# METAGENOMIC DISCOVERY AND CHARACTERISATION OF RESTRICTION ENDONUCLEASE FROM KOGELBERG BIOSPHERE RESERVE

**by**

**Sibongile Mtimka**

Submitted in accordance with the requirements

for the degree

**Masters of Science**

in the subject

LIFE SCIENCES

at the

**UNIVERSITY OF SOUTH AFRICA**

SUPERVISOR: PROF. S. GILDENHUYS

JOINT SUPERVISOR: DR. T TSEKOA

CO-SUPERVISOR: DR. P PILLAY

May 2018

# Declaration

Name: ___ Sibongile Mtimka
Student number: __58528962__
Degree: Masters in Life Sciences_
**METAGENOMIC DISCOVERY AND CHARACTERISATION OF RESTRICTION NDONUCLEASE FROM KOGELBERG BIOSPHERE RESERVE**

I **Sibongile Mtimka (Student no. 58528962)** hereby declare that the dissertation, which I hereby submit for the degree of **Masters in Life Sciences** at the University of South Africa, is my own work and has not previously been submitted by me for a degree at this or any other institution.

I declare that the dissertation does not contain any written work presented by other persons whether written, pictures, graphs or data or any other information without acknowledging the source.

I declare that where words from a written source have been used the words have been paraphrased and referenced and where exact words from a source have been used the words have been placed inside quotation marks and referenced.

I declare that I have not copied and pasted any information from the Internet, without specifically acknowledging the source and have inserted appropriate references to these sources in the reference section of the dissertation or thesis.

I declare that during my study I adhered to the Research Ethics Policy of the University of South Africa, received ethics approval for the duration of my study prior to the commencement of data gathering, and have not acted outside the approval conditions.

I declare that the content of my dissertation/thesis has been submitted through an electronic plagiarism detection program before the final submission for examination.

Student signature: _____ Date:

## Acknowledgements

To my supervisor Prof. Glidenhuys, a big thank you for all the support and encouragement she gave me. Without her guidance and constant feedback this M.Sc would not have been achievable.

I would like to sincerely thank Dr. Tsekoa for his guidance, understanding, patience, and most importantly, his mentorship. He encouraged me to not only grow as a student but also as an instructor and an independent thinker. Thank you for affording me this opportunity to conduct my MSc studies.

To Dr P. Pillay thank you for your time, effort, expertise and support you have displayed throughout my research project

I would also like to thank Dr. Rashamuse for his assistance and guidance and providing me with the foundation.

I would also like to thank my friends and colleagues for our scientific debates, exchanges of knowledge, skills, and venting of frustration during, which helped enrich the experience.

I would also like to say a heartfelt thank you to my aunts (Nomangesi Malinga and Khuselwa Malinga-Nopote) for always believing in me and encouraging me to follow my dreams. For helping in whatever way they could during this challenging period.

"Mbonge uYehova, mphefumlo wam"

-Psalm 103

## Abstract

Restriction endonucleases are a group of enzymes that cleave DNA at or around specific sequences, which are typically palindromic. A fosmid library was constructed from a metagenome isolated from soil from the Kogelberg Nature Reserve, Western Cape and was functionally screened for restriction endonucleases. Next-generation (NGS) Illumina sequencing technology was used to identify putative endonucleases. The sequence data generated was assembled and analysed using CLC Bio Genomics Workbench and bioinformatics tools (NCBI BLAST, REBASE and MG-RAST). Using these tools, genes encoding restriction-modification systems and endonuclease homologues were discovered. Three genes were identified and were recombinantly produced in Rosetta™ (DE3) pLysS and purified with IMAC using Ni-TED resin and subsequently characterised. These three genes were selected based on the identity percentage when compared to sequences on the NCBI database. Production of Endo8 was scaled up using 2 l fermenter and the purification done using ÄKTA Avant 150 FPLC using a HiScale 50 column packed with Ni-TED resin and the total amount of protein achieved was 58.82 $mg.g^{-1}$. The productivity achieved at 17 hours (8 h harvest) was 2-fold greater than at 12 hours. Endonuclease activity of endo8 and endo52 was tested, both exhibited strong non-specific activity at 37 °C with an incubation period of 30 min. This work demonstrates that environmental soil samples are a valuable source for discovery of novel enzymes and also the utility of functional metagenomics to discover and purify these enzymes. These endonucleases may contribute to the next generation of reagent enzymes for molecular biology research.

**Key words**: Bacteriophage; Fosmid Library; Functional Screening; Kogelberg Nature Reserve Restriction Endonuclease; Restriction Enzymes; Soil Metagenome.

.

## Table of Contents

# List of Abbreviation

| | |
|---|---|
| AdoMet | S-Adenosyl methionine |
| AMP | Adenosine monophosphate |
| ATCC | American Type Culture Collection |
| ATP | Adenosine Triphosphate |
| BlastP | Basic Local Alignment Search Tool, Protein database |
| B-Per | Bacterial Protein Extraction Reagent |
| Bp | Base pairs |
| BSA | Bovine Serum Albumin |
| CAGR | Compound Annual Growth Rate |
| Cfu | Colony forming units |
| Chlor | Chloramphenicol |
| COG | Cluster of Orthologous Groups |
| DEAD/DEAX | Helicase Family Proteins |
| dNTP | Deoxyribonucleoside triphosphate |
| FPLC | Fast protein liquid chromatography |
| GMO | Genetically Modified Organisms |
| HE | Homing Endonuclease |
| HgDNA | Human Genomic DNA |
| Hr(s) | Hour(s) |
| IMAC | Immobilized metal ion affinity chromatography |
| IPTG | Isopropyl β-D-1-thiogalactopyranoside |
| Kan | Kanamycin |
| KBR | Kogelberg Biosphere Reserve |
| KEGG | Kyoto Encyclopedia of Genes and Genomes |
| Λ | Lambda DNA |
| LB | Lysogeny broth |
| M | Molar |
| MCS | Multiple Cloning Site |
| MgDNA | Mouse Genomic DNA |
| M systems | Modification systems |
| MTase | Methyltransferase |
| $_{met}$DNA | Metagenomic Deoxyribonucleic acid |
| Min | Minutes |
| mol | Moles |
| Nm | Nanometres |
| NGS | Next generation sequence |
| $OD_{600}$ | Optical Density (wavelength is given in the subtext) |
| ORF | Open Reading Frame |
| R&D | Research and Development |
| RE(s)/REase | Restriction Enzyme(s)/ Restriction Endonuclease |

| | |
|---|---|
| R-M systems | Restriction-Modification systems or Restriction and Modification |
| Rpm | Revolutions per minute |
| SAM | S-adenosyl methionine |
| SDS-PAGE | Sodium Dodecyl Sulphate Polyacrylamide Gel Electrophoresis |
| TAE | Tris, Acetic Acid and EDTA |
| TEMED | Tetramethylethylenediamine |
| UV | Ultra violet light |
| V | Volts |
| v/v | Volume to volume ratio |
| w/v | Weight to volume ratio |
| w/w | Weight to weight ratio |
| × *g* | Centrifugal force (gravity) |

# List of Figures

# List of Tables

# 1 Introduction

Deoxyribonucleic acid (DNA) is the molecule responsible for storage and transmission of genetic hereditary information. DNA is incredibly important because of its essential role in carrying genetic instructions for characteristics of each and every living being. Included in those instructions are elements relating to growth, development and reproduction of the organism.

The DNA sequence encodes untranslated regions, genes (open reading frames) and regulatory elements that can encompass various recognition and binding sites to which a range of enzymes bind to effect a range of functions. These functions include DNA replication and division, recombination and repair, transcription, epigenetic modification and encoding protein (Alberts *et. al.,* 2002; Wilson *et. al.,* 2012). These naturally occurring and biologically important enzymes have been exploited for biotechnology and molecular biology and have created an important commercial market for nucleic acid manipulating enzymes.

Enzymes with the ability of manipulating nucleic acids *in vivo* and *in vitro* are very important in molecular biology research and development, and have led to the rapid expansion of revolutionary recombinant technology and genetic engineering. The sustained and growing interest in recombinant technology, molecular diagnostics, genomic and gene-expression analysis and gene editing is accelerating demand for products related to manipulation and analysis of nucleic acids. According to Brown (2002) these nucleic acid manipulating enzymes can be clustered into the following broad categories:

**(i) Polymerase**

DNA polymerases synthesize new polynucleotides that correspond to an existing DNA or RNA template. They are molecular vehicles that guide the synthesis of DNA from nucleotides. There are seven different groups to which DNA polymerases are categorised in, viz. A, B, C, D, X, Y and RT (Rothwell and Waksman, 2005). Polymerases structural framework consists of three subdomains:(i) the fingers, (ii) palm, and (iii) thumb subdomains (Rothwell and Waksman, 2005).

