture was incubated for 16 hrs for the assays done at temperatures 4 °C and 16 °C. Different time intervals; 30 min, 1 hr, 1hr 30 min and 2 hrs were used for the assays done at 22 °C. The protein was also tested for its ability to ligate fragments at higher temperatures. In this case pUC19 (100 ng) cut with *Bam*HI was used as the substrate, and the assays were performed at 30, 40, 50, 60 and 70°C and reaction mixtures were incubated for 30 minutes.

The blunt end ligation was done with and without polyethylene glycol 6000 (PEG), and increasing concentrations (0 to 25 %) of PEG were used. The ability to ligate fragments into a vector was also tested using a ~ 650 bp fragment cut with *Nde*I and *Xho*I and a pET vector cut with similar enzymes. L igation reactions were used to transform *E. coli* Top10 cells after incubation for 2 hrs at 22°C. The plasmids were cut with similar enzymes to check the presence of the insert. Commercial ligases from Fermentas, Promega and Kapa Biosystems were compared with the cloned ligase when doing the above assays and 1 unit as described the manufacture was used to do the assays.

The DNA ligase was also assayed for its ability to ligate the 3' and the 5' cohesive ends using pUC 19 cut with *Kpn*I and *BamH*I, respectively. The plasmid (200 ng) was cut with these enzymes separately and the assays were done at 22°C for 1hr with the same amount of protein.

The reaction buffer (50 mM Tris-HCl, pH 7.5, 10 mM DTT) contained additional 5 mM ATP and 10 mM $MgCl_2$. The enzyme was inactivated by incubation at 65 °C for 10 minutes and the ligation products were monitored on 1% agarose gel.

5.2.4.2. SegB homing endonuclease

The expressed endonuclease protein was purified manually using High density Nickel 6BCL-HQNi-500 resin (Agarose Bead Technologies) and the protein was eluted with 50 mM Tris-HCl buffer (pH 8) containing 1M NaCl and 250mM imidazole. Fractions containing the purified protein were identified with 12.5% SDS-PAGE, followed by dialysis against 50 mM Tris buffer (pH 8) which was changed twice. The protein concentration was determined using Quick Start™ Bradford Protein Assay (Bio-Rad) method following instructions as described by the manufacturer. To assay the activity of the endonuclease, lambda DNA (150 ng) was used. The endonuclease activity was assayed at 37 °C for 1 hr using 0.11 µg of the enzyme. The following buffer conditions; 10 mM Tris-HCl (pH 8), 10 mM $MgCl_2$ 100 mM KCl and 0.1 mg/ml BSA were used. The reaction was stopped by heat inactivation at 65 °C for 10 minutes. The digestion products were immediately analyzed on 1% agarose gel and the uncut lambda DNA was compared to with the digested DNA.

5.2.4.3. Phosphatase kinase

Primers; LKinF and LKin2R (Table 5.1) were used to amplify a kinase fragment, and the expression was done as described above. The protein was manually purified using His-select ™ Nickel affinity gel (Sigma) and eluted with 250 mM imidazole concentration by collecting 2 mL fractions. Fractions containing purified products were analyzed by SD S-PAGE and protein concentrations were determined with the BCA protein determination kit (Pierce Protein Research Products). The enzyme contains both the phospatase and the kinase domains, and the activity of the domains was assayed differently using pUC19 plasmid cut with *Bam*HI. The ability of the phosphatase domain to remove the 5' phosphate group was determined using digested pUC19. *Bam*HI cut pUC19, 100 ng was treated with the expressed protein at 37 °C for 1 hr. The reaction mixture (10 µl) contained the following buffer system, 50 mM Tris-HCl, 1 mM $MgCl_2$ and 0.1 mM $ZnCl_2$ at pH 8.0. A negative control was also included and contained no enzyme, only the reaction buffer and the cut plasmid. Phosphatase treated products were ligated using T4 DNA ligase (Fermentas Life Sciences) at 22 °C for 2 hrs. Ligation mixtures

were then used for the transformation of *E. coli* Top 10 cells. The digested vector was also included and used as the transformation control and it contained the reaction buffer as a way of checking effect of the buffer on the transformation efficiency.

Kinase activity was determined using commercially available pUC19 (100 ng) cut with *Bam*HI and dephosphorylated (Fermentas Life Sciences). The plasmid was treated with kinase protein using the following buffer assay conditions; 70 mM Tris-HCl [pH 7.8], 10 mM $MgCl_2$, 0.5 mM DTT and 10 mM ATP in a total volume of 10 µl. The assays were performed at 25 °C for 1 hr, and the phosphorylated reaction mixtures were then were then ligated at 22 °C for 2 hrs. *E. coli* Top 10 competent cells were then transformed with ligation mixtures using methods as described by Russell and Sambrook, (2001). Unligated *Bam*HI digested vector containing the reaction buffer was included to check effect of the buffer on transformation efficiency.

5.3. Results and Discussions

5.3.1. Expression studies

5.3.1.1. DNA ligase

A band of approximately 1100 bp (Figure 5.1) encoding a DNA ligase was amplified from Black beauty biofilm phage DNA. The elution profile (Figure 5.2) showed that fractions 5 and 6 possibly contained the protein; hence they were analyzed on SDS-PAGE (Figure 5.3). The protein was about 47 kDa with the His tag (approximately 41 kDa without the tag) which is the expected size of phage DNA ligases. Sequencing of the cloned gene revealed a novel ATP-dependent ligase. From the results it is evident that a novel T7 phage related DNA ligase has been cloned from Beatrix mine biofilm phage DNA. Though the protein showed low similarity to the available ligases at amino acid level as expected with ligases, the conserved motive KLDGVR containing the catalytic lysine was identified (Figure 5.4). The second most conserved motif; SLRFPRFIRIR is mostly found in eukaryotic and archaeal ATP ligases and not in phage ligases as indicated previously (Kletzin, 1992). In this study the second motif was also not detected with the cloned phage ligase from the mine (Figure 5.5).

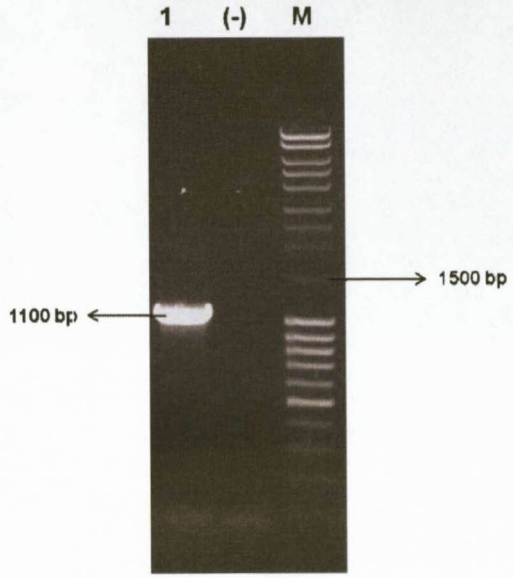


Figure 5.1: PCR amplification of DNA ligase, lane 1 represents the product, lanes (-) and M corresponds to the negative control and the DNA marker used, respectively.

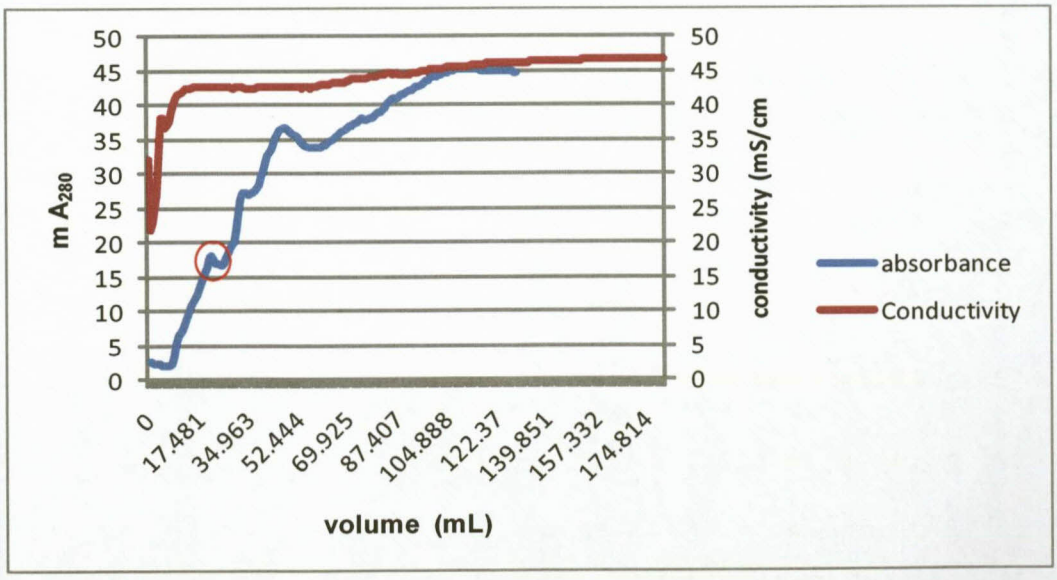


Figure 5.2: Elution profile of the DNA ligase purified with His-Trap column, the flow rate was set at 2 ml/minute and 5 ml were collected. The fraction containing the protein is circled.

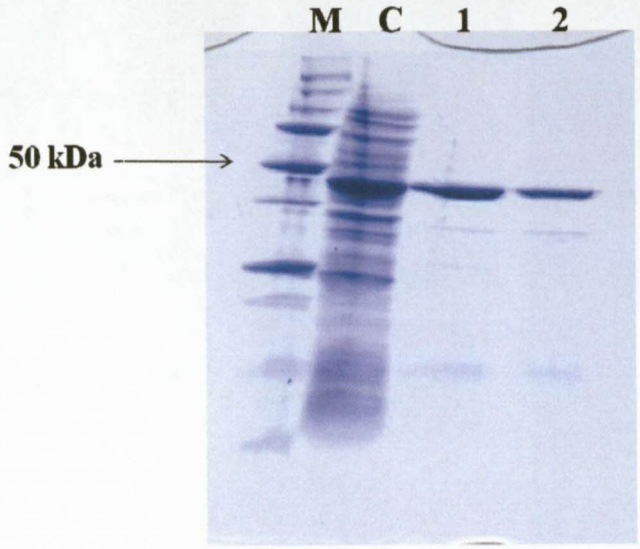


Figure 5.3: Purification of the DNA ligase using the His-Trap column, lane M is the protein ladder, lane C is the crude and 1 and 2 are the purified products from fractions 5 and 6, respectively.

```
NT01VM2994      --MSFDTLYHKGSSGAIYSWKVWTEGADVCTEYGQVDGKKQLARKTAEAKNVDKSNATTP 58
PhageRv5        MHQFKTTLYARNKDGSIQMWKVDTSGNEVIVVFGRVGGAVQSKTTKCEAKNIGRSNETTA 60
                      *** :...*:*  *** *.* :* . :*:*.*  *   ...****:.:** **.

NT01VM2994      EEQAVLEATSMWTHRVERKYRKTLDEANNDEIFLPMLASEFEKRRGQKKDGHTYKCDVQP 118
PhageRv5        EEQAVLEAQSKWEKQVRLGYKENVEDLEAIDIS-PMLAQDASK----KPHAIVYPCHLQH 115
                ******** * * ::*.  *::.::: :  :*  ****.: .*    * .. .* *.:*

NT01VM2994      KLDGVRCLAYWDGEEVKLLSRGGKDYN-CPHIVEVLKNVLPHN--LVLDGELYLHGVGFQ 175
PhageRv5        KLDGNRCFVKFVDGVPKFISRGNKVYEPKGNILRELQDLHEETGFDEFDGEFYIHGLPLQ 175
                **** **:. : .   *::***.* *:   :*:. *::: ..    :***:*:**: :*

NT01VM2994      TITSWVKRLQ------------------------------ENTSKIQYHVYDCVLLD-- 202
PhageRv5        KITSLVKKWRSLEDIEKEIDKDFMADIKRREKAIKAGEETWKDFNKVDHPVYEEPVRDSD 235
                .*** **: :                                :: .*::: **:  : *

NT01VM2994      -----------------ERKAQWPSRYSSLHDFFSTRHKRLNGVIHLVDTYEAQNEDEV 244
PhageRv5        RYGGYCSYDLKLMVFDVPDCIHRWEDRATNLQEVIDYCEVNHLTNVEGVPFKISYNEEEV 295
                          :   :* .* :.*::.:.  . .    :. *    : **:**

NT01VM2994      LRVHNQVVSEGYEGAIVRMYDNSEYRFAYRSKRLLKVKSFSDNEYRVVGYTTGKGRFEDC 304
PhageRv5        RNSIGQYMQEGYEGSIIRNFRG-TYEYGQRSTDLLKWKLFTDVEAKVIGVEKDKN--GEG 352
                 .  .* :.*****:*:* : .  *.:. **. *** * *:* * :*:*  ..*.   :

NT01VM2994      AVWVCQAGDHTFQVVSKGTMEERRKMLEEADSYVGQLLKVKYFELTDEGIPRFPVGVGFR 364
PhageRv5        VLICEEKDGTQCKCKMKGTFAVRN--YEKCCSYVGKFITIKFQQRTVDGVPQFPVGIAFR 410
                .:   : ..   :   ***:  *.   *:. ****:::.:*: : * :*:*:****:.**

NT01VM2994      MEEDL------ 369
PhageRv5        NLNPETWEVLE 421
                  :
```

Figure 5.4: Protein alignment of the expressed DNA ligase (NT01VM2994) to its closest hit *E. coli* phage Rv5 ligase, the catalytic site is highlighted with red.