**(ii) Nucleases**

Nucleases disrupt DNA and RNA backbones by cleaving phosphodiester bonds. These enzymes are organized into three sub-groups, namely: Endonucleases have the ability to

nucleic acid from internal sites; Exonucleases have the ability to digest nucleic acid from 5' end or 3' end and lastly; the endo-exonucleases have the ability to cleave DNA and/or RNA from both internal and terminal positions. Endonucleases play an important role in the repairing of DNA. They were first available commercially in the early 1970's and the number continues to grow. Restriction endonucleases cleave DNA at particular point and are an essential tool in the molecular biology toolkit (Williams, 2003). The point at which the enzyme cuts in a nucleotide sequence is known as a restriction site. Numerous endonucleases exist; however, they differ in mechanism, structure, and applications, and they remain a great area of interest (Williams, 2003). These enzymes recognise a particular nucleotide sequence and cleave it apart with the addition of a water molecule which assists with breaking the bond into two nucleotides. However, some DNA sequences have repetitive sequences, therefore, the restriction enzymes then cleave the DNA into a thousand or even million fragments. In the number of RE known, recognition sequence is generally four to eight base pairs in length and palindromic (Robinson *et. al.,* 2001)

**(iii) Ligases**

According to Pascal (2008), ligases fix breaks in the nucleic acids' backbone structure by creating a phosphoester bond between opposing 3' hydroxyl and 5' phosphate ends. DNA ligases can be categorised into two classes, the NAD+ and ATP dependant and molecular weight ranges from 70-80 kDa and 30 - >100 kDa respectively (Doherty and Suh, 2000). Modes of action of ligases can be split into three, (i) the activation of the enzyme from formation of protein-adenosine monophosphate (AMP) intermediate, (ii) transfer of the AMP moiety ligase to 5' phosphate group on the break site (single strand), and (iii) the ligation step is catalysed with loss of free AMP (Doherty and Suh, 2000; Pascal, 2008).

**(iv) End-modification enzymes**

End-modification enzymes are capable of making alterations to the end of DNA molecules by adding or removing a phosphate group. Unlike DNA polymerases, they are template independent, they can synthesize new DNA without base-pairing (Brown, 2002). The most widely used end-modification enzymes are alkaline phosphate and T4 polynucleotide kinase.

DNA manipulating enzymes are extensively used in the field of life sciences, especially in research and diagnostics. They are used for amplification, detection, cloning, recombinant expression, mutagenesis, and analysis of nucleic acids. The global market for alternative

restriction endonucleases is estimated to be $155 million in 2018 (Dewan, 2014). New developments and research within the market have resulted in excellent market growth. The market is projected to grow at a compound annual growth rate (CAGR) of 7.41 % in the next five years (Dewan, 2014). Technological advances in DNA manipulation and diagnostic enzymes are believed to drive the growth of research and biotechnology. New and improved techniques are constantly needed to facilitate rapid and efficient discovery, expression, and characterization of bacterial genes for different purposes including producing useful proteins for laboratory research. .

## 1.1 Restriction endonucleases and restriction-modification systems

According to Robinson, Walsh and Bonventre, (2001), R-M systems developed in bacteria and archaea to protect them from invasion by viruses or bacteriophages, acting as a defensive mechanism (Figure 1.1). Restriction-Modification (R-M) systems comprising restriction enzymes (RE's) are mainly found in bacteria and archaea (Wilson *et al.*, 2012; Vasu and Nagaraja, 2013). Some exceptions include HsaI, a RE that was isolated from humans (Lao and Chen, 1986). Restriction involves the breakage of DNA by hydrolysing the phosphodiester backbone of both strands while modification involves methyltransferase activity, which adds a methyl group at a position that blocks the paired restriction activity (Blumenthal and Cheng, 2002). Restriction and Modification genes are usually closely linked and often overlap or are separated by a few nucleotides (Wilson and Murray, 1991). The R-M activities are in most systems separate but in others the two can be combined in a multi-subunit or single enzyme. The MTases are in charge of exchanging methyl group from S-adenosyl methionine moving it to carbon number 5 or N4 amino group of cytosine or to the N6 amino group of adenine (Vasu and Nagaraja, 2013) Classical R-M systems consists of a R-subunit (restriction) and M-subunit (methylation) and these two generally comprise two enzymatic functions: 1) the endonuclease enzyme activity, which recognises and cleaves unmethylated (therefore unprotected) DNA at the restriction site, and 2) the methyltransferase enzyme activity, which protects endogenous DNA by modifying it at the recognition sequence by methylation of either adenosyl or cytosyl residues within the sequence (Bickle and Kruger, 1993). The R- and M- subunits usually occur on separate open reading frames, however, in some cases, from two separate enzyme molecules, while in other instances are subunits on one oligomeric enzyme structure.

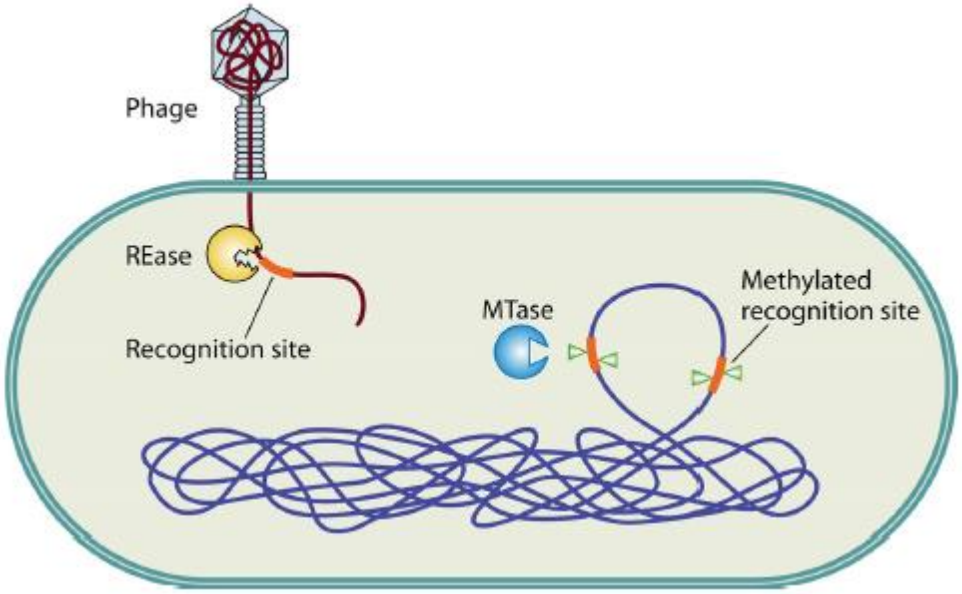



**Figure 1.1: Image illustrating the defence mechanism of restriction-modification systems was taken from Vasu and Nagaraja, 2013.**

.Restriction-Modifications systems identify the methylation status of incoming foreign DNA. Incoming DNA that is unmethylated is identified as forgein and therefore cleaved by restriction endonuclease (REase). Methylated DNA on the other hand is recognised as part of the DNA within. The methylation status at the genomic recognition sites is maintained by the cognate methyltransferase (MTase) of the R-M system

## 1.2 The different types of restriction enzymes

Restriction endonucleases (REs) are classified into six groups (I-VI): Type I, II, III, and IV (Loenen *et al.*, 2014), Type V and artificial restriction enzyme; however, four types (Type I, II, III & IV) of the six are well researched (Vasu and Nagaraja, 2013). All these classifications came about due to differences in the enzymes' recognition sites, subunit composition, cleavage position, co-factor requirements and substrate preference (Williams, 2003). Among the various types of REs, type II REs are the most extensively studied and used enzymes. Figure 1.2 illustrates the ends generated by various restriction enzymes. In recent advances, artificial nucleases have been created resulting in the production of two nucleases (zinc-finger nucleases and TAL-effector nuclease, or TALENs), and in the emergence of revolutionary genome editing techniques (Loenen *et al.*, 2014) using RNA-guided restriction.

### 1.2.1 Type I

According to Weiserová & Ryu (2008), this group of enzymes is the most complex group of REs. Type I enzymes cut the DNA far from their recognition site (Loenen *et. al.,* 2014) and require two co-factors (ATP and S-adenosyl-L-methionine) to function. Type Is are large, multifunctional and oligomeric enzymes. These proteins are encoded by three genes (*h*ost *s*pecificity *d*eterminant, *hsd*): restriction (R), modification (M) and recognition (S, specificity) gene (Loenen *et. al.,* 2014). They are versatile proteins with both restriction and modification active sites in a single protein (Vasu and Nagaraja, 2013). Type I were the first type of REW to be identified and it were first isolated from *E. coli* K-12 (Wilson *et. al.,* 2012). Type I enzymes have four subgroups Type IA (EcoKI), Type IB (EcoAI), Type IC (EcoR124I) and Type ID (StySBLI) (Blumenthal and Cheng, 2002; Weiserová and Ryu, 2008). They are generally large because they possess a MTase, ATPase, and DNA translocase and endonuclease activity. According to Davies *et. al.,* (1999), enzymes in this group are considered smart because of their capability to distinguish their own particular DNA methylation on the target sequence and counter it with a different action which encompasses extensive DNA translocation.

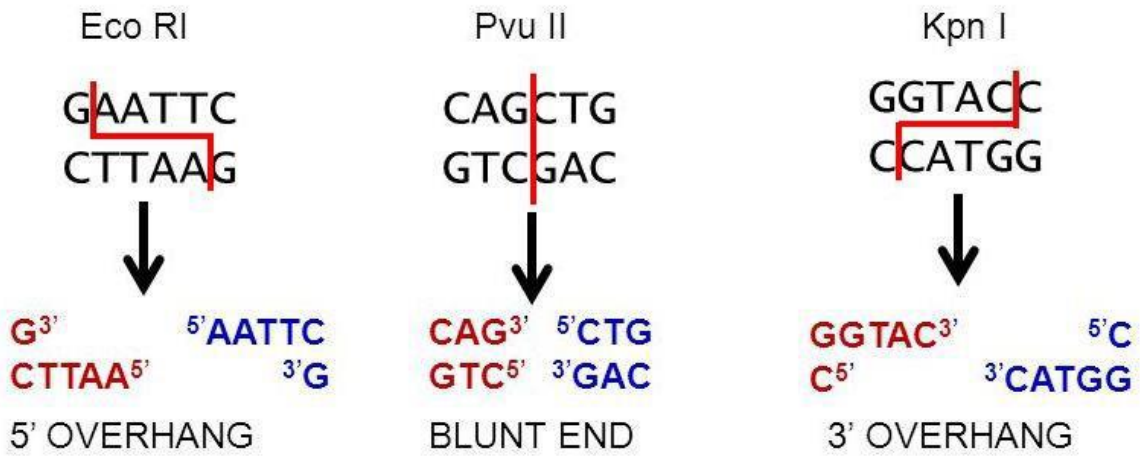



**Figure 1.2: Image was taken from Divan and Royds, (2013). The images show the different ends generated by different types of restriction enzymes.**

*Eco*RI is globular type II restriction enzyme found in the bacteria *E. coli*. It creates 4 nucleotide sticky ends with 5' end overhangs of AATT. *Pvu*II is a type II restriction endonuclease from *Proteus vulgaris* and cleaves DNA between the central GC base pair of its recognition sequence (5′-CAGCTG-3′) in a $Mg^{2+}$-dependent reaction. It generates blunt-ended products. *Kpn*I is from *Klebsiella pneumoniae*. It recognizes double-stranded sequence GGTACC and cleaves after C-5.

### 1.2.2 Type II

Type II is the largest of the four major groups, extensive studies have been done on this particular group and they are highly useful in recombinant DNA (Wilson *et al.,* 2012). Type II REases differ greatly and they arise in various structural forms. With more studies and discoveries done on this group and with the distinct variation amongst the more recently characterised enzymes, subgroups were established (Table 1.1).

Although these various type II RE subgroups exist, the enzymes have a comparable centre, which harbours the dynamic site (one for every subunit) and additionally serves as a critical structure stabilization factor (stabilization centre) (V. Pingoud *et al.*, 2005). REs have a core structure that comprises a five-stranded mix of β-sheets lined by α-helices (V. Pingoud *et al.*, 2005), the second and third strand of the β-sheet serve as a scaffold for the catalytic residue of the PD-D/ExK motif (Figure 1.3). The fifth β-strand can be parallel or anti-parallel to the fourth strand. Enzymes belonging to the PD-D/ExK superfamily can be divided into two branches: (i) one branch cleaves DNA bonds and generates 5’ overhangs of four-bases and the other (ii) branch cleaves DNA bonds and generates “blunt” ends.

**Table 1.1: Different subtypes of type II REases and their short description(s). The table was adapted from Loenen *et al,*. 2014.**

As much as these enzymes are classified into different classes, they are not exclusive; one enzyme can belong to several other subgroup and/or class. One example is the *Bcg*I, it belongs to different subgroups

| Subtype | Features of restriction enzymes | Examples |
|---|---|---|
| **Type IIA** | Asymmetric recognition sequence | *Fok*I |
| **Type IIB** | Cleavage on both sides of the recognition sequence | *Bcg*I |
| **Type IIC** | Single, combination R-M polypeptide | *Hae*IV |
| **Type IIE** | Two sequences required for cleavage, one serving as allosteric effector | *Eco*RII, *Sau*3AI |
| **Type IIF** | Two sequences required for cleavage concerted reaction by homotetramer | *Sfi*I |
| **Type IIG** | Requires AdoMet cofactor for both R-M | *Eco*57I |
| **Type IIH** | Separate M and S subunits; MTase organization similar to Type I systems | *Bcg*I |
| **Type IIM** | Require methylated recognition sequence; Type IIP or Type IIA | *Dpn*I |
| **Type IIP** | Palindromic recognition sequence; recognized by both homodimeric and monomeric enzymes; cleavage occurs symmetrically, usually within the recognition sequence | Prototypes *Eco*RI & *Eco*RV |
| **Type IIS** | Asymmetric recognition sequence; cleavage at fixed positions usually outside recognition sequence | *Fok*I |
| **Type IIT** | Heterodimeric restriction enzyme | *Bpu*10I, *Bsl*I |
| **Putative** | All subtypes | |
| **Control** | Control proteins of Type II restriction enzymes | C.*Bam*HI, C.*Pvu*II |