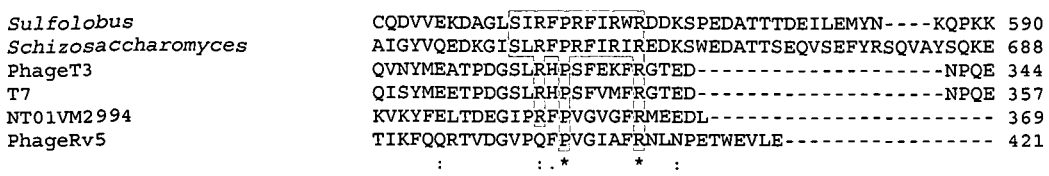


Figure 5.5: Amino acid sequence alignment of the N-terminal of the DNA ligases, the second motif is highlighted in archaea and eukaryotic ligases as well as conserved amino acids in phage ligases.

The protein concentration was determined using the BCA protein assay method (Figure 5.6), 0.3 µg of enzyme efficiently ligated sticky ended fragments at different temperatures. Lambda fragments cut with *EcoR*I and *Hind*III were ligated at 4 °C and 16 °C after 16 hrs (Figure 5.7). Sticky end ligations performed with Fermentas ligase resulted in bands with highest molecular weight at all temperatures (Figure 5.7) indicating that the enzyme is more effective. Since the activity was monitored by joining of lambda fragments, the Beatrix mine ligase was more effective than Kapa ligase as the enzymes joined more bands under the same reaction conditions. When assays were done at 22 °C with sticky ended lambda fragments the bands were ligated within the first 30 minutes of the reaction (Figure 5.8). All the assays were not quantitative and activity was monitored by electrophoresis.

Blunt ended DNA was not ligated as efficiently as the sticky ends. Initial ligations done with blunt ended DNA failed at all temperatures that were tested indicating that the expressed enzyme has low affinity towards blunt end fragments. Commercial ligases from Fermentas and Kapa were however able to ligate blunt ends as a smear was obtained indicating some form of ligation (Figure 5.7 and Figure 5.9). Addition of the PEG enhanced the ligation reaction with increasing concentration showing improved ligation of the bands (Figure 5.10). The highest ligation was obtained with 10% PEG and further increase in the concentration did not improve the ligation efficiency. Results indicate that the enzyme is effective in ligating sticky ends and blunt ends at different temperatures although the blunt end ligation efficiency is low. This was however expected as literature shows that most ligases require PEG to efficiently ligate blunt ends (Pheiffer *et al.*, 1983). In addition commercial ligases used in this study also require PEG at different concentrations as indicated by the manufacturer.

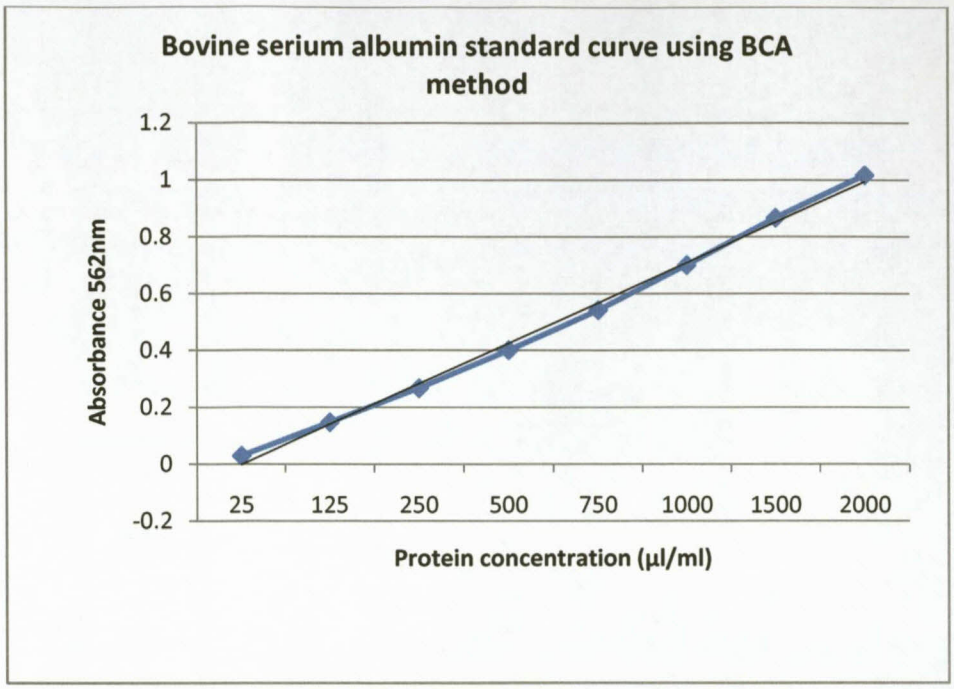


Figure 5.6: BSA standard curve constructed using BCA protein assay method.

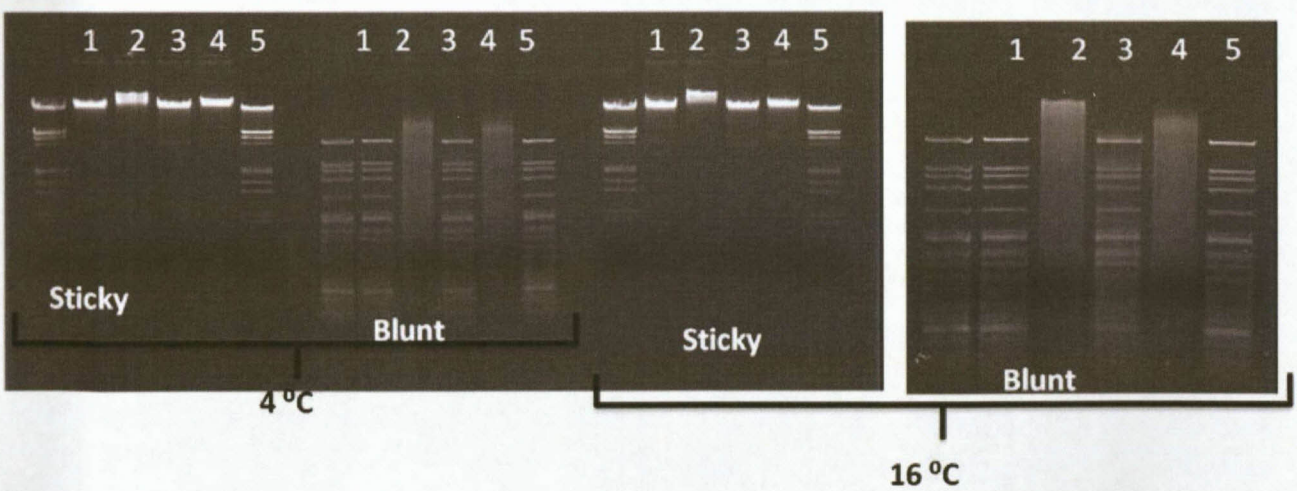


Figure 5.7: DNA ligase assays done at 4°C and 16°C for sticky and blunt ended lambda DNA, lanes 1, 2, 3, 4 and 5 corresponds to the assays done with cloned protein, Fermentas, Kapa, Promega and the negative control, for different reaction conditions as indicated.

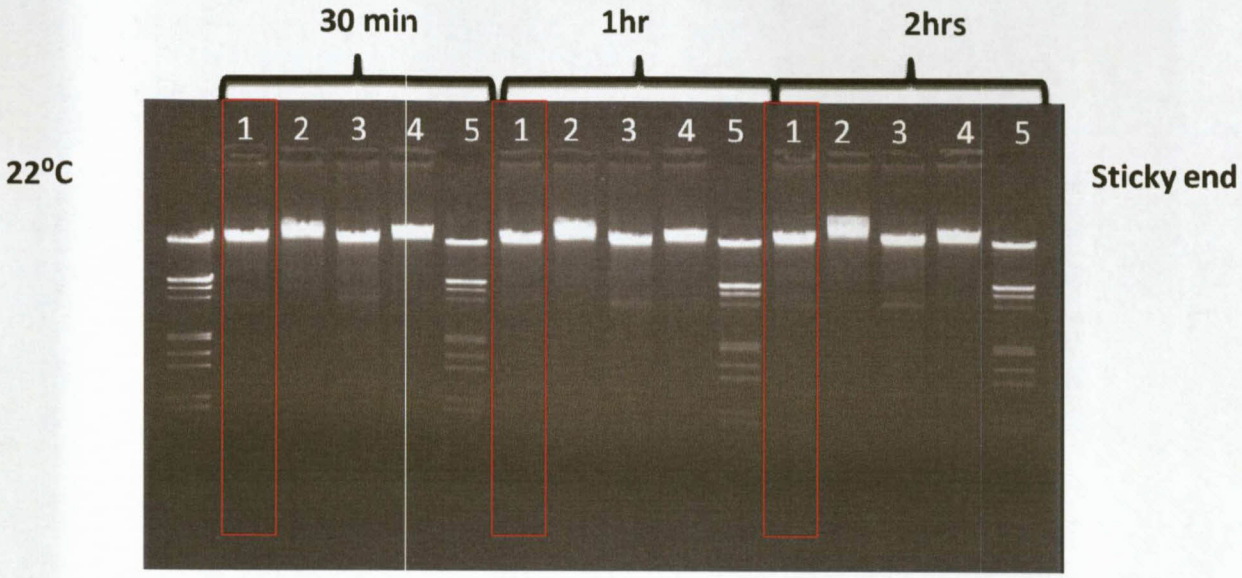


Figure 5.8: DNA ligase assays done at 22°C for sticky ended lambda DNA, lanes 1, 2, 3, 4 and 5 corresponds to the assays done with cloned protein, Fermentas, Kapa, Promega and the negative control, respectively.

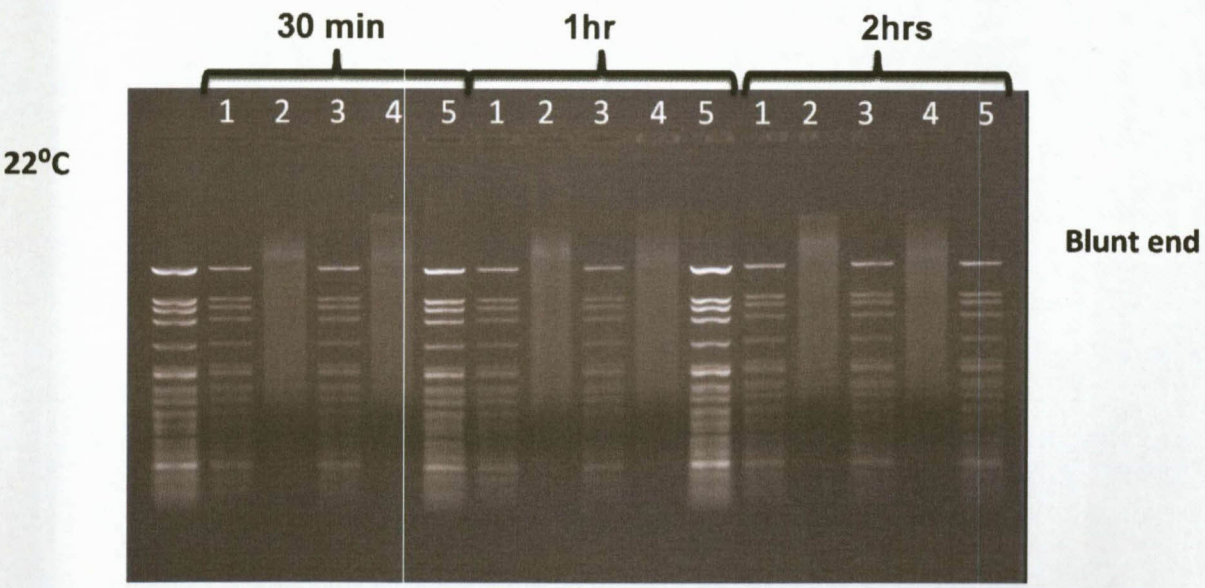


Figure 5.9: DNA ligase assays done at 22°C for blunt ended lambda DNA, lanes 1, 2, 3, 4 and 5 are cloned protein, Fermentas, Kapa, Promega and the negative control which were used for the assays.

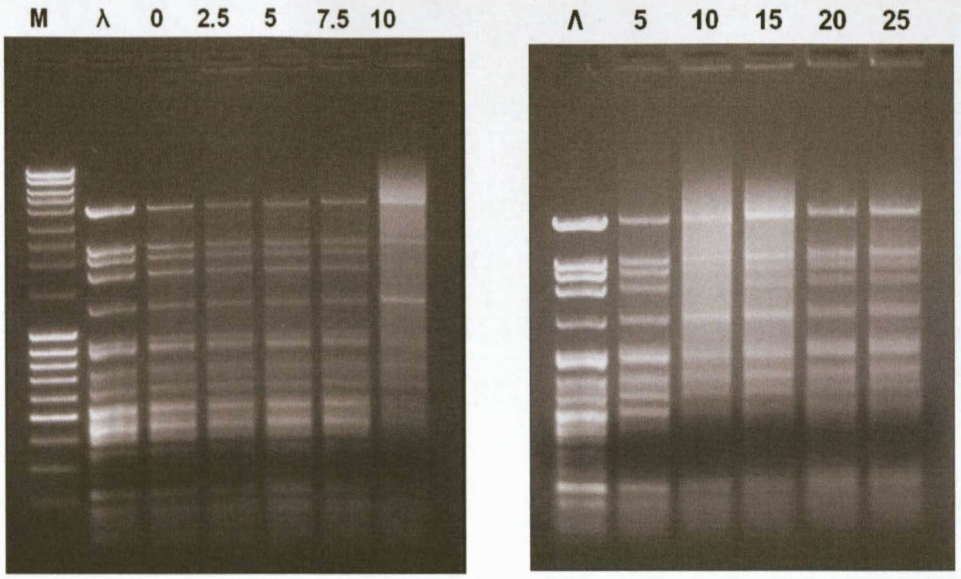


Figure 5.10: Blunt end DNA ligation with added PEG, lanes 0 to 25 are the increasing % (w/v) concentrations of PEG. The DNA ladder used is indicated on lane M and on lanes λ are unligated cut lambda DNA.

Ligation reactions performed at high temperatures showed that the protein is functional at high temperatures. Minimal ligation of the pUC19 vector was obtained at 30 and 60°C, however endonuclease activity was also observed when ligation was performed at these high temperatures (Figure 5.11). The expressed ligase showed activity at temperatures as high as 70 °C. Though thermo-stability studies were not done the results demonstrate that the protein could be thermo-stable as the reaction mixtures were incubated for 30 minutes at respective temperatures.

In addition the protein was able ligate fragments into a vector, as the inserts of the correct size were obtained after restriction analysis of the extracted plasmids (Figure 5.12).

The DNA ligase was also assayed for its ability to ligate the 3' and the 5' cohesive ends using pUC 19 cut with *Kpn*I and *BamH*I, respectively. The results showed that the DNA ligase is capable of ligating both 5' and 3' cohesive ends. Ligation of the 5' sticky ends resulted in all the three forms of the plasmid and the 3' only produced the nicked form (Figure 5.13).

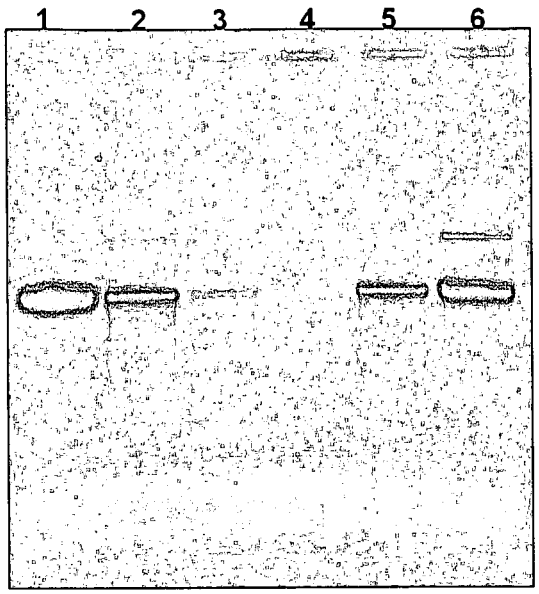


Figure 5.11: DNA ligase assays done at different temperatures, *Bam*HI cut pUC19 is on lane 1 and lanes 2 - 6 are reaction temperatures 30, 40, 50, 60 and 70 °C, respectively.

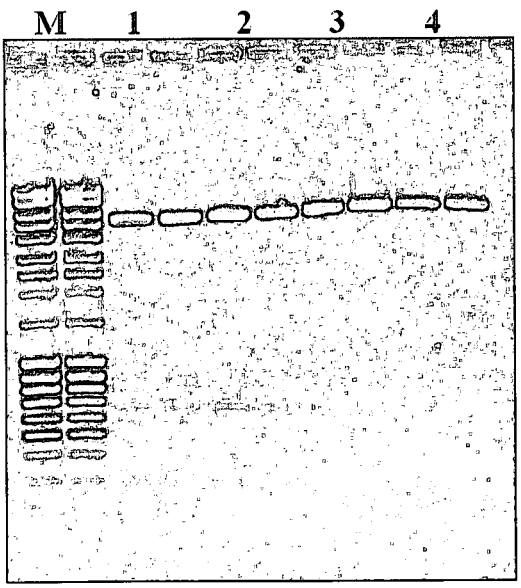


Figure 5.12: DNA ligase activity assay indicating the ligation of fragments into a vector, DNA ladder used is indicated on lane M. Different ligases: Beatrix mine ligase, Fermentas, Kapa and Promega are on lanes, 1, 2, 3 and 4, respectively.

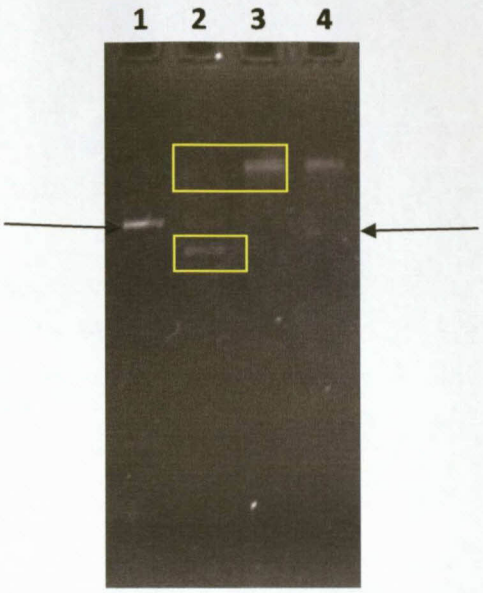


Figure 5.13: Ligase assays for the ability to ligate the 3' and the 5' cohesive ends, lanes 1 is pUC 19 cut with *Bam*HI (for 5' cohesive ends) and on lane 2 is the ligated product. The *Kpn*I (3' cohesive ends) digestion is represented on lane 4 and the ligation on lane 3. Digested plasmids are indicated with arrows and ligation products with yellow rectangles.

5.3.1.2. The endonuclease

A band of approximately 654 bp encoding the SegB homing endonuclease was PCR amplified (Figure 5.14) from Beatrix mine phage DNA. The endonuclease gene was expressed at low levels in *E. coli* as the gene product was expected to be toxic and the purified protein (Figure 5.15) was used for the assays. Digestion of lambda DNA resulted in a smear indicating that the protein is functional (Figure 5.16). The endonuclease is part of the GIY-YIG family; therefore it cleaves DNA with relaxed specificity. The enzyme is tolerant of base-pair changes in its homing site and a variety of insertions and/or deletions between the cleavage and insertion sites are permitted (Stoddard, 2006). However some sites are still highly preferable for the cleavage than others. This phenomenon was observed with our results whereby complete digestion of lambda DNA was not obtained after an hour of incubation suggesting that the protein does not cut DNA randomly. Furthermore the *Seg*B endonuclease from the T4 phage has been shown to be site specific, recognizing a 27 bp sequence (Brok-Volchanskaya *et al.*, 2008). The preferred or recognition site for cleavage was not determined as this would require a detailed study.

Catalytic motif and conserved residues of the GIY-YIG family are located within the first 70-100 amino acids of the protein. The expressed protein contained YRI-YVG and the alignments

further indicate that most phage endonucleases comprise of YXI-YVG instead of the typical GIY- YIG motif. The three conserved amino acid residues Phe56, Glu75 and Asn90 which forms part of the catalytic residues were identified with the expressed endonuclease from Beatrix mine (Figure 5.17). The protein also had Arg at position 27 (Figure 5.17), though it has been shown that some T4 phage *Seg* proteins are intergenic endonucleases and they do not have this conserved residue (Kowalski *et al.*, 1999).

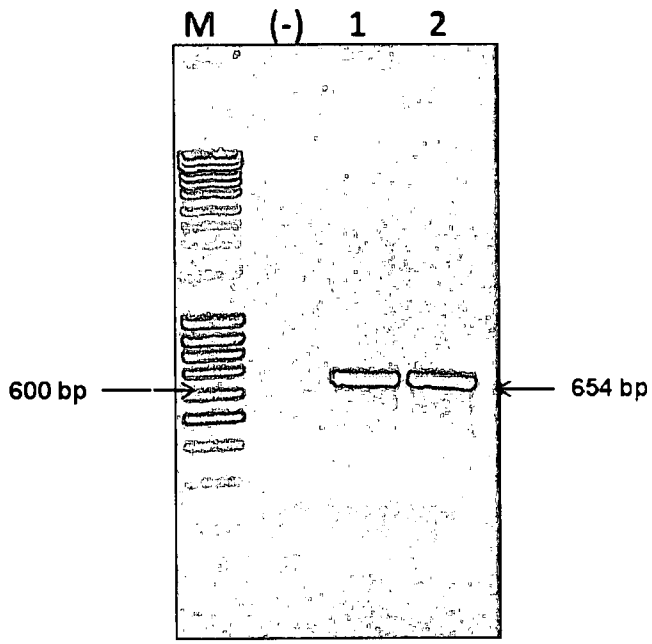


Figure 5.14: PCR amplification of the endonuclease, the DNA ladder used is represented on lane M, the negative control with lane (-), and amplified endonuclease product on lane 1 and 2.

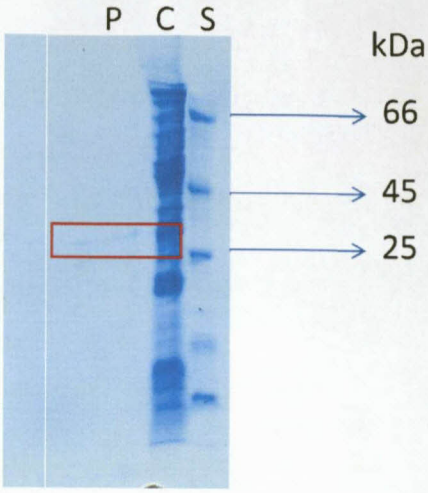


Figure 5.15: Expression and purification of the endonuclease, the protein ladder is represented with lane S and the purified product and crude with lanes P and C, respectively. The expected band is in the red rectangle.

Figure 5.16: Endonuclease activity assays, the uncut lambda DNA is on lane 1 and the lambda cut with the expressed endonuclease is represented on lane 2.

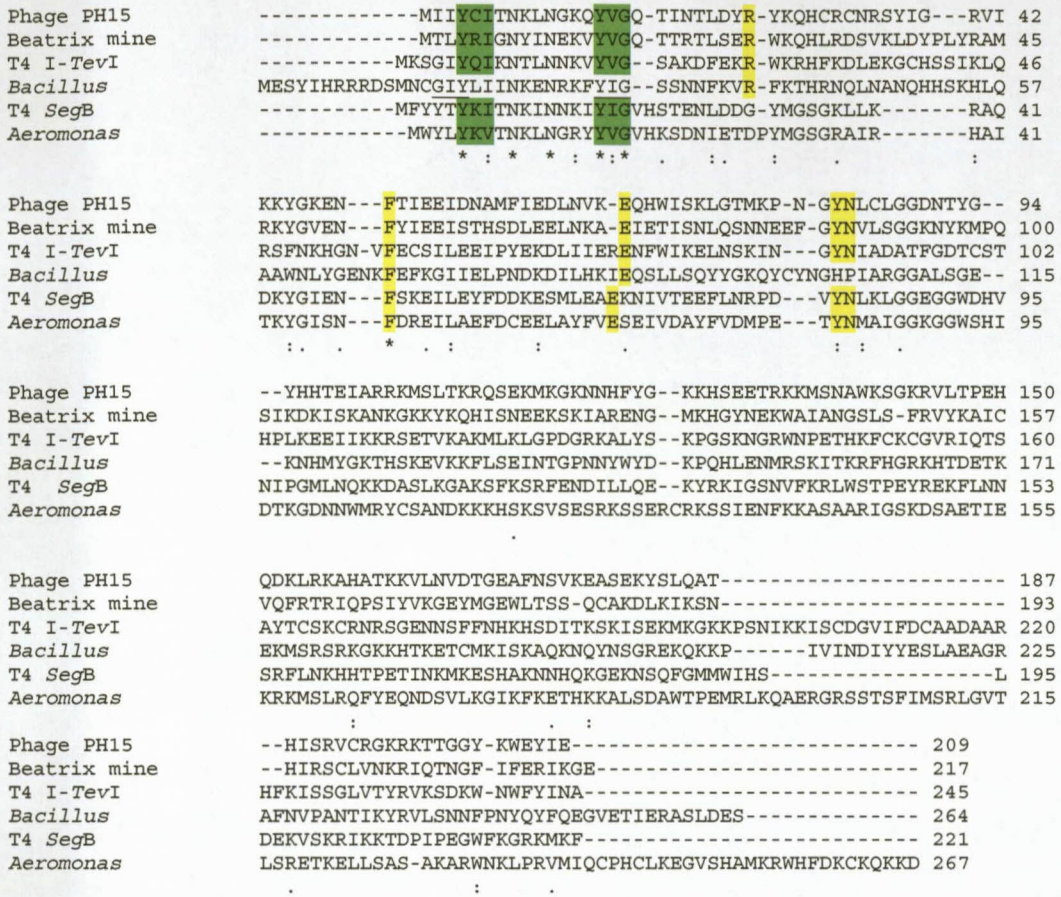


Figure 5.17: Amino acid sequence alignment of the endonucleases, the protein expressed in this study is indicated as Beatrix mine. Other sequences were obtained from the following sources, phage PH15 is from *Staphylococcus* phage PH15 (accession number YP950665), *Bacillus* – *Bacillus cereus* (accession number ZP04250233), I-TevI and SegB are both sequenced T4 phage genome (accession number NC000866). All conserved catalytic residues are highlighted in yellow, and the GIY-YIG moti is underlined with *Bacillus* sp. The typical motif, YXI-YVG obtained with phages is highlighted in green.

5.3.1.3. Phosphatase kinase

Polynucleotide kinase/phosphatase is a bi-functional protein and a gene fragment of 924 bp encoding these domains was PCR amplified (Figure 5.18) from Beatrix mine phage metagenome DNA. The protein was expressed at low levels and various attempts to increase the level of expression and yields of purified protein were unsuccessful. When the purified protein (Figure 5.19) was used for the assays no activity could be detected with both domains. Colonies were obtained as expected with the kinase treated vector; however the negative

control also resulted in colonies indicating that the activity is not due to the expressed protein. The following residues are essential for the kinase activity; Lys15, Ser16, Asp35, Arg38 and Arg126 (Wang and Shuman, 2001, 2002). The first two amino residues are located within the nucleotide binding (Waller A box) motif; GxxGxGKST. Sequence alignment to the T4 phage polynucleotide kinase/phosphatase revealed that the expressed protein had the following substitutions; D35I, R38D and R126E (Figure 5.20). Suggesting that these substitutions are responsible for the inactivation of the kinase protein.