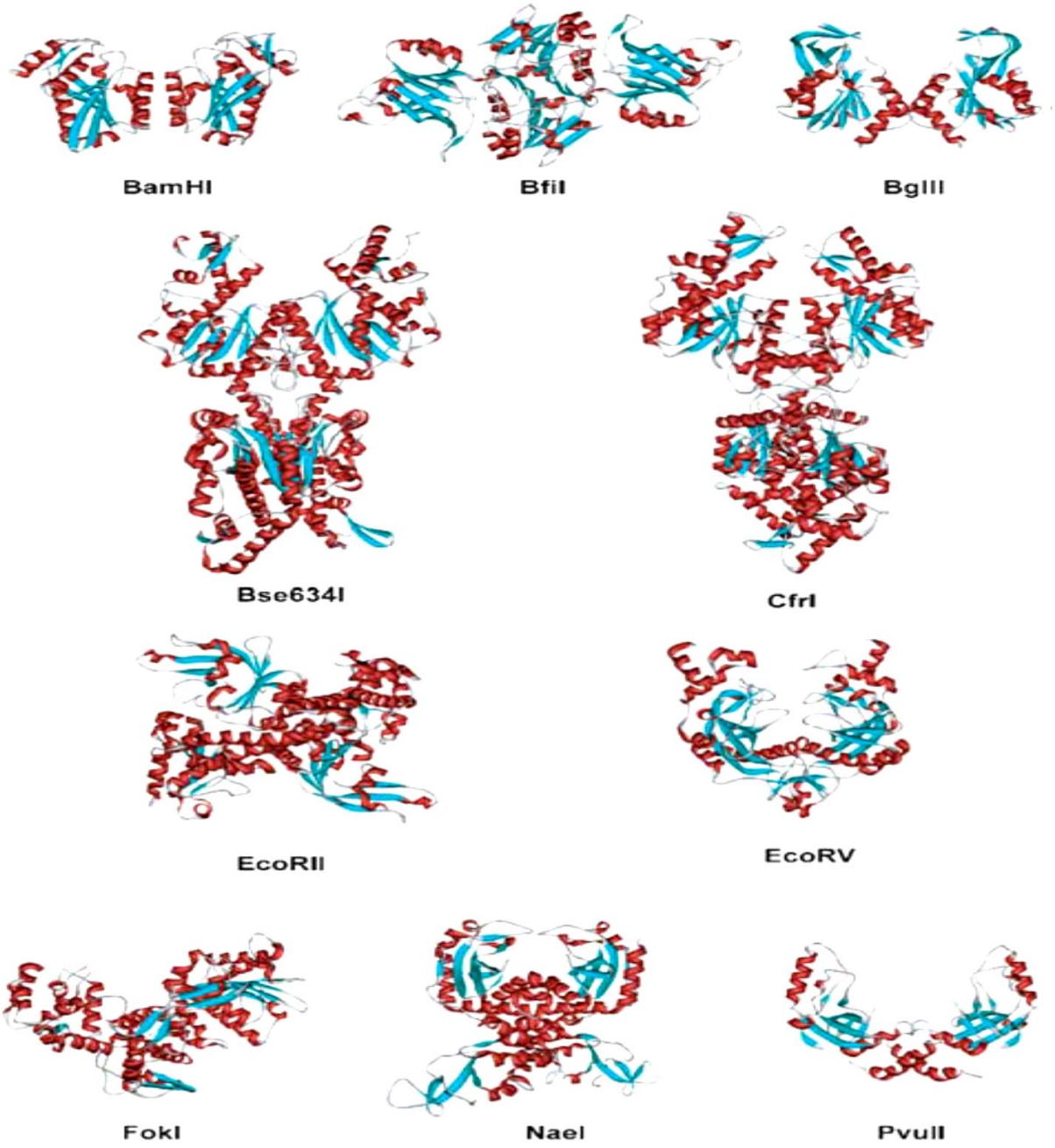



**Figure 1.3: Crystal structures of Type II restriction endonucleases adapted from Pingoud *et al.*, (2005).**

*Bam*HI (1*BAM*), *Bfi*I [V. Siksnys, unpublished], *Bgl*II(1ES8), *Bse*634I (1KNV), *Cfr*10I (1CFR), *Eco*RII (1NA6), *Eco*RV (1RVE), *Fok*I (2*FOK*), *Nae*I (11EV7), *Pvu*II (1*PVU*). α-helices are indicated in red, β-strands in blue. Note that *Bse*634I, *Cfr*10I are homotetrameric enzymes, RII and *Nae*I as Type IIE enzymes have an extra domain and *Fok*I and *Msp*I are monomeric enzymes in the co-crystal.

### 1.2.3 Type III

Type III enzymes have a place in the beta-subfamily of N6 adenine methyltransferases which are known to have nine motifs. Type III enzymes restrict the DNA away from the actual recognition sites and require two co-factors (ATP and SAL), although the cofactors are not a definite requirement. Properties of enzymes within this group are said to be intermediate between Type I and II group and are known to recognise asymmetric sequences (Loenen *et al.,* 2014). The MTase subunit of type III enzymes is active independently of the REase subunit but the REase is only active as a unit (MTase and REs) in the presence of a stimulant (AdoMet) (Blumenthal and Cheng, 2002). These enzymes contain two subunits (restriction and modification) thus making the proteins multifunctional, hetero-oligomeric. They perceive two separate non-palindromic sequences that are conversely situated (Vasu and Nagaraja, 2013).

### 1.2.4 Type IV

Enzymes that are in this subgroup identify asymmetrical DNA sequences and target DNA that has either been methylated, hydroxymethylated or glucosyl-hydroxymethylated (Tock and Dryden, 2005). This group of enzymes has an unusual characteristic in that it cuts on both ends of the restriction site (Williams, 2003). Lepikhov *et al.,* (2001) suggested that this group of enzymes is made up of enzymes that are in-between type IIS and type III. This group of enzymes were initially classified under type IIB enzymes but due to later discoveries and their need for co-factors ($Mg^{2+}$ and AdoMet) for cleavage, they were then proposed into a new group designated type IV.

### 1.2.5 Type V

These enzymes cleave DNA of flexible lengths; given a reasonable guide RNA is used. CRISPR enzymes are part of the type V group and unlike other REases, they use guide RNAs. Due to this system, they are very flexible and can be easily programmed to target virtually any DNA or RNA substrate (Wilkinson and Wiedenheft, 2014).

### 1.2.6 Artificial restriction enzymes

These enzymes are a set of proteins that exist from having a nuclease space intertwined with either a characteristic or a built DNA restricting area (Nwankwo, 2014). Enzymes in this group have the ability to target up to 36 bp and can be also be modified so that they can bind to a required DNA sequence. The most commonly used enzymes in this group are the Zinc finger

nucleases and is normally used in genetic engineering, although it can also be used in cloning (Nwankwo, 2014). Transcription activator-like effector nucleases (TALENs) are made up of 33-34 amino acid with two mutable regions that are for specific nucleotides (Ansai *et al.*, 2013). TALENs are fusion of nuclease domain of FokI and DNA recognition domain of transcription activator like effector (Ansai *et al.*, 2013). The flexibility of ZFNs and TALENs emerges from the capacity to alter the DNA-binding domain to distinguish almost any sequence (Gaj *et al.*, 2013). The most common motifs for DNA-binding in eukaryotes is the Cys2-His2 zinc-finger domain and in the human genome, it is the second regularly encoded protein domain (Gaj *et al.*, 2013).

## 1.3 R-M functions and importance in Bacteria and Archaea

The well-known function of restriction–modification is the ability to protect bacteria and archaea against foreign DNA invasion. The increase in occurrence and variety of restriction-modification systems highlights their success in acting as defence mechanisms in the world of bacteria and archaea (Vasu and Nagaraja, 2013). R-M systems importance in bacteria and archaea aid in maintaining them and this is importance is regulated by the selfish gene within the R-M systems. According to Kobayashi (2004), the substitution of R-M systems within bacteria/ archaea can result in cell death. The function of the selfish gene act is supported by a number of gene experiments and analysis (Naito *et al.*, 1995; Kobayashi, 2001; Engelberg-Kulka *et al.*, 2006). R-M systems play a role in gene variation and they can in the generation of gene discovery.

### 1.3.1 Gene arrangement and regulation of R-M systems

R-M systems have genes that are well organised and linked spatially (Figure 1.4). Arrangement of the adjacent gene can be: 1) parallel, having the 5’ end from one gene following the 3’ end of the other; 2) convergently, having the 3’ end in proximity, or 3) divergently, having the 5’ ends close one another. Regulation of R-M systems is crucial to prevent auto-restriction. The composition of R-M systems consists of restriction enzyme (toxin) and the modification enzyme (anti-toxin). Regulation of R-M systems is not fully understood, however, the MTase is said to appear before the 5’ of REase in the gene (Rimšelienė *et al.*, 1995). R-M systems limit the genetic transition among lineages with special epigenetic identities, as described with the aid of DNA methylation that’s sequence-specific (Mruk and Kobayashi, 2014), i.e. the regulation of the REase and MTase ensures that the cell's own DNA is protected before any REase activity appears (Tao *et. al.,* 1991). To date, the regulation of type II R-M systems has been described

to involve three mechanisms (Česnavičiene *et. al.,* 2003). The first mechanism uses the helix-turn-helix (HTH) motif of the methyltransferase which is on the N-terminus.