The activity of the expressed phosphatase protein was also not detected when using *BamH*I digested pUC19 as the substrate. Colonies were obtained after transformation indicating the vector was not dephosphorylated. Equal number of colonies were obtained with the phosphatase treated vector and the negative controls. The following amino acid residues; Asp165 and Asp167 are located within FDLDGTL phosphatase motif. In addition the amino acid residues; Arg176, Arg213, Asp254 and Asp278 are also important for the phosphatase activity (Wang and Shuman, 2002). The protein had R176N substitution suggesting that the mutation is responsible for the inactivation of the phosphatase domain as the other catalytic residues were still conserved (Figure 5.20). The results demonstrate that a number of residues are involved in the active sites of both phosphatase and kinase domains. Mutations at these sites could be responsible for the loss of activity of this bifunctional protein from Beatrix mine.

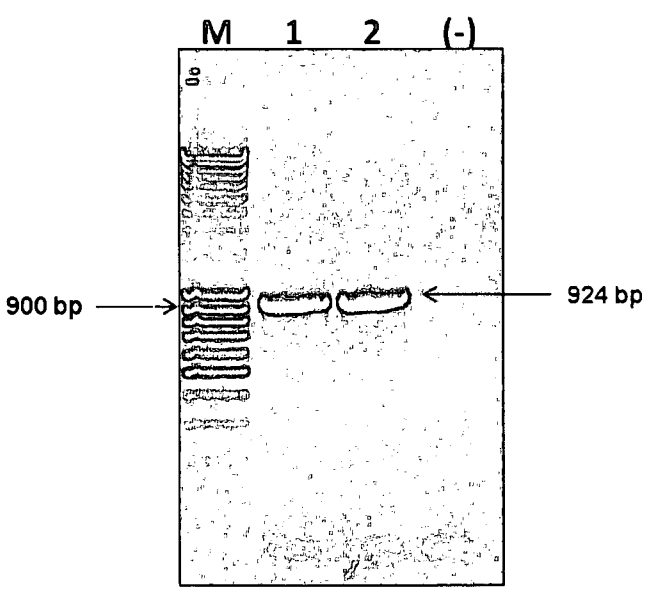


Figure 5.18: Amplification of polynucleotide kinase, the PCR products are represented on lanes 1 and 2. On lane is the DNA ladder used and the negative control on lane (-).

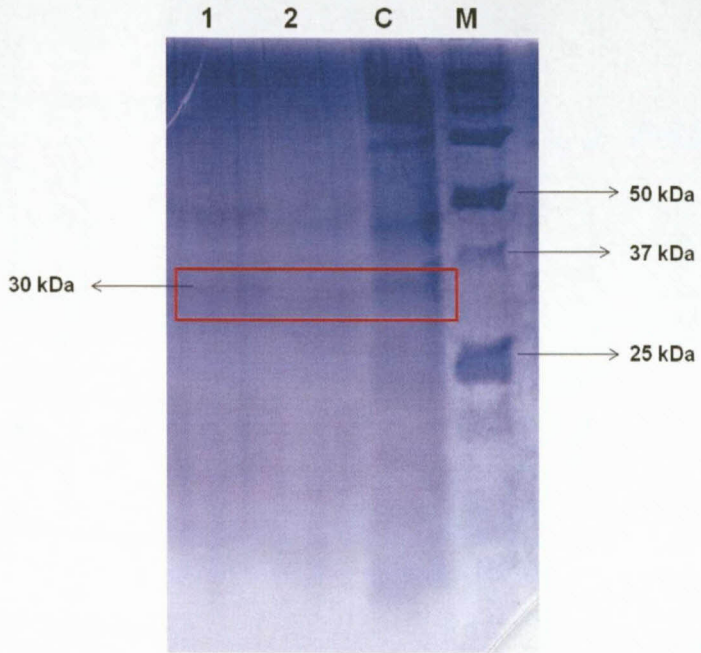


Figure 5.19: Expression of the kinase phosphatase gene, lanes 1 and 2 represent the purified protein and lane C is the crude, the expected band is within the red rectangle. Lane M is the standard protein ladder.

```
ORF00571        --MITLMVGPPGSGKSTLAEEYCRKHPCPPNCTRISQDDQGKEGHMGLFNTALMARSDLV 58
NC000866        MKKIILTIGCPGSGKSTWAREFIAKNPGFYNINRDDYR-QSIMAHEERDEYKYTKKKEGI 59
                   * * :* ******* *.*:  *:*   * .* .   *.  .*    :     :.: :

ORF00571        IDRMNFDKNQRNRYLEPARKAGYATRIIVVHCPLDTCLERCGKRENHPTIKDSKAASQAV 118
NC000866        VTGMQFDTAKSILYGGDSVKG-------VIISDTNLNPERRLAWETFAKEYGWKVEHKVF 112
                :  *:**. :   *   : *.       *: .  :   **    *....  . *.  :..

ORF00571        NFFFSHYERVEDNRADEVIRKGWAGDYAPSAIICDLDGTLCNVEHRRHFVRPPKDFYIKT 178
NC000866        DVPWTELVKRNSKRGTKAVP---------------IDVLRSMYKSMREYLGLPVYNGTPG 157
                :. ::.  : :.:.. :.:               :*   .  :  *.::  *

ORF00571        LGVLEPVDPDAPLPVFKKNWPAFFKGIKDDAVNQWCAGILRAMSSK-HIIVYCSGRSDNE 237
NC000866        KPKAVIFDVDGTLAKMNGRGPYDLEKCDTDVINPMVVELSKMYALMGYQIVVVSGRESGT 217
                     .* *..*. :: . *  ::  . *.:*   . : :  :   : ** ***...

ORF00571        RKATVEWLEKHGLDVMDGAEVPLFMRNRQ---DSRRDDVVKEIILDFEILPRYNPYFMID 294
NC000866        KEDPTKYYRMTRKWVEDIAGVPLVMQCQREQGDTRKDDVVKEEIFWKHIAPHFDVKLAID 277
                :: ..:: .     * * * ***.*: ::   *:*:****** *:  .* *:::  : **

ORF00571        DRQQVVDMWRRRGY---------- 308
NC000866        DRTQVVEMWRRIGVECWQVASGDF 301
                ** ***:**** *
```

Figure 5.20: Protein alignment of the polynucleotide kinase/phosphatase to the T4 phage endonuclease, T4 sequence is obtained from the sequenced T4 phage genome (accession number NC000866). Catalytic residues are highlighted in green for kinase and in yellow for the phosphatise domain. The conserved motifs are underlined.

5.4. Conclusions

Random genome or metagenome sequencing provides access to genetic and genomic information of a variety of microbial and viral communities from the environment. In this study phage proteins encoding different metabolic and cellular functions were identified with Beatrix mine phage sequencing project. Three DNA processing phage proteins; DNA ligase, *Seg*B homing endonuclease and polynucleotide kinase, were selected for expression studies in *E. coli*. The proteins play crucial roles in most basic cellular functions, including DNA repair, DNA replication, and transcription. DNA ligases are useful in closing the nick between the 3'-OH and 5'-phosphate group of the DNA. The endonuclease initiates transfer of an intron or inteins in a process called homing, leading to gene conversion or inheritance of the mobile element. Kinases catalyze the transfer of the phosphate from ATP to a variety of molecules, while phosphatases catalyze the removal of phosphoryl groups.

Most eukaryotic ligases together with the archaeal and bacteriophage ligases are ATP-dependent. The proteins range in size from approximately 100 kDa to about 41 kDa with the T7 phage ligases being the smallest in this group. A novel ATP-dependent T7-phage DNA ligase was cloned from Beatrix mine phage DNA. As expected with all ligases the Beatrix mine ligase also contained the KXDG sequence motif involved in the formation of the adenylated enzyme intermediate. The enzyme was expressed functionally in *E. coli* and was able to ligate sticky ends created with restriction endonucleases recognizing 5' and 3' sites. Blunt ends were ligated in the presence of PEG, activity was not observed in absence of PEG. It was interesting to note that the protein is capable of ligating sticky ends at temperatures as high as 70 °C.

The cloned polynucleotide kinase/phosphatase protein showed similarity to the Pnkp from the T4 phage, however the expressed protein was not functional when tested using *Bam*HI cut pUC19 vector. The protein failed to remove the phosphate group from the cut DNA, furthermore phosphodiester bonds were also not formed suggesting the kinase domain is not functional. Sequence analysis revealed that the cloned protein lacked essential residues for the activity of both domains. The protein had the following substitutions; D35I, R38D and R126E, and R176N. These amino acid residues have been proved to be responsible the activity of the kinase and phosphatase domains when using alanine scan (Wang and Shuman, 2002). The first three mutations are deactivated the N-terminal kinase domain and the C-terminal phospatase was inactivated by R176N substitution.

The *Seg*B homing endonuclease cloned from was successfully expressed in *E. coli*. A smear was obtained when the enzyme was used to cut lambda DNA. Though studies have shown that this family of endonucleases has relaxed recognition sequence, our results indicate that the Beatrix mine endonuclease might recognize a specific site. In addition Brok-Volchanskaya *et al.* (2008) also showed that the T4-phage *Seg*B homing endonuclease recognizes a 27 bp sequence, and longer incubation periods resulted in the protein cutting other sites. In this study lambda DNA was not completely digested after an hour of incubation suggesting that the endonuclease does not cut randomly. The recognition sequence for the expressed protein was however not determined as this will require detailed analysis to determine the recognition sequence. Furthermore the protein contained the YXI-YVG motif instead of the GIY-YIG, on the other hand aligments of the protein from other phages also comprised of YXI-YVG. All the other catalytic residues were present and they were located within the first 90 amino acid residues.

Overall results demonstrate that viral communities are great source of novel proteins including phage communities from Beatrix mine. The proteins were also successfully expressed in *E. coli*.

Chapter 6

Summary

Bacteriophages are viruses that infect both bacteria and archaea, and they are the most abundant microbial communities in the ecosystem. Phages have unique applications as they have the ability to control the mortality rate of the hosts, and they are also useful in molecular biology techniques as a many enzymes that are utilized in this field have a phage origin. Though phages are the most abundant they still the most unexplored, especially from the environment. This is due to the fact that approximately 99% of their bacterial hosts cannot be cultured using the standard techniques and phages require these hosts for propagation and replication. Development of culture-independent techniques has managed to circumvent problems associated with prokaryotic diversity studies by using the 16S rDNA sequencing. Viruses however do not have a common gene or sequence fragment that can be used for phylogeny. Development of the phage proteomic tree has facilitated PCR detection of different families or clades of unculturable phages. In this study tailed phages (T4-like phages and T7-like podoviruses) were targeted because of their abundance in the environment. The major capsid protein (g23) and DNA polymerase fragment were used to detect T4-like phages and T7-like podoviruses, respectively. Transmission electron microscopy was also used to study and determine morphology of phages, though the technique does not give a true reflection of the number of viral like particles from the environment. The methods were first optimized with water and sediments from Loch Logan pond where phage counts were expected to be high. These methods were used preliminarily as a way to check the presence of phages in the samples. Water samples collected from four South African mines, Masimong, Beatrix, Star Diamonds and Tau Tona (level 100 and level 118) were also subjected to these two techniques. Phage particles were only observed with Beatrix mine when using transmission electron microscopy. The samples were further characterized as the TEM requires high viral counts. Uncultured T7-like podoviruses were detected with all mine samples indicating the presence of these groups of phages from the mines, and the T4-like phages only in Beatrix and Tau Tona. Phylogenetic analysis revealed that the DNA polymerase fragment from the T7-like phages is highly conserved. The sequences were similar to the clones obtained with marine, water and soil samples from different locations. In contrast to the T7 phages the T4-like phages were highly diverse with very few clones showing similarity to the known capsid protein.

The main focus of the project however was the genomic analysis of viral communities from deep mines and also to identify novel phage proteins that might have biotechnological applications. Hence shotgun libraries were also constructed to get genomic information of the viral assemblages from the mines. Sequencing revealed untapped viral communities with the majority of the clones showing no similarity to the known proteins. Very few phage proteins were obtained and the data was not enough to identify many the novel genes from the phage communities. This is due to the fact that only 20 clones from each mine were sequenced. Therefore the use of high throughput sequencing technique was necessary to obtain large amount of sequencing data. Biofilm from Beatrix mine was selected for pyrosequencing and 2X quarter runs of the GS FLX were done using isolated viral DNA. The viral communities from Beatrix mine were unique with ~75% of the proteins not showing similarity to any known proteins. Microbial analysis of Beatrix mine revealed that most the diversity from this mine is clustered within the classes *Beta* and *Gamma Proteobacteria* and the phyla *Firmicutes.* Hence the portion of the known phages were all three families of dsDNA phages infecting *Enterobacteria* phages and few of the phage proteins were from different *Bacillus* sp. Proteins from *Acanthamoeba polyphaga mimivirus* were also detected from Beatrix mine. Seven prophages were detected with possible genome sizes ranging from 5 kb to more than 12 kb. These prophages contained high number of hypothetical proteins.

Novel proteins were identified from the Beatrix mine sequencing and three proteins; DNA ligase, *Seg*B homing endonuclease and polynucleotide kinase were selected for expression studies. The ATP-dependent DNA ligase was able to ligate sticky ends at temperatures 4 °C, 16 °C and 22 °C. The enzyme also ligated sticky ends at temperatures as high as 70 °C. Blunt end fragments were also ligated in presence of polyethylene glycol. The expressed polynucleotide kinase had the following substitutions; D35I, R38D and R126E, and R176N, no activity could be detected, possibly a result of the substitutions. The endonuclease digested lambda DNA, and early indications are that that the endonuclease recognizes a specific sequence and does not cut randomly. The overall results shows that uncultured phage communities, including South African phage metagenome are the largest untapped source of genomic information in the biosphere. In addition they are the source of novel biotechnologically important proteins.

Opsomming

Bakteriofage is virusse wat bakterieë en archaea infekteer en hulle verteenwoordig die volopste entiteit in die ekosisteem. Fage het unieke toepassings en hulle beskik oor die vermoë om die mortaliteitstempo van gasheerbevolkings te beheer. Die toepassing van fage en hulle ensieme in molekulêre biologie tegnieke is lank reeds gevestig. Al is fage die volopste in die natuur, is hulle ook die mees onontginde, hoofsaaklik omdat hulle bakteriese gashere onbekend is en ongeveer 99% van alle bakterie nie in kultuur gekweek kan word nie. Kultuur-onafhanklike tegnieke verskaf nou die moontlikheid om hierdie beperkings aan te spreek, soos in die geval van die gebruik van 16S rDNA basisopeenvolgings om bakteriese diversiteit te bestudeer. Virusse beskik egter nie oor 'n algemeen gekonserveerde geen wat vir die doeleindes gebruik kan word nie.

Die ontwikkeling van die faag proteoom boom het gene geïdentifiseer wat gebruik kan word om seker faag families waar te neem. In hierdie studie is fage met sterte (T4-agtige en T7-agtige podofage) ondersoek omdat hulle die volopste in die natuur is. Die hoof kapsiedproteïen (g23) en 'n DNA polimerase fragment is vir die T4-agtige en T7-agtige fage, onderskeidelik gebruik. Transmissie elektronmikroskopie is ook toegepas om die morfologie van die virusse te ondersoek; die tegniek verskaf ongelukkig nie betroubare kwantitatiewe data nie. Metodes is geoptimaliseer met water en sediment van Loch Logan waar faagtellings waarskynlik hoog sou wees. Watermonsters van vier Suid-Afrikaanse myne, Masimong, Beatrix, Star Diamonds en Tau Tona (vlak 100 en vlak 118) is ook bestudeer. Faagpartikels is in Beatrix water waargeneem met transmissie elektronmikroskopie. Ongekweekte T7-agtige podofage is in al die myne waargeneem maar T4-agtige fage net in Beatrix en Tau Tona. Filogenetiese analise het getoon dat die T7-agtige fage hoogs gekonserveerd en byna identies aan die van oseaan, varswater en grondmonsters is. Die T4-agtige fage is daarenteen hoogs divers, baie min klone het ooreenkoms met die bekende kapsiedproteïene getoon.

Die hoof fokus van die projek was egter om 'n ontleding van die virusgemeenskappe in die diep myne op genoomvlak te doen en om nuwe faagproteïene op te spoor wat moontlik nuwe toepassings in biotegnologie kan hê. Genoom biblioteke is daarom saamgestel om verdere inligting oor die virusbevolking in te win. DNA basisopeenvolgingdata het aangedui dat die meeste klone geen ooreenkoms met bekende volgordes getoon het nie. Die data was ook

onvoldoende om beduidende hoeveelhede nuwe gene waar te neem. Dit het die gebruik van hoë deurvloei tegnologie genoodsaak.

Phi 29 geamplifiseerde faag DNA uit biofilm van Beatrixmyn is onderwerp aan volgordebepaling met GS FLX tegnologie. Data is saamgestel en aan die annoteringstelsel van TIGR onderwerp. Die faagbevolkings uit Beatrixmyn is klaarblyklik uniek omdat ongeveer 75% van die proteïene geen ooreenkoms met enige bekendes in databasisse getoon het nie. Mikrobiese analise van die Beatrix mynwater het aangedui dat die meeste divetsiteit in die myn saamgroepeer met die klasse Beta en *Gamma Proteobakterië* en die filum *Firmicutes.* Die bekende komponente van die fag bevolking was dus van al drie families dsDNA fage wat *Enterobakterië infekteer.* Al drie families dsDNA virusse is waargeneem, maar die meeste was van *Enterobacteria* fage. Proteïene van *Acanthamoeba polyphaga mimivirus* is ook waargeneem. Sewe profaag opeenvolgings is waargeneem met groottes wat wissel tussen 5kb en 12kb, elk het 'n groot aantal hipotetiese proteïen bevat.

Nuwe proteïene is in die Beatrixmyn monster waargeneem. Drie proteïene, 'n DNA ligase, *Seg*B homing endonukease en polinukleotied kinase is geselekteer vir uitdrukkingstudies. Die ATP-afhanklike DNA ligase kon klewerige punte by 4 °C, 16 °C en 22 °C ligeer. Die ensiem kon ook klewerige punte by 'n temperatuur van 70 °C ligeer. Stomp punt ligering kon in die teenwoordigheid van poli-etileenglikol plaasvind. Die polinukleotied kinase het die volgende substitusies in aminosuurvolgorde getoon; D35I, R38D, R126E en R176N. Die substitusies kon moontlik verklaar waarom die ensiem nie in aktiewe vorm uitgedruk kon word nie. Die endonuklease kon lambda DNA hidroliseer en vroeë aanduidings is dat die ensiem spesifieke basisopeenvolgings herken.

Die ondersoek bevestig dat faag gemeenskappe 'n groot onontginde bron van genetiese materiaal in die biosfeer verteenwoordig en dat dit as uitstekende bron van nuwe, biotegnologies belangrike proteïene kan dien.

References

Ackermann H W and Dubow M S, (1987). Viruses of Prokaryotes. CRC Press,Boca Raton.

Ackermann H W and Krisch HM, (1997). A catalogue of T4-type bacteriophages. *Archives of Virology*, **142**: 2329–2345.

Ackermann H W, (2003). Bacteriophage observations and evolution. *Research in Microbiology*, **154**: 245–251.

Ackermann H W, (2007). 5500 Phages examined in the electron microscope. *Archives of Virology*, **152**: 227–243.

Ahmadian A, Ehn M and Hober S, (2006). Pyrosequencing: history, biochemistry and future. *Clinica Chimica Acta*, **363**: 83-94.

Akulenko N V, Ivashina T V, Shaloiko L A, Shliapnikov M G and Ksenzenko V N, (2004). Novel site-specific endonucleases F-TfII, F-TfIII and F-TfIIV encoded by the bacteriophage T5. *Journal of Molecular Biology*, **38**: 632–64.

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

Altschul S F, Madden T L, Schäffer A A, Zhang J, Zhang Z, Miller W and Lipman D, (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research,* **2***:* 3389–3402.

Angly F E, Felts B, Breitbart M, Salamon P, Edwards R A, Carlson C, Chan A M, Haynes M, Kelley S, Liu H, Mahaffy J M, Mueller J E, Nulton, Olson J, Parsons R R, Rayhawk S, Suttle C A and Rohwer F, (2006). The Marine Viromes of Four Oceanic Regions. *PLOS Biology*, **4**: 2121-2132.

Angly F, Rodriguez-Brito B, Bangor D, McNairnie P, Breitbart M, Salamon P, Felts B, Nulton J, Mahaffy J and Rohwer F, (2005). *BMC Bioinformatics*, **6**: 41.

Baback G, Mehran G and Pål N, (2007). Pyrosequencing technology for short DNA sequencing and whole genome sequencing. *Seibutsu Butsuri*, **47**:129-132.

Balding C, Bromley S A, Pickup R W and Saunders J R, (2005). Diversity of phage integrases in *Enterobacteriaceae*: development of markers for environmental analysis of temperate phages. *Environmental Microbiology*, **7**: 1558–1567.

Bateman A, Birney E, Cerruti L, Durbin R, Etwiller L, Eddy S R, Griffiths_Jones S, Howe K L, Marshall M and Sonnhammer E L L, (2000). The Pfam protein families database. *Nucleic Acid Research*, **30**: 276-280.

Bau M, Romer R L, Luders V and Beukes N J, (1999). Pb, O and C isotopes in silicified Mooidraai dolomite (Transvaal Supergroup, South Africa): Implications for the composition of the Palcoproterozoic seawater and "dating" the increase of oxygen in the Precambrian atmosphere. *Earth Planet Science Letters*, **174**: 43–57.

Belle A, Landthaler M and Shub D A, (2001). Intronless homing: site-specific endonuclease SegF of bacteriophage T4 mediates localized marker exclusion analogous to homing endonucleases of group I introns. *Genes and development*, **16**: 351–362.

Benhar I, (2001). Biotechnological applications of phage and cell display. *Biotechnology Advances*, **19**: 1-33.

Bennett S, (2004). Solexa Ltd. *Pharmacogenomics*, **5:** 433–438.

Bernhardt T G, Wang I-N, Struck D K and Young R, (2001). A protein antibiotic in the phage Q virion: Diversity in lysis targets. **Science, 292**: 2326 – 2329.

Blanco L, Bernad A, Lázaro J M, Martins G, Garmendia C and Salas M, (1989). Highly efficient DNA synthesis by the phage 429 DNA polymerase, symmetrical mode of DNA replication. *The Journal of Biological Chemistry*, **264**: 8935-8940.

Bonfield J K, Smith K F and Staden R, (1995). A new DNA sequence assembly program. *Nucleic Acid Research*, **23**: 4992-4999.

BØrsheim K Y, Bratbak G and Heldal, (1990). Enumeration and biomass estimation of planktonic bacteria and viruses by transmission electron microscopy. *Applied and Environmental Microbiology*, **56**: 352-356.

Breitbart M and F Rohwer, (2005). A method for discovering novel DNA viruses in blood. *Biotechniques*, **39**: 729-736.

Breitbart M and Rower F, (2005). Here a virus, there a virus, everywhere the same virus. *TRENDS in Microbiology*, **13**: 278-284.

Breitbart M, Hayes M, Kelley S, Angly F, Edwards R A, Felts B, Mahaffy J M, Mueller J, Nulton J, Rayhawk S, Rodriguez-Brito B, Salamon P and Rohwer F, (2008). Viral diversity and dynamics in an infant gut. *Reseaech in Microbiology*, **159**: 367-373.

Breitbart M, Hewson I, Felts B, Mahaffy J M, Nulton J, Salamon P and Rohwer F, (2003). Metagenomic analyses of an uncultured viral community from human feces. *Journal of Bacteriology*, **185**: 6220-6223.

Breitbart M, Miyake J H and Rohwer F, (2004). Global distribution of nearly identical phage-encoded DNA sequences. *FEMS Microbiology Letters*, **236**: 249-256.

Breitbart M, Salamon P, Andresen B, Mahaffy J M, Segall A M, Mead D, Azam F and Rohwer F, (2002). Genomic analysis of uncultured marine viral communities. *Proceedings of the National Academy of Sciences (USA)*, **99**: 14250–14255.

Brok-Volchanskaya V S, Kadyrov F A, Sivogrivov D E, Kolosov P M, Sokolov A S, Shlyapnikov M G, Kryukov V M and Granovsky I E, (2008). Phage T4 SegB protein is a

homing endonuclease required for the preferred inheritance of T4 tRNA gene region occurring in co-infection with a related phage. *Nucleic Acid Research*, **29**: 1-12.

Brussaard C P D, (2004). Optimization of procedures for counting ciruses by Flow Cytometry. *Applied and Environmental Microbiology*, **70**: 1506–1513.

Brussow H and Hendrix R W, (2002). Phage Genomics: Small Is Beautiful. *Cell*, **108**: 13–16.

Casas V and Rohwer F, (2007). Phage metagenomics. *Methods in enzymology*, **421**: 259-68.

Chauthaiwale V M, Therwath A and Vasanti V D, (1992). Bacteriophage lambda as a cloning vector. *Microbiological Reviews*, **56**: 577-591.

Chen F, Lu J, Binder B J, Liu Y and Hodson R E, (2001). Application of digital image analysis and flow cytometry to enumerate marine viruses stained with SYBR gold. Applied and *Environmental Microbiology*, **67**: 539-545.

Chen K and Pachter L, (2005). Bioinformatics for whole-genome shotgun sequencing of microbial communities. *PLoS Computational Biology*, **1**: e24.

Cherepanov A V and de Vries S, (2001). Binding of nucleotides by T4 DNA Ligase and T4 RNA ligase: optical absorbance and fluorescence studies. *Biophysical Journal Volume*, **81**: 3545–3559.

Chevalier B S and Stoddard B L, (2001). Homing endonucleases: structural and functional insight into the catalysts of intron/intein mobility. *Nucleic Acid Research*, **29**: 3757-3774.

Clark J R and March J B, (2006). Bacteriophages and biotechnology: vaccines, gene therapy and antibacterials. *Trends in Biotechnology*, **24**: 212-218.

Click E M and Webster R E, (1997). Filamentous Phage Infection: Required Interactions with the TolA Protein. *Journal of bacteriology*, **179**: 6464-6471.

Comeau A M and Krisch H M, (2008). The capsid of the T4 Phage superfamily: The evolution, diversity, and structure of some of the most prevalent proteins in the biosphere. *Molecular Biology and Evolution*, **25**: 1321–1332.

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

Delcher A L, Harmon D, Kasif S, White O and Salzberg S L, (1999). Improved microbial gene identification with Glimmer. *Nucleic Acid Research*, **27**: 4636-4641.

Demuth J, Neve H and Witzel K, (1993). Direct electron microscopy study on the morphological diversity of bacteriophage populations in Lake Pluβsee. *Applied and Environmental Microbiology*, **59**: 3378-3384.

Dinsdale E A, Edwards R A, Hall D, Angly F, Breitbart M, Brulc J M, Furlan M, Desnues C, Haynes M, Li L, McDaniel L, Moran M A, Nelson K E, Nilsson C, Olson R, Paul J, B R Brito, Ruan Y, Swan B K, Stevens R, Valentine D L, Thurber R V, Wegley L, White B A and Rohwer F, (2008). Functional metagenomic profiling of nine biomes. *Nature*, **455**: 629-632.

Doherty A J and Suh S W, (2000). Structural and mechanistic conservation in DNA ligases. *Nucleic Acid Research*, **21**: 4051-4058.

Doherty A J, Ashford S R, H Subramanya S and Wigley D B, (1996). Bacteriophage T7 DNA Ligase: Overexpression, purification, crystallization and characterization. *The Journal of Biological Chemistry*, **271**: 11083-11089.

Dorigo U, Jacquet S and Humbert J, (2004). Cyanophage Diversity, Inferred from *g20* Gene Analyses, in the Largest Natural Lake in France, Lake Bourget. Applied *and Environmental Microbiology*, **70**: 1017–1022.