This motif suppresses expression of the MTase gene by binding to the promoter region. The second mechanism takes advantage of the modulation promoter activity, which depends on the methylation status of the target sequences. The R-M system located within the promoter region of the two genes recognises the target sequence. Methylation of the promoter can inhibit the transcription binding factors thus resulting in gene silencing. The third mechanism depends on the activity of the Control protein (C-protein) and its' ORF which can be found in the regions of the restriction and modification genes. C-protein is an additional protein that has been known to be involved in the process of controlling the expression of the R gene. Studies that best illustrate the function of the C-protein have been conducted in *Pvu*II and *Bam*HI systems. They serve as transcriptional activators, which aid in the prevention of the expression of the R gene until the C protein has accumulated and there is sufficient methylase for protection of the REs activity during expression (Roberts *et. al.*, 2003). These proteins also have the ability to control their own expression. A number of Type II R-M systems make use of the C protein to manage the genetic switch of the toxin endonucleases (Mruk and Kobayashi, 2014). C proteins also have the ability to suppress MTase expression allowing the RE to cleave the incoming infection without being modified by the MTase

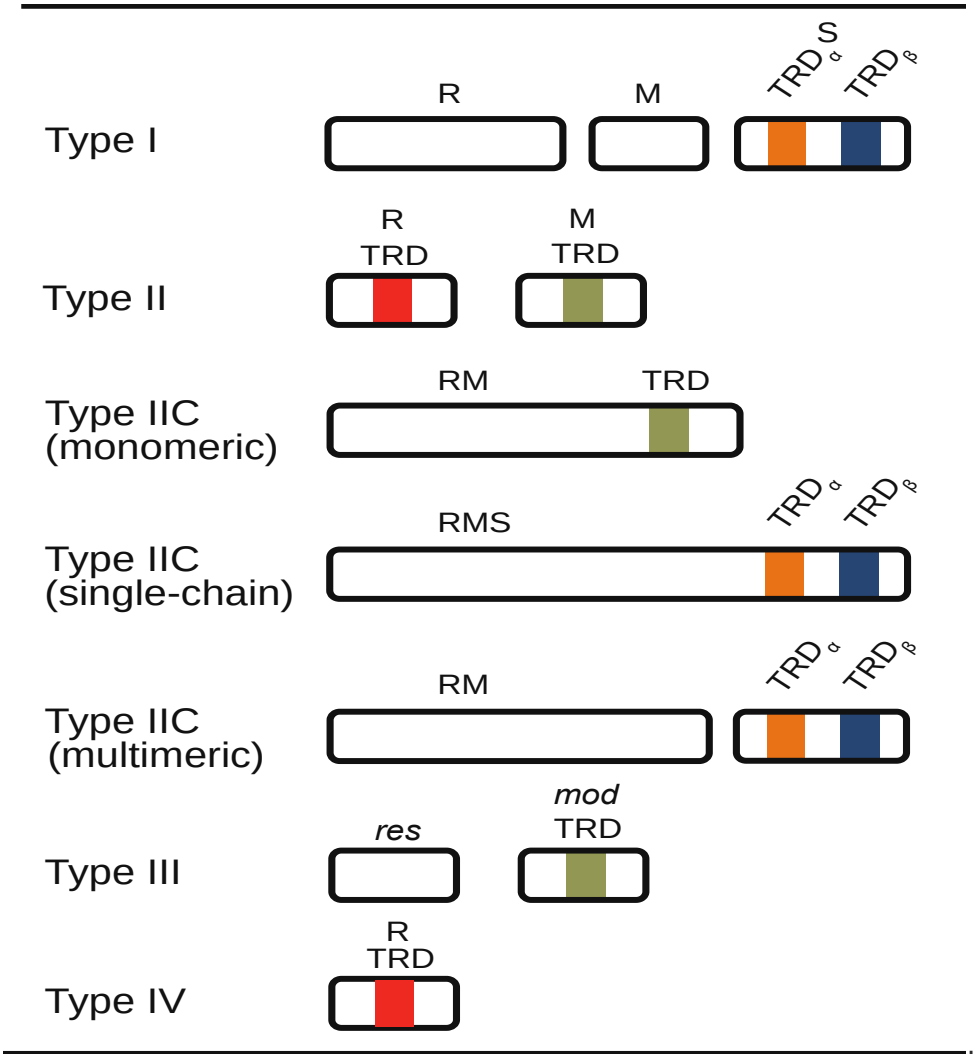



**Figure 1.4: Diagram illustrating gene arrangement adapted from Oliveira *et al*., (2014)**

R-M systems have a gene arrangement that shows how the genes are within an ORF. The picture shows some of the gene arrangements of some of the R-M systems. Restriction (R), modification (M), specificity (S) subunit and target recognition domains (TRDs, shown as colored regions).

## 1.4 Homing Endonucleases

These enzymes share characteristics with restriction endonucleases, in that they also have the ability to cleave DNA at a specific site, but vary in terms of structural and recognition properties and genomic location (Belfort and Roberts, 1997). Unlike restriction enzymes, they fall in six groups with specific sequence motifs namely LAGLIDAD, GIY-YIG, H-N-H, PD-(D/E)xK, His-Cys, and the latest to be discovered being the EDxHD motif (Stoddard *et. al.*, 2014). Figure 1.5 show the different families of Homing endonucleases. HEs have parasitic components which exploit the double strand break-repair mechanism used for propagation by the host DNA (Chevalier and Stoddard, 2001; Hafez *et. al.,* 2012). A number of REs in fact belong to the HNH family and these include *Pac*I, *Hpy*99I, and *Kpn*I.

According to Jurica & Stoddard (1999) and Chevalier & Stoddard (2001), homing refers to an high frequency event of site specific gene conversion. The site specific gene conversion occurs when a mobile intervening sequence from an intron (group I or II) or an intein, this sequence is then copied and shifted to a specific site within a cognate allele which is missing the intervening sequence (Jurica and Stoddard, 1999; Chevalier and Stoddard, 2001). The final product of homing is the replication of the introns.

Homing endonucleases can be clustered into various groups based on their distinct structural characteristics; however, these groups share a common ancestor with different host proteins with an unrelated function (Scalley-Kim *et. al.,* 2007). Those proteins encompass restriction endonucleases, DNA mismatch repair proteins, transcription factors, four-way junction resolving enzymes, and colicins (Taylor and Stoddard, 2012). According to Chevalier & Stoddard (2001), even though homing endonucleases have differences in structure and mechanisms, they all have related functional requirements, which have a linkage with the wide variety of genomic and biological hosts. One of the first homing endonucleases (I-Ssp6803I) to be presented to have the PD-(D/E)xK motif and to also resemble REs is from the PD-(D/E)xK motif family (Zhao *et. al.*, 2009). Though REs fall in the PD-(D/E)xK superfamily, a number of type II REs have either the GIY-YIG or the HNH motifs. Table 1.2 below contrasts the properties of restriction enzymes and homing endonucleases and was adopted from Belfort and Roberts (1997).

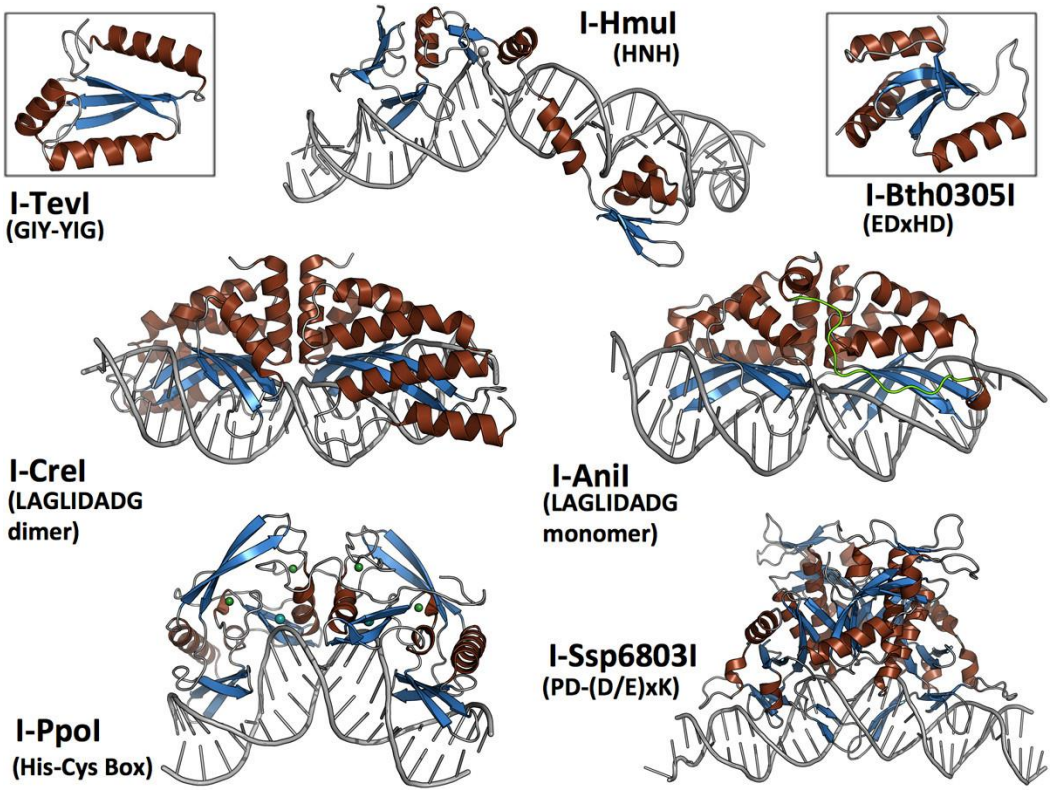



**Figure 1.5: Image representing the structural representation of families and subfamilies of homing endonuclease was adopted fromStoddard et. al., (2014)**

Top: these three GIY-YIG, HNH and EDxHD different catalytic nuclease domains are found in a number of phage-encoded homing endonucleases. Middle: two closely related types of LAGLIDADG homing endonucleases, they correspond to homodimeric and monomeric enzymes, are encoded within organellar and archaeal genomes. Bottom: These enzymes are mutlimers; one is a homodimer and the other a tertramer. Bottom: Endonuclease with the His-Cys box which harbour a variant of the HNH active site and PD-(D/E)xK endonucleases are found in protist and cyanobacterial genomes. Both enzymes are multimers (a homodimer and a tetramer, respectively)

**Table 1.2 Comparison of homing endonucleases and restriction enzymes adopted from Belfort and Roberts, (1997).**

Homing endonucleases share the same ability to cleave double stranded DNA at specific sites, however, they differ in a number of things such as (i) structure, (ii) recognition properties, and (iii) genomic location.

| Property | Homing endonuclease | Restriction enzyme |
|---|---|---|
| **Conserved protein motifs** | i. LAGLIDADG<br>ii. GIY-YIG<br>iii. H-N-H<br>iv. His-Cys | a) None definitive |
| **Recognition sequences** | i. Lengthy (12–40 bp)<br>ii. Asymmetric<br>iii. Sequence-tolerant | a) Short (3–8 bp)<br>b) Symmetric and asymmetric<br>c) Sequence-specific |
| **Accessory molecules** | i. Some require protein or RNA for full activity | a) Some require methyltransferase components or specificity sub-units |
| **Genomic location** | i. Intron, intein, or intergenic<br>ii. All three biological | a) Flanking modification gene<br>b) Confined to archaea, bacteria and some kingdoms eukaryotic viruses |

## 1.5 Applications of restriction enzymes in biotechnology & molecular biology

REs are widely utilised in molecular biology as key reagents in a number of applications. One of the most prominent uses of REs is for cutting DNA into fragments that facilitate the identification and characterisation of genes. Other applications include genetic engineering, recombinant DNA cloning, molecular husbandry, Southern blotting analysis, genetic engineering and Restriction Fragment Length Polymorphism analysis (RFLP)(Robinson *et al.*, 2001; Loenen *et al.*, 2014). Table 1.3 indicate the different applications of restriction enzymes and their uses in molecular genetics. Some of the uses of REs in revolutionary recombinant technology have led to companies such as Genentech being able to produce insulin from bacteria and yeast and Biogen to recombinantly produce a vaccine for Hepatitis B (Loenen *et al.*, 2014). When working on restriction enzymes, it is critical to focus more on REs that are potentially useful (and profitable) in industry, medicine and agriculture. These fields are poised to continue to grow. In gene therapy, obtaining enzymes that can target several genes at once are more desirable in this application. Availability of enzymes that can target specific sites in a double strand can also aid in the disruption of genes and target genes in refractory cells (Certo and Morgan, 2016).

**Table 1.3: Molecular genetic techniques and their explanation adopted from Espinoza-miranda *et. al*., (2012).**

The tables give details on the key applications of RE and their importance to the world of molecular genetics.

| Approach | Usages |
|---|---|
| **AFLP (Amplified Fragment Length Polymorphisms)** | Works by using both REase and polymerase chain reaction (PCR). It has been used in studying the genetic variation in the natural population of *Pinus oorcarpa* from Nicaragua utilising two restriction enzymes. |
| **PCR/REA (Polymerase Chain Reaction/ Restriction Enzyme Analysis):** | This technique focuses on amplifying variable regions of 23S rRNA and is later accompanied by the use of two restriction enzymes for digestion. It is said to be a simple technique to carry out. When using this technique, one can easily separate and distinguish bacterial species that are intimately linked |
| **PFGE (Pulsed Field Gel Electrophoresis):** | A gold standard technique used in the molecular identification of chromosomal DNA. For the separation of 10-800kb fragments, results from this method are evaluated using software. However, the downside to this method is that results of samples of large quantity will be obtained two to three days. |
| **RFLP (Restriction Fragment Length Polymorphisms)** | Is frequently used as a marker for genetic linkage analysis in forensic genetics, hereditary diseases, and population genetic studies. It was the first to be used in tests relating to human identity |
| **RSM assay (Restriction Site Mutation Assay)** | This method is utilised in studying mutations in codons in genes that linked with the growth of tumours. |

## 1.6 Metagenomics and its application in enzyme bioprospecting

Culture-enrichment methods have been used to gain access to novel enzymes from environmental sources. This method comprises of the cultivation of micro-organisms and the subsequent screening of the pure strains for desired activities (Wahler and Reymond, 2001). For a very long time, restriction enzymes were isolated/detected using traditional microbiology methods and that has limited new discoveries (Espinoza-miranda *et al.*, 2012). Metagenomics have now been adopted to facilitate the discovery of novel restriction enzymes and/or to improve the existing ones. With this approach, microbial DNA can be extracted directly from the environmental sample. Various techniques have been developed to access these novel enzymes from various environmental samples including metagenomics, PCR-based sequence-independent and genome sequencing approaches (Wilson and Piel, 2013). Handelsman describes metagenomics as the study of genomic DNA that is directly isolated from the natural environment (Handelsman, 2004). The approach is highly favoured as it eliminates the need to culture microorganisms for DNA extraction. Using metagenomics, access can be gained into characterising and quantifying unexplored and diverse microbes. This strategy can also be used for the discovery of novel genes and resulting in enzymes that are relevant for biotechnology and pharmaceutical industries (Culligan *et. al.,* 2014). Apart from revealing novel enzymes from nature, the use of metagenomics results in enzymes/proteins that can be used in other fields such as bioremediation, personalised medicine and xenobiotic metabolism (Bashir *et. al.*, 2014)**.**

### 1.6.1 Applied metagenomics and gene discovery

Applied metagenomics refers to strategies that comprise of a combining both functional and sequence-based techniques. It is aimed at rapidly isolating, cloning and over-expressing genes with desired activity from metagenomic libraries (Mathur *et. al.,* 2005). In practice, these techniques involve the extraction of community or environmental DNA (eDNA) from a selected environment, the extracted eDNA is then processed and cloned into a vector (either a plasmid, cosmids, fosmids or bacterial artificial chromosomes depending on the size of the clone) to create a metagenomic library (Li *et al.*, 2009). Based on the target, the library can either be sequenced directly or functionally screened for desired properties. Figure 1.6 illustrates the metagenomics approaches for discovery of novel genes. It gives a clear picture on each of the various steps taken in the two different techniques (functional-based vs sequence-based).