Díez B, Pedrós-Alió C, Marsh T L and Massana R, (2001). Application of denaturing gradient gel electrophoresis (DGGE) to study the diversity of marine picoeukaryotic assemblages and comparison of DGGE with other molecular techniques. *Applied and Environmental Microbiology*, **67**: 2942–2951.

Eddy S R, (1998). Profile hidden markov models. *Bioinformatics,* **14**: 755-763.

Edgell D R., Belfort M and Shub D A, (2000). Barriers to intron promiscuity in bacteria. *Journal of Bacteriology*, **182**: 5281–5289.

Edwards R A and Rohwer F, (2005). Viral metagenomics. *Nature Reviews*, **3**: 504-510.

Edwards R A, Rodriguez-Brito B, Wegley L, Haynes M, Breitbart M, Peterson D M, M O Saar, S Alexander, E C A Jr and Forest Rohwer, (2006). Using pyrosequencing to shed light on deep mine microbial ecology. *BMC Genomics*, **7**: 57.

Edwards R, (2006). Random Community Genomics. Open paper

Filée J, Bapteste E, Susko E and Krisch H. M, (2006). A selective barrier to horizontal gene transfer in the T4-Type bacteriophages that has preserved a core genome with the viral replication and structural genes. Molecular Biology and Evolution, 23:1688-1696.

Filée J, Tétart F, Suttle C A and Krisch H M, (2005). Marine T4-Type bacteriophages, a ubiquitous component of the dark matter of the biosphere. *Proceedings of the National Academy of Sciences (USA)*, **102**: 12471-12476.

Frenkel D and Solomon B, (2002). Filamentous phage as vector-mediated antibody delivery to the brain. *Proceedings of the National Academy of Sciences (USA)*, **99**: 5675-5679.

Galburt E A and Stoddard B L, (2002). Catalytic mechanisms of restriction and homing endonucleases. *Biochemistry*, **41**: 13851-13860.

Gentile M and Gelderblom H R, (2005). Rapid viral diagnosis: role of electron microscopy. The *New Microbiologica*, **28**: 1-12.

Gillespie D E, Brady S F, Bettermann A D, Cianciotto N P, Liles M R, Rondon M R, Clardy J, Goodman R M and Handelsman J, (2002). Isolation of Antibiotics Turbomycin A and B from a metagenomic library of soil microbial DNA. *Applied and Environmental Microbiology*, **68**: 4301–4306.

Grabow W O K, (2001). Bacteriophages: update on application as models for viruses in water. *Water SA*, **27**: 251-268.

GS FLX Titanium series reagents: New reagents for today's Genome Sequencer FLX instrument. Roche Applied Science, (2008).

Gunther W, Loessner M J and Scherer S, (1996). Bacteriophage receptors on Listeria monocytogenes cells are the N-acetylglucosamine and rhamnose substituents of teichoic acids or the peptidoglycan itself. *Microbiology*, **142**: 985-992.

Haft D H, Loftus B J, Richardson D L, Yang F, Eisen J A, Paulsen I T and White O (2001). TIGRFAMs: a protein family resource for the functional identification of proteins. *Nucleic Acids Research*, **29**: 41–43.

Haft D H, Selengut J D and White O, (2003). The TIGRFAMs database of protein families. . *Nucleic Acids Research,* **31**: 371-373.

Hagens S and Loessner M J, (2007). Application of bacteriophages for detection and control of foodborne pathogens. *Applied Microbiology and Biotechnology*, **76**: 513–519.

Hambly E and Suttle C A, (2005). The viriosphere, diversity, and genetic exchange within phage communities. *Current Opinion in Microbiology*, **8**: 444-450.

Handelsman J, (2004). Metagenomics: Application of genomics to uncultured microorganisms. *Microbiology and Molecular Biology Reviews*, **68**: 669–685.

Hendrix R W, Smith M C, Burns R N, Ford M E and Hatfull G, (1999). Evolutionary relationships among diverse bacteriophages and prophages: all the world's a phage. *Proceedings of the National Academy of Sciences (USA)*, **96**: 2192-2197.

Henne A, Daniel R, Schmitz R A and Gottschalk G, (1999). Construction of environmental DNA libraries in *Escherichia coli* and screening for the presence of genes conferring utilization of 4-hydroxybutyrate. *Applied and Environmental Microbiology*, **65**: 3901-3907.

Hsiao W W L, Ung K, Aeschliman D, Bryan J, Finlay B B and Brinkman F S L, (2005). Evidence of a large novel gene pool associated with prokaryotic genomic islands. *PLoS Genetics*, 1:e62.

Hughes J B, Hellmann J J, Ricketts T H and Bohannan, (2001). Counting the uncountable: Statistical approaches to estimating microbial diversity. *Applied and Environmental Microbiology*, **67**: 4399–4406.

Huse S M, Huber J A, Morrison H G, Sogin M L and Welch D M, (2007). Accuracy and quality of massively parallel DNA pyrosequencing. *Genome Biology*, **8**: R143.

Huson D H, Auch A F, Qi J and Schuster S C, (2007). MEGAN analysis of metagenomic data. *Genome Research*, **17**: 377-386.

Häring M, Peng X, Brügger K, Rachel R, Stetter K O, Garrett R A and Prangishvili D, (2004). Morphology and genome organization of the virus PSV of the hyperthermophilic archaeal genera *Pyrobaculum* and *Thermoproteus*: a novel virus family, the *Globuloviridae*. *Virology*, **323**: 233–242.

Häring M, Rachel R, Peng X, Garrett R A and Prangishvili D, (2005). Viral diversity in hot springs of Pozzuoli, Italy, and characterization of a unique bottle-shaped archaeal virus, *Acidicanus* bottle-shaped virus, from a new family, the *Ampullaviridae*. *Journal of Virology*, **79**: 9904–9911.

Iacono M, Villa L, Fortini D, Bordoni R, Imperi F, Bonnal R J P, Sicheritz-Ponten T De Bellis G, Visca P, Cassone A and Carattoli A, (2008). Whole-Genome Pyrosequencing of an Epidemic Multidrug-Resistant *Acinetobacter baumannii* Strain Belonging to the European Clone II Group. *Antimicrobial agents and chemotherapy*, **52**: 2616–2625.

Jeon S-J and Ishikawa K, (2003). A novel ADP-dependent DNA ligase from *Aeropyrum pernix* K1. *FEBS Letters*, **550**: 69-73.

Jia Z, Ishihara R, Nakajima Y, Asakawa S and Kimura M, (2007). Molecular characterization of T4-type bacteriophages in a rice field. *Environmental Microbiology*. **9**: 1091–1096.

Karlin S and Burge C, (1995). Dinucleotide relative abundance extremes: a genomic signature. *Trends in Genetics*, **11**: 283-290.

Kemp P F and Aller J Y, (2004). Bacterial diversity in aquatic and other environments: what 16S rDNA libraries can tell us. *FEMS Microbiology Ecology*, **47**: 161-177.

Kimura N, (2006). Metagenomics: Access to unculturable microbes in the environment. *Microbes and Environments*, **21**: 201-215.

Kletzin A, (1992). Molecular characterisation of a DNA ligase gene of the extremely thermophilic archaeon *Desulfurolobus ambivalens* shows close phylogenetic relationship to eukaryotic ligases. *Nucleic Acids Research,* **20**: 5389-5396.

Kowalski J C, Belfort M, Stapleton M A, Holpert M, Dansereau J T, Pietrokovski S, Baxter S M and Derbyshire V, (1999). Configuration of the catalytic GIY-YIG domain on intron endonuclease I-*Tevl*: coincidence of computational and molecular findings. *Nucleic Acid Reaserch*, **27**: 2115-2125.

Kühlmann U C, James GR and Hemmings A M, (1999). Structural parismony in endonuclease active sites: Should the number of homing endonuclease families be refined? *FEBS Letters,* **463:** 1–2.

Lane D J, (1991). 16S/23S rRNA sequencing in: Nucleic Acid techniques in bacterial systematics (Stackebrandt, E and Goodfellow, M. Eds), pp. 115-175. Wiley, Chichester, UK.

Lane D J, Pace B, Olson G J, Stahl D A, Sogin M L and Pace N R, (1985). Rapid determination of 16S ribosomal RNA sequences for phylogenetic analyses. *Proceedings of the National Academy of Sciences (USA)*, **82**: 6955–6959.

Lee S, Won K, Lim H E, Kim J, Choi G J and K Y Cho, (2004). Screening for novel lipolytic enzymes from uncultured soil microorganisms. *Applied and Microbiology Biotechnology*, **65**: 720–726.

Lehman I R, (1974). DNA ligase: structure, mechanism, and function. *Science*, **186**: 790–797.

Leplae R, Hebrant A, Wodak S J and Toussaint A, (2004). ACLAME: A CLAssification of Mobile genetic Elements. *Nucleic Acids Research*, 32(Database issue): D45-49.

Lewin B, (1997). *Genes VI*. Oxford University Press.

Lima-Mendez G, J Van Helden, A Toussaint and R I Leplae, (2008). Prophinder: a computational tool for prophage prediction in prokaryotic genomes. *Bioinformatics Applications Note*, **24**: 863-865.

Lima-Mendez G, Toussaint A and Leplae R, (2007). Analysis of the phage sequence space: The benefit of structured information. *Virology*, **365**: 241-249.

Lovmar L and Syvänen A, (2006). Multiple displacement amplification to create a long-lasting source of DNA for genetic studies. *Human Mutation*, **27**: 603-614.

Maniloff J, (1995). Identification and classification of viruses that have not been propagated. *Archives in Virology*, **140**: 1515-1520.

Marie D, Brussaard C, Thyrhaug R, Bratbak, G and Vaulot D, (1999). Enumeration of marine viruses in culture and natural samples by flow cytometry. *Applied and Environronmental Microbiology*, **65**: 45–52.

Martin I V and McNeill S A, (2002). ATP-dependent DNA ligases. *Genome Biology*, **3**: 3005.1-3005.7.

Marvin D A, (1998). Filamentous phage structure, infection and assembly. *Current Opinion in Structural Biology*, **8**: 150-158.

Murray N E and Murray K, (1974). Manipulation of restriction targets in phage lambda to form receptor chromosomes for DNA fragments. *Nature*, **251**: 476-481.

Ng T F F, Manire C, Borrowman K, Langer T, Ehrhart L and Breitbart M, (2009). Discovery of a novel single-stranded DNA virus from a sea Turtle Fibropapilloma using viral metagenomics. *Journal of Virology*, **83**: 2500–2509.

Nicomrat D, Dick W A and Tuovinen O H, (2006). Assessment of the microbial community in a constructed wetland that receives acid coal mine drainage. *Microbial Ecology*, **51**: 83-89.

Noble R T and Fuhrman J A, (1998). Use of SYBR Green I for rapid epifluorescence counts of marine viruses and bacteria. *Aquatic Microbial Ecology*, **14**: 113-118.

Ochsenreiter T, Selezi D,Quaiser A, Bonch-Osmolovskaya L and Schleper C, (2003). Diversity and abundance of Crenarchaeota in terrestrial habitats studied by 16S RNA surveys and real time PCR. *Environmental Microbiology*, **5**: 787-797.

Onstott T C, Lin L H, Davidson M, Mislowack B, Borcsik M, Hall J, Slater G, Ward J, Lollar B S, Lippmann-Pipke J, Boice E, Pratt L M, Pfiffner S, Moser D, Gihring T, Kieft T L, Phelps T J, van Heerden E, Litthauer D, Deflaun M, Rothmel R, Wanger G and G Southam, (2006). The Origin and Age of Biogeochemical Trends in Deep Fracture Water of the Witwatersrand Basin, South Africa. *Geomicrobiology Journal*, **6**: 369-414

Onstott T C, Moser D P, Pfiffner S M, Fredrickson J K, Brockman F J, Phelps T J, White D C, Peacock A, Balkwill D, Hoover R, Krumholz L R, Borscik M, Kieft T L and Wilson R, (2003). Indigenous and contaminant microbes in ultradeep mines. *Environmental microbiology*, **5**: 1168-1191.

Onstott T C, Tobina K, Dong H, DeFlaun M F, Fredrickson J K, Bailey T, Brockman F, Kieft T, Peacock A, White D C, Balkwill D, Phelps T J and Boone D R, (1997). The deep gold mines of South Africa: Windows into the subsurface biosphere. **Proc. SPIE.**, **3111**: 344-357.

Paul JH, Jiang SC and Rose JB, (1991). Concentration of viruses and dissolved DNA from aquatic environments by vortex flow filtration. *Applied and Environronmental Microbiology*, **57**: 2197–2204.

Pfiffner S M, Cantu J M, Smithgall A, Peacock A D, White D C, Moser D P, Onstott T C and van Heerden E, (2006). Deep Subsurface Microbial Biomass and Community Structure in Witwatersrand Basin Mines. Geomoicrobiology Journal, 6: 431 – 442.

Pheiffer B H and Zimmerman S B, (1983). Polymer-stimulated ligation: enhanced blunt- or cohesive-end ligation of DNA or deoxyribooligonudeotides by T4 DNA ligase in polymer solutions. *Nucleic Acids Research*, **11**: 7853- 7871.

Proctor L M and Furhman J A, (1990). Viral mortality of marine bacteria and cyanobacteria. *Nature*, **343**: 60-62.

Quinto M and Bender R A, (1984). Use of bacteriophage P1 as a vector for Tn5 insertion mutagenesis. *Applied and Environ Microbiology*, 47: 436-438.

Rachel R, Bettstetter M, Hedlund B P, H□ring M, Kessler A, Stetter K O and Prangishvili D, (2002). Remarkable diversity of viruses and virus-like particles in hot terrestrial environments. *Archives in Virology*, **147**: 2419–2429.

Reysenbach, A L and Pace, N R, (1995). In: Robb, F.T., Place, A.R. (Eds.), Archaea: A Laboratory Manual—Thermophiles. Cold Spring Harbour Laboratory Press, New York, pp. 101–107.