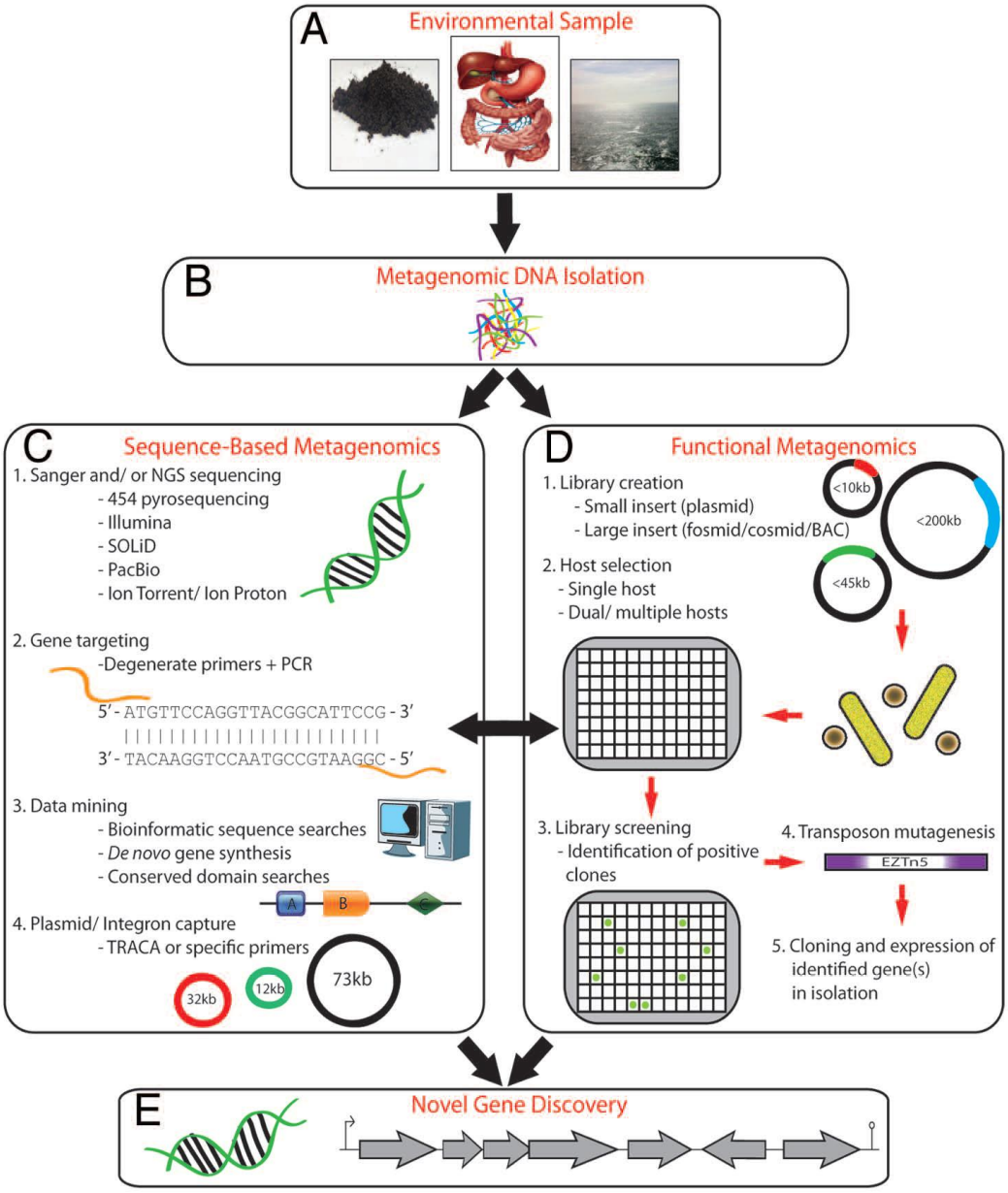



**Figure 1.6:Schematic illustration of metagenomic approaches that are used to discover novel genes adapted from Culligan *et al.*, (2014).**

From (A) sample collection from a given sample, (B) isolation of metagenomic DNA from the sample either by using kits that are available or conversational approaches, (C) and (D) sequencing the isolated metagenomic DNA using Next Generation Sequencing platforms and us of functional techniques which include construction of library using a vector suitable for the study respectively and finally (E) the discovery of the gene discovered from both (C) and (D) or with just one.

Functional metagenomics is an approach that is designed to identify unknown genes and their encoded enzymes from a metagenomic library through heterologous expression screening. The phenotype presented by the expression host as a result of the cloned DNA is detected directly (Kakirde *et al.*, 2010). In functional-based approach, the target gene is accessed without any prior knowledge of the sequence information, unlike sequence-based approach which requires prior knowledge. Using this approach has led to the discovery of novel gene products. Numerous enzymes that are available have been discovered through the use of functional metagenomics, however, highly effective screening techniques to discover DNA manipulating enzymes are currently underdeveloped. The discovery of such efficient screening techniques with specificity for target enzyme is crucial.

In sequence-based metagenomics, the whole gene of interest could be identified through direct sequencing combined with sequence homology studies or by using Polymerase Chain Reaction (PCR) with primers targeting conserved sequences (Li *et al.*, 2009). As opposed to the functional-based approach, the sequence-based approach cannot easily identify new genes. The sequence-based approach is limited to discovery of new members of already existing genes (Kakirde *et al.*, 2010). This metagenomics approach allows the whole microbial genome to be acquired in a complex environmental sample.

Next generation sequencing (NGS) is a major player in this sector of metagenomics discovery of enzymes. With a single run, NGS can produce sequence output with a great deal of nucleotides (Voelkerding *et al.*, 2009). According to Voelkerding *et. al.,* (2009) and Liu *et al.,* (2012), the widely used NGS platforms are: Roche/454 FLX pyrosequencer, Illumina system, SOLID system, Helicos system and PacBio RS II. According to Buermans and den Dunnen (2014) the use of the above platforms has played an instrumental role in the acquisition of sequence data from DNA/RNA. Roche platforms uses pyrosequencing and produces long reads which makes it easy to map genomes, however it has a challenge of having a high error rate in homopolymer (Buermans and den Dunnen, 2014; van Dijk *et al.*, 2014). PacBio RS is similar to the Roche but produces longer reads and it doesn't require PCR step, it makes use of real-time sequencing (Liu *et al.*, 2012). PacBio is relatively expensive and isn't right for small laboratories (van Dijk *et al.*, 2014). Illumina is the leading platform with NGS industry followed by SOLID system (van Dijk *et al.*, 2014). According to van Dijk *et al.,* (2014), in all the NGS platforms, Illumina offers high-through at low cost and SOLID has the lowest error rate.

### 1.6.2 Molecular biology enzyme discovery by metagenomics

The screening assay technique has to be capable of screening a large number of clones in parallel (i.e, high-throughput) and have fairly limited false positives or false negatives (Schoenfeld *et. al.*, 2010). The most used and simple technique for functional screening is the plate-based colony assay. The discovery of the gene of interest also depends on a number of linked factors. These factors include: gene target size, host-vector system, it's within the metagenome, approach used for screening, and the successful expression of the gene in a selected host (Uchiyama and Miyazaki, 2009)**.** The approaches used to screen for enzymes is based on activity and techniques that can be used include: agar plates assays, growth assays, reporter assays, cell lysate assays, pH indicator assays, growth inhibition assays and liquid based assays. Walder *et al.*, (1981) successfully cloned and expressed *Pst*I in *E. coli* and the screening of *Pst*I positive clones were selected based on resistances acquired from bacteriophage lambda phage. This same principle was also used by Mann *et al.,* (1978) to screen for and clone *Hha*II.

## 1.7 Kogelberg biosphere reserve

The Kogelberg biosphere reserve is found in the Western Cape of South Africa (Figure 1.7). It is known as one the world's greatest biodiversity hotspots and was the first to be registered as a biosphere reserve in South Africa (Pool-Stanvliet, 2013). The reserve is known for its natural beauty and floral diversity. Its extraordinary biodiversity and quality of fynbos mean that it is considered the heart of the Cape Floral Kingdom. The Cape Floral Kingdom has the world's highest plant biodiversity (Goldblatt, 1997). It is characterised by greatness in the plant species and high endemism. The Cape Floral Kingdom itself is thus known as one of the world's biodiversity hotspots. The area has dry summers and wet winters. It is subjected to factors such as periodical fires, low nutrients and summer droughts (Onstein *et. al.,* 2014). The pH of the soil found in the Cape Floral Kingdom reported to be acidic and also lack crucial nutrients such as phosphorus, potassium and nitrogen compounds (Stafford *et. al.*, 2005). The region has a soil environment that harbours a remarkable quantity of undiscovered microbial biodiversity. One environment on earth that is known to perhaps contain the greatest microbial diversity on is the soil habitat (Torsvik *et al.*, 1990; Delmont *et al.*, 2011). Investigations conducted by Stafford *et. al.,* (2005) and Segobola *et al.*, (2018) highlights the possibility of discovering unique microorganisms because of the unique plant biodiversity in the region. Temperature in this area varies with every season. During the summer months, it is reported to be usually hot, dry and

windy. Average maximum temperature is reported to be 22 °C, with the monthly temperature ranging from 27 °C in February to 18 °C in June-Aug (Turpie et al., 2009). The average minimum being 11°C monthly, and average minimum temperature ranging from 7° in July to 16° in January and February (Turpie et al., 2009).