Riesenfeld C S, Schloss P D and Handelsman J, (2004). METAGENOMICS: Genomic Analysis of Microbial Communities. *Annual Reviews in Genetics*, **38**: 525–552.

Rincón B, Raposo F, Borja R, Gonzalez J M, Portillo M C and Saiz-Jimenez C, (2006). Performance and microbial communities of a continuous stirred tank anaerobic reactor treating two-phases olive mill solid wastes at low organic loading rates. *Journal of Biotechnology*, **121**: 534–543.

Roesch L F W, Fulthorpe R R, Riva A, Casella G, Hadwin A K M, Kent A D, Daroub S H, Camargo F A O, Farmerie W and Triplett E W, (2007). Pyrosequencing enumerates and contrasts soil microbial diversity. *The ISME Journal*, **1**: 283–290.

Rohwer F and Edwards, (2002). The Phage Proteomic Tree: a genome-based taxonomy for phage. *Journal of Bacteriology*, **184**: 4529-35.

Rohwer F, (2003). Global Phage Diversity. *Cell*, **113**:141.

Rohwer F, (2005). Tangential Flow Filtration. Rohwer lab manual.

Ronaghi M, (2001). Pyrosequencing sheds light on DNA sequencing. *Genome Research*, **11**: 3-11.

Rondon M R, August P R, Bettermann A D, Brady S F, Grossman T H, Liles M R, Loiacono K A, Lynch B A, MacNeil I A, Osburne M S, Clardy J, Handelsman J and Goodman R M, (2000). Cloning the soil metagenome: a strategy for accessing the genetic and functional diversity of uncultured microorganisms. *Applied and Environmental Microbiology*, **66**: 2541–2547.

Rusch D B, Halpern A L, Sutton G, Heidelberg K B, Williamson S, Yooseph S, Wu D, Eisen J A, Hoffman J M, Remington K, Beeson K, Tran B, Smith H, Baden-Tillson H, Stewart C, Thorpe J, Freeman J, Andrews-Pfannkoch C, Venter J E, Li K, Kravitz S, Heidelberg J F, Utterback T, Rogers Y, Falcón L I, Souza V, Bonilla-Rosso G, Eguiarte L E, Karl D M, Sathyendranath S, Platt T, Bermingham E, Gallardo V, Tamayo-Castillo G, Ferrari M R, Strausberg R L, Nealson K, Friedman R, Frazier M and J Craig Venter, (2007). The Sorcerer II Global Ocean Sampling expedition: Northwest Atlantic through eastern tropical Pacific. *PLoS Biology*, **5**: e77.

Russel M, (1995). Moving through the membrane with filamentous phages. *Trends in Microbiology*, **3**: 223-228.

Sahl J W, Schmidt R, Swanner D E, Mandernack K W, Templeton A S, Kieft T L, Smith R L, Sanford W E, Callaghan R L, Mitton J B and Spear J R, (2008). Subsurface Microbial Diversity in Deep-Granitic-Fracture Water in Colorado. *Applied and Environmental Microbiology*, **74**: 143–152.

Sambrook J and Russell D W, (2001). Molecular cloning: a laboratory manual, 2rd Ed., Cold Spring Harbor Laboratory Press. Cold Spring Harbor, NY.

Sambrook J, Fritsch E F and Maniatis T, (1989). Molecular cloning: a laboratory manual, 2nd Edn., Cold Spring Harbor. NY.

Sanger F, Nicklen S and Coulson A R, (1977). DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences (USA),* **74:** 5463–5467.

Sano E, Carlson S, Wegley L and Rohwer F, (2004). Movement of Viruses between Biomes. *Applied and Environmental Microbiology*, **70**: 5842–5846.

Schloss P D and Handelsman J, (2003). Biotechnological prospects from metagenomics. *Current Opinion in Biotechnology*, **14**: 303–310.

Sidhu S S, (2000). Phage display in pharmaceutical biotechnology. *Current Opinion in Biotechnology*, **11**: 610–616.

Smith T F and Waterman M S, (1981). Identification of common molecular subsequencees. *Journal of Molecular Biology*, **147**: 195-197.

Spinelli S, Campanacci V, Blangy S, Moineau S, Tegoni M and Cambillau C, (2006). Modular structure of the receptor binding proteins of *Lactococcus lactis* Phages: The RBP structure of the temperate phage TP901-1. *The journal of biological chemistry*, **281**: 14256–14262.

Steward G F, Montiel J L and Azam F, (2000). Genome size distributions indicate variability and similarities among marine viral assemblages from diverse environments. **Limnology and Oceanography**, **45**: 1697-1706.

Stoddard B L, (2006). Homing endonuclease structure and function. *Quarterly Reviews of Biophysics*, **38**: 49–95.

Stummeyer K, Schwarze D, Claus H, Vogel U, Gerardy-Schahn R and Muhlenhoff M, (2006). Evolution of bacteriophages infecting encapsulated bacteria: lessons from *Escherichia coli* K1-specific phages. *Molecular Microbiology*, **60**: 1123–1135.

Takai K, Moser D P, DeFlaun M, Onstott T C and Fredrickson J K, (2001). Archaeal diversity in waters from deep South African gold mines. *Applied and Environmental Microbiology*, **67**: 5750-5760.

Tamura K, Dudley J, Nei M & Kumar S, (2007). MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. *Molecular Biology and Evolution*, **24**: 1596-1599.

Tamura K, Nei M and Kumar S, (2004). Prospects for inferring very large phylogenies by using the neighbor-joining method. *Proceedings of the National Academy of Sciences (USA)* **101**: 11030-11035.

Tétart F, Repoila F, Monod C and Krisch H M, (1996). Bacteriophage T4 Host Range is expanded by Duplications of a Small Domain of the Tail Fiber Adhesin. *Journal of Molecular Biology*, **258**: 726–731.

Timson D J, Singleton M R and Wigley D B, (2000). DNA ligases in the repair and replication of DNA. *Mutation Research*, **460**: 301–318.

Toussaint A, Lima-Mendez G and Leplae R, (2007). PhiGO, a phage ontology associated with the ACLAME database. *Research in Microbiology*, **58**: 567-571.

Tyson G W, Chapman J, Hugenholtz P, Allen E E, Ram R J, Richardson P M, Solovyev V V, Rubin E M, Rokhsar D S and Banfield J F, (2004). Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature*, **428**: 25-26.

Venter J C, Remington K, Heidelberg J F, Halpern A L, Rusch D, Eisen J A, Wu D, Paulsen I, Nelson K E, Nelson W, Fouts D E, Levy S, Knap A H, Lomas M W, Nealson K, White O, Peterson J, Hoffman J, Parsons R, Baden-Tillson H, Pfannkoch C, Rogers Y and Smith H O, (2004). Environmental genome shotgun sequencing of the Sargasso Sea. *Science*, **304**: 66-74.

Vos AT and Roos JC, (2005). Causes and consequences of algal blooms in Loch Logan, an urban impoundment. *Water SA*, **31**: 385-392.

Wang G, Hayashi M, Saito M, Tsuchiya K, Asakawa S and Kimura M, (2009a). Survey of major capsid genes (g23) of T4-type bacteriophages in Japanese paddy field soils. *Soil Biology and Biochemistry*, **41**:13–20.

Wang G, Jin J, Asakawa S and Kimura M, (2009b). Survey of major capsid genes (g23) of T4-type bacteriophages in rice fields in Northeast China. *Soil Biology and Biochemistry*, **41**: 423-427.

Wang L K and Shuman S, (2001). Domain structure and mutational analysis of T4 polynucleotide kinase. *The Journal of Biological Chemistry*, **276**: 26868–26874.

Wang L K and Shuman S, (2002). Mutational analysis defines the 5'-kinase and 3'-phosphatase active sites of T4 polynucleotide kinase. *Nucleic Acid Research*, **30**: 1073-1080.

Warren R L, Sutton G G, Jones S J M and Holt R A, (2008). Assembling millions of short DNA sequences using SSAKE. *Bioinformatics Applications Note*, **23**: 500-501.

Wegley L, Edwards R, Rodriquez-Brito B, Liu H and Rohwer F, (2007). Metagenomic analysis of the microbial communities associated with the coral Porites astreoides. *Environmental Microbiology*, **9**: 2707-2719.

Weinbauer M G, (2004). Ecology of prokaryotic viruses. *FEMS Microbiology Reviews*, **28**: 127-181.

Wen K, Ortmann A C and Suttle C A, (2004). Accurate estimation of viral abundance by epifluorescence microscopy. *Applied and Environmental Microbiology*, **70**: 3862-3867.

Wendlinger G, Loessner M J and Scherer S, (2002). Bacteriophage receptors on *Listeria monocytogenes* cells are the *N*-acetylglucosamine and rhamnose substituents of teichoic acids or the peptidoglycan itself. *Microbiology,* **142**: 985-992.

Westwater C, Kasman L M, Schofield D A, Werner P A, Dolan J W, Schmidt M G, and Norris J S, (2003). Use of genetically engineered phage to deliver antimicrobial agents to

bacteria: an alternative therapy for treatment of bacterial infections. *Antimicrobial Agents and Chemotherapy*, **47**:1301–1307.

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

Williamson K E, Radosevich M and Wommack K E, (2005). Abundance and diversity of viruses in six Delaware soils. *Applied Environmental Microbiology*, **71**: 3119-3125.

Williamson S J, Rusch D B, Yooseph S, Halpern A L, Heidelberg K B, Glass J I, Andrews-Pfannkoch C, Fadrosh D, Miller C S, Sutton G, Frazier M and Venter J C, (2008). The Sorcerer II Global Ocean Sampling Expedition: Metagenomic Characterization of Viruses within Aquatic Microbial Samples. *PLoS ONE*, **3**: e1456

Wommack K E, Ravel J, Hill R T, Chun J and Colwell R R, (1999). Population Dynamics of Chesapeake Bay Virioplankton: Total-Community Analysis by Pulsed-Field Gel Electrophoresis. *Applied and Environmental Microbiology*, **65**: 231–240.

Wommack KE and Colwell RR, (2000). Virioplankton: viruses in aquatic ecosystems. *Microbiology and Molecular Biology Reviews*, **64**: 69-114.

Yang Q and Catalano C E, (2003). Biochemical characterization of bacteriophage genome packaging in vitro. *Virology*, **305**: 276-287.

Yooseph S, Sutton G, Rusch D B, Halpern A L, Williamson S J, Remington K, Eisen J A, Heidelberg K B, Manning G, Li W, Jaroszewsk L i, Cieplak P, Miller C S, Li H, Mashiyama S T, Joachimiak M P, van Belle C, Chandonia J, Soergel D A, Zhai Y, Natarajan K, Lee S, Raphael B J, Bafna V, Friedman R, Brenner S E, Godzik A, Eisenberg D, Dixon J E, Taylor S S, Strausberg R L, Frazier M and Venter C, (2008). The Sorcerer II Global Ocean Sampling Expedition: Expanding the Universe of Protein Families. *PLoS Biology*, **5**: e16.

Young R, (1992). Bacteriophage lysis: mechanism and regulation. *Microbiological Reviews*, **56**: 430-481.

Zhang T, Breitbart M, Lee W H, Run J, Wei C L, Soh S W L, Hibberd M L, Liu E T, Rohwer F and Ruan Y, (2006). RNA viral community in human feces: Prevalence of plant pathogenic viruses. *PLoS Biology,* **4**: e3.

Zhong Y, Chen F, Wilhelm S W, Poorvin L and Hodson R E, (2002). Phylogenetic diversity of marine cyanophage isolates and natural virus communities as revealed by sequences of viral capsid assembly protein gene *g20*. *Applied and Environmental Microbiology*, **8**: 1576:1584.

Zhu H, S Yin and Shuman S, (2007). Characterization of Polynucleotide Kinase/Phosphatase Enzymes from Mycobacteriophages Omega and Cjw1 and Vibriophage KVP40. *The Journal of Biological Chemistry*, **279**: 26358-26369.

Zhu H, Smith P, Wang L K and Shuman S, (2007). Structure–function analysis of the 3′ phosphatase component of T4 polynucleotide kinase/phosphatase. *Virology*, **366**: 126-136.

Appendix A

Proteins contained in each prophage region

Prophage 1

DNA polymerase III, subunits gamma and tau

One conserved hypothetical protein

Three hypothetical proteins

Metal dependent phosphohydrolase

Prophage 2

Four conserved hypothetical proteins

Six hypothetical proteins

Exonuclease

Intein N-splicing region domain protein

Replicative DNA helicase DnaB, intein containing

Deoxyuridine 5-triphosphate nucleotidohydrolase

DNA polymerase I, thermostable (Tth polymerase I), putative

Prophage 3

Intron encoded BmoI, putative

Six hypothetical proteins

Phage tail tape measure protein, family putative

Membrane protein, putative

Prophage 4

BNR /Asp-Box repeat domain protein

Three hypothetical proteins

Phage tail tape measure protein, family, core region

Prophage 5

Chaperonin GroEL

Seven hypothetical proteins

One conserved hypothetical protein

Intron encoded Bmol, putative

Major capsid protein g23

Putative protein

Polypeptide deformylase

Prophage 6

Intron associated endonuclease I (I-*Tev*I) (IRF protein)

Seven hypothetical proteins

Group I intron GIY-YIG endonuclease

Major facilitator superfamily MFS I, putative

TnsA endonuclease N family

Intron encoded Bmol

One conserved hypothetical protein

Prophage 7

Three hypothetical proteins

gp18

Appendix B

Phage proteins obtained with TIGR annotation of Beatrix viral metagenome

Gene id	gene name
ORF00034	phage portal protein, putative
ORF00058	HNH endonuclease
ORF00117	DNA helicase
ORF00150	ATP-dependent DNA ligase
ORF00156	hypothetical prophage lsa1protein
ORF00180	putative replicative DNA helicase
ORF00191	RNA polymerase sigma factor RpoD
ORF00198	dna-directed rna polymerase subunit beta (rnap subunitbeta) (transcriptase subunit beta) (rna polymerase subunit beta), putative
ORF00203	DNA polymerase III subunit alpha
ORF00272	HNH endonuclease, putative
ORF00285	phage major capsid protein
ORF00288	DNA polymerase III subunit alpha
ORF00289	DNA polymerase III subunit alpha
ORF00300	phage portal protein
ORF00302	DNA polymerase III subunit epsilon, putative
ORF00306	putative phage tail protein
ORF00331	H-N-H endonuclease F-TflIV
ORF00406	putative bacteriophage protein
ORF00416	exonuclease SbcC
ORF00419	DNA polymerase
ORF00426	phage portal protein
ORF00429	phage conserved protein
ORF00458	uv-damage endonuclease (uvde)
ORF00470	endonuclease III, putative
ORF00496	DNA ligase, ATP-dependent
ORF00501	hybrid signal transduction histidine kinase I, putative
ORF00524	putative prophage protein
ORF00552	replicative DNA helicase
ORF00571	PseT polynucleotide 5'-kinase and 3'-phosphatase
ORF00596	prophage MuMc02, head decoration protein, putative

ORF00633	putative endonuclease SegE (Endodeoxyribonuclease segE)
ORF00694	group I intron endonuclease subfamily
ORF00760	DNA polymerase I (POL I), putative
ORF00773	group I intron GIY-YIG endonuclease
ORF00778	crossover junction endodeoxyribonuclease RuvC (Hollidayjunction nuclease ruvC) (Holliday junction resolvase ruvC)
ORF00781	phage DNA modification methylase
ORF00789	polynucleotide kinase/ligase, putative
ORF00795	putative SegB homing endonuclease
ORF00796	phage major capsid protein, HK97 family
ORF00797	phage major capsid protein, HK97 family
ORF00798	phage major capsid protein, HK97
ORF00802	phage endonuclease protein
ORF00802	phage endonuclease protein
ORF00807	group I intron GIY-YIG endonuclease
ORF00822	exonuclease SbcD, putative
ORF00883	deoxyribonuclease
ORF00923	HNH endonuclease
ORF00929	probable Inorganic polyphosphate/ATP-NAD kinase
ORF00943	group I intron GIY-YIG endonuclease
ORF00959	HNH endonuclease
ORF01003	phage major capsid protein, HK97 family
ORF01025	helicase conserved C- domain protein
ORF01073	prophage Lp3 protein 18, putative
ORF01098	putative endonuclease
ORF01108	cytidylate kinase (CK) (Cytidine monophosphate kinase)(CMP kinase), putative
ORF01136	replication-associated protein (ATP-dependent helicase Rep) (RepP), putative
ORF01153	phage head-tail adaptor, putative
ORF01209	HNH endonuclease domain protein
ORF01246	MobE homing endonuclease
ORF01250	group I intron endonuclease subfamily
ORF01317	phage head-tail adaptor, putative
ORF01328	phage tail tape measure protein lambda, putative
ORF01362	RNA polymerase sigma factor
ORF01376	DNA primase-helicase subunit, putative

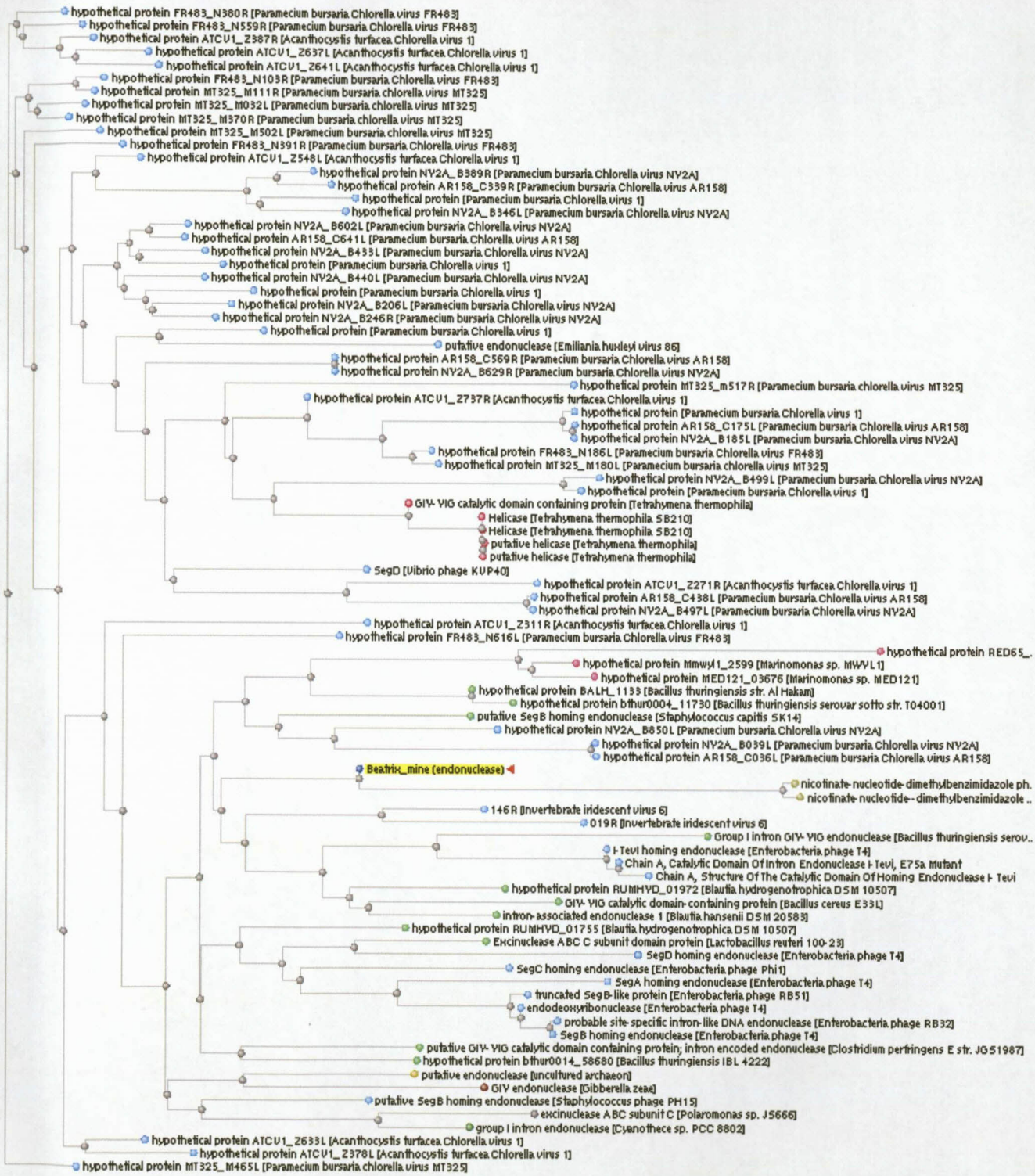
ORF01379	intron-associated endonuclease 1 (I-TevI) (IRF protein), putative
ORF01394	bacteriophage replication gene A protein (GPA)
ORF01411	H-N-H endonuclease F-TfllV, putative
ORF01437	replicative DNA helicase
ORF01475	phage Gp37Gp68
ORF01476	phage protein Gp37/Gp68
ORF01537	terminase large subunit
ORF01537	terminase large subunit
ORF01539	terminase large subunit
ORF01539	terminase large subunit
ORF01595	DNA polymerase I
ORF01676	phage conserved protein
ORF01705	group I intron GIY-YIG endonuclease
ORF01723	intron-associated endonuclease 1 (I-TevI) (IRF protein)
ORF01736	intron-associated endonuclease 1 (I-TevI) (IRF protein)
ORF01740	DNA ligase (Polydeoxyribonucleotide synthase [NAD+])
ORF01743	Eco57I restriction endonuclease domain protein
ORF01746	intron-associated endonuclease 1 (I-TevI) (IRF protein)
ORF01748	putative SegB homing endonuclease
ORF01781	phage Tail Collar Domain family
ORF01798	phage integrase, putative
ORF01800	phage Tail Collar
ORF01810	exonuclease, RNase T and DNA polymerase III, putative
ORF01810	exonuclease, RNase T and DNA polymerase III, putative
ORF01823	phage conserved protein
ORF01828	phage conserved protein
ORF01829	HNH endonuclease domain protein
ORF01833	phage conserved protein
ORF01914	intron-associated endonuclease 1 (I-TevI) (IRF protein)
ORF01919	DNA ligase
ORF01920	DNA ligase (Polydeoxyribonucleotide synthase [ATP])
ORF01989	exonuclease
ORF01994	group I intron GIY-YIG endonuclease
ORF02011	phage-related protein, gp16
ORF02014	phage Tail Collar Domain family

ORF02028	nuclease
ORF02107	DNA polymerase III subunit alpha
ORF02178	helicase, Snf2 family
ORF02184	DNA polymerase III subunit alpha
ORF02188	DNA polymerase III DnaE, putative
ORF02199	intron-associated endonuclease 1 (I-TevI) (IRF protein)
ORF02217	phage tail tape measure protein, family, putative
ORF02224	chain A, Catalytic Domain Of Intron Endonuclease I-Tevi, E75a Mutant
ORF02230	terminase large subunit, putative
ORF02230	terminase large subunit, putative
ORF02252	phage tail tape measure protein, family, core region
ORF02407	ATP-dependent RNA helicase Dbp3, putative
ORF02587	intron-associated endonuclease 1 (I-TevI) (IRF protein)
ORF02589	group I intron GIY-YIG endonuclease
ORF02595	TnsA endonuclease N family
ORF02694	exonuclease
ORF02696	replicative DNA helicase DnaB, intein-containing
ORF02698	DNA polymerase I, thermostable (Tth polymerase 1), putative
ORF02769	exonuclease SbcD
ORF02820	DNA polymerase I (POL I), putative
ORF02829	putative deoxynucleotide monophosphate kinase (DNK)(dNMP kinase)
ORF02847	HNH endonuclease domain protein
ORF02854	DNA polymerase III, subunits gamma and tau
ORF02901	putative DNA ligase
ORF02917	ribonuclease HI

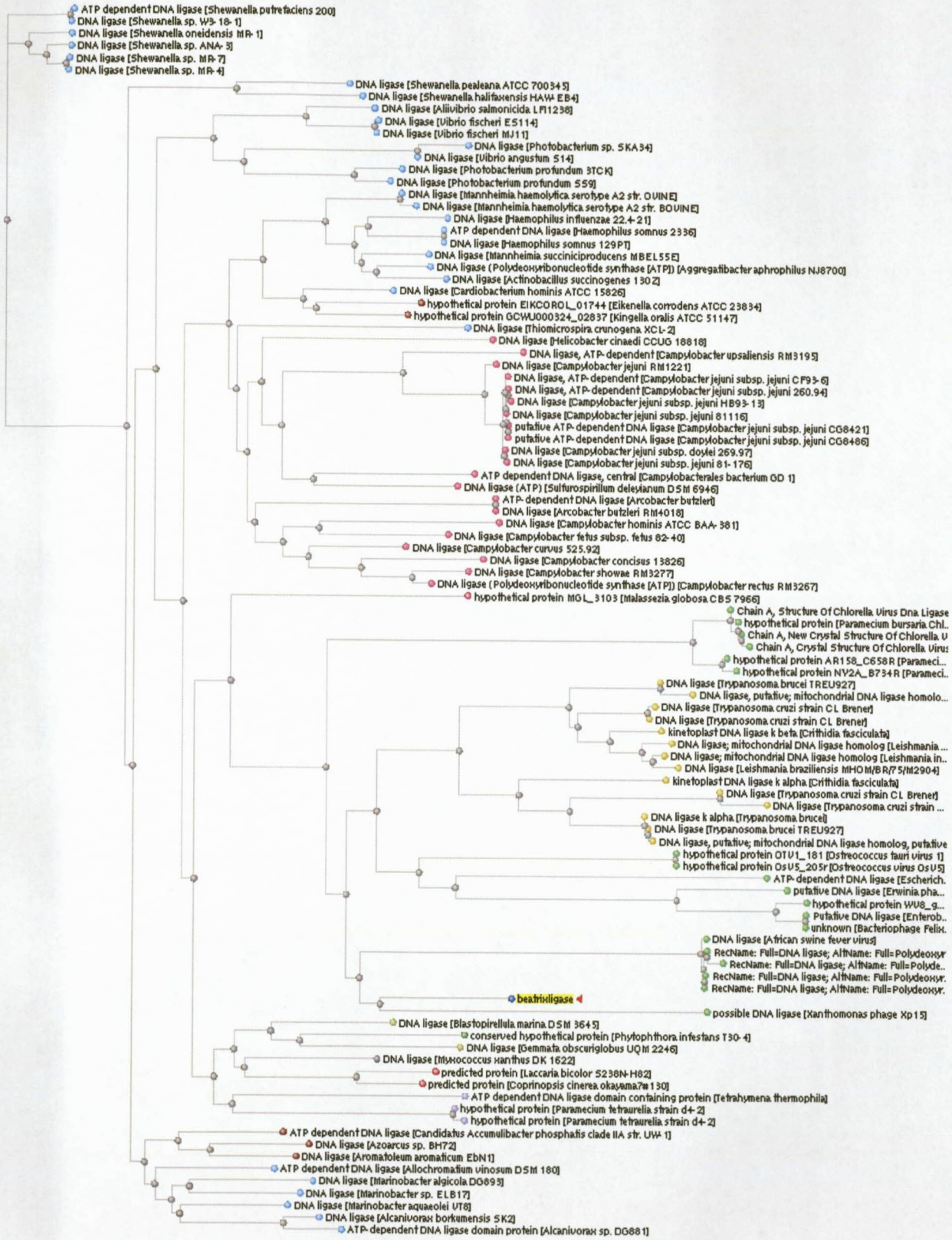
Appendix C

Phylogenetics of the three expressed proteins from Beatrix mine

GIY-YIG endonuclease



ATP-dependent DNA ligase



PseT polynucleotide kinase phosphatase