**Figure 1.7: Graphical representation of the location Kogelberg Nature Biosphere within the Western Cape, Courtesy of AfriGIS (Pty) Ltd, Google 2018.**

The Kogelberg Biosphere Reserve, situated to the east of Cape Town, the Western Cape province of South Africa in the Boland Mountains. It is recognized as one of the world's biodiversity hotspots

## 1.8 Discovery and recombinant expression of restriction endonucleases/enzymes

Restriction enzymes evolved from bacteria and archaea as defensive mechanisms against invading viruses/bacteriophages (Robinson *et. al.*, 2001). The first restriction enzyme was documented in the early 1950's by Luria and his colleagues, with their investigation of the interaction of phage infection and host specificity. They observed that phage lambda grew poorly on *E. coli* K (Szalay *et. al.,* 1979). The discovery of Eco B and Eco K was done by using an assay that detected degradation/digestion of unmodified DNA. Szalay and co-workers were able to digest the unmodified DNA while modified DNA could not be degraded (Arber, 1978).

R-Ms systems have a selfish genetic part and this may underlie the function and structure of R-M enzymes. When expressing R-M systems, the host has to be protected from restriction. Therefore, in order to protect the host from restriction, ideally the R-M system in its entirely should be cloned and expressed as a single recombinant DNA molecule, however, the restriction gene has to be closely linked to the methylase gene (Longo *et al.*, 2002). However, with recent advances in the expression systems, vectors such as pETDuet-1 can have the two genes on different MCSs and express as a single DNA molecule. A study done by Ichige and Kobayashi (2005) shows that co-expressing the R and M components can lead to higher expression of RE, presumably by protecting the host DNA from digestion by the overexpressed RE. It is also important to design the construct so that when co-expressing, the M system is expressed first followed by the R system.

There are a variety of prokaryotic hosts available for recombinant reproduction of proteins; however, the most favoured for R & D is *E. coli* (Longo *et. al.,* 2002). *E. coli* has systems that limit the degradation of methylated DNA, either on cytosine residues or adenine residues (Longo *et. al.,* 2002) yeast allows digestion of unmethylated DNA. This affects the success of restriction digestion and bacterial transformations (Casali, 2003). The recommended host for the expression of restriction enzymes is one where such types of restriction systems are inactivated through mutation or loss (Longo *et. al.*, 2002). The *E. coli* strains used in many laboratories for recombinant expression are able to overcome this problem (Casali, 2003). These *E. coli* strains are deficient in R-M systems and therefore, they can be used without concerns around limiting or digesting the foreign transformed DNA. The nature of RE can also cause problems when expressed, by destroying the host cells but *E. coli* can stabilize the restriction modification of REs. Rosetta™ (DE3) expression hosts cells were selected over the others (Moon *et al.*, 2008; Bergmann *et al.*, 2014) because they carry a rare plasmid which facilitates the expression of

rare codons and increases expression after IPTG induction (Casali, 2003). Rosetta™ (DE3) has an *E. coli* basically BL21 (DE3) genetic background, but with the pRARE plasmid, which encodes for several rare codon tRNAs. The pRARE becomes an advantage in a case where target gene has several rare codons such as AUA, AGG, AGA, CUA, CCC, and GGA (Casali, 2003). Codon optimisation of protein sequences can improve the production of a particular recombinant protein (Yadava and Ockenhouse, 2003). However, the art of optimising codons is not fully understood. Codon optimisation is essentially the altering of codons within the sequence (Yadava and Ockenhouse, 2003). Codon optimisation comes through the recognition of synonymous codon substitutions. Initially, codon substitutions were regarded irrelevant however, now it has been established that the approach can have a significant effect on heterologous protein expression (Webster *et al.*, 2017).

The pET expression system is an effective and is highly efficient under the strong control of T7 promoter (Rosano and Ceccarelli, 2014). The T7 promoter within the pET vector system is extremely common for expression of proteins. The pET expression system is inducible with lactose or IPTG. The DE3 cells carry a lambda lysogen which carries the T7 gene that is utilised by T7 promoter within the expression vector (Casali, 2003). The most efficacious combination between a host and vector can then be used to produce the recombinant enzyme at bench-scale prior to scale up using fermentation processes.

### 1.8.1 Microbial fermentation

Fermentation processes can either by done under aerobic conditions or anaerobic conditions; this is purely based on the type of protein being produced or microorganism being cultivated. The recombinant protein can either be produced intracellularly or extracellularly. The enzyme desired is produced using the process of utilising the energy source provided Bioreactors are where the microorganisms convert substrates to the desired products. Bioreactors come in different designs but they are usually cylindrical and with different sizes ranging from a litre to thousands of litres. Bioreactors provide an environment that's favourable and that can be monitored and achieve optimal growth and product litre with the expression system used (Singh *et al.*, 2014). There are several types of bioreactors that can be used in fermentation processes and these include continuous stirred tank bioreactor, bubble column bioreactor, airlift bioreactor, fluidized bed bioreactor, packed bed bioreactor, and photo-bioreactors (Singh *et al.*, 2014). For recombinant protein production, one of the mostly used bioreactor is the continuously stirred

tank reactor. Small stirred tank reactors are normally cylindrical and made of glass while larger ones are typically made of stainless steel.

Fed-batch fermentation is a process whereby fresh media is fed into the bioreactor without removing the initial media. According to Cheng *et al.*, (2009) using fed batch results increased productivity, reduced fermentation time, increased concentration of oxygen dissolved in the medium and reduced toxic effects from the components that make up the medium. The continuous addition of fresh media dilutes the media in the reactor and this reduces the concentration of metabolites that effects the growth of the cells and formation of the recombinant product (Chen *et al.*, 2004). During the run, the feed rate can be manipulated to add in sources of carbon, nitrogen, phosphates, nutrients, precursors, or inducers into the culture (Lim and Shin, 2013).

## 1.9 Characterisation of restriction endonuclease

The enzymes functionality is determined by the amount of enzyme necessary to effectively cleave a DNA substrate eg. λDNA. Adequate storage conditions are essential for the storage of the purified enzyme. Numerous factors contribute to the optimal function of the restriction enzymes within the endonuclease reaction such as the ionic strength, buffer pH, co-factors temperature and reaction time (Jutur and Reddy, 2007; Wei *et al.*, 2008). The activity of any given enzyme is dependent on the substrate and this is essentially because of the influence of sequence flanking the recognition site. The first enzyme extracted was from *Haemophilus Influenza* and showed endonuclease activity when tested against foreign native DNA (Smith and Welcox, 1970). Its activity however, could not be seen on *H. influenza* DNA but showed endonuclease activity on T7 phage DNA (Smith and Welcox, 1970). A number of enzymes have been characterised and among them thses few are reported to be active at different temperature points: *Sma*I at 30 °C, *Spe*I at 50 °C, *BstE*II at 55 °C and *Taq*I 65 °C (Karcher, 1995).

Standard techniques used to carry out functional characterisation of restriction enzymes are similar to those described by Sambrook and Russell (2001), unless state otherwise. A typical reaction entails reaction buffer, $dH_2O$, substrate (DNA) and finally restriction enzyme. A study conducted by Chung *et al.*, (2011), isolated and functionally characterised an enzyme (*Cbe*I) on pDCW68 DNA and within 10 min complete digestion was observed. The enzymes activity was similar to that of *Hae*III enzymes. Various factors (buffer conditions, temperature and incubation period) aiding to optimal functioning of the enzyme were test. Comparison studies of the two

proved on different plasmids indicates that *Cbe*I acts like *Hae*III-like enzymes (Chung *et al.*, 2011).

## 1.10 Scope of the study

The DNA manipulating enzymes are routinely used within the life sciences and play a crucial role as reagents in laboratory research and diagnostics. The enzyme reagent market continues to grow due to the need for novel enzyme functions and/or improved existing enzyme functions and characteristics. Numerous studies using metagenomic screening have yielded enzymes with potential for biocatalytic, reagent and other applications (Adrio and Demain, 2014). Hence, this study will focus on the utilization of the metagenomic approach for the screening and discovery of novel restriction endonucleases from the Kogelberg Biosphere Reserve in the Cape Floral Kingdom.

The inability to capture full biodiversity lowers the rate of discovery of novel DNA/RNA manipulating enzymes. New and improved techniques are needed to facilitate rapid and efficient discovery, expression, and characterization of viral genes to produce useful proteins for laboratory research. Nucleic acid manipulating enzymes are some of the major drivers of advances in molecular biology research.

### 1.10.1 Aim and Objectives

#### 1.10.1.1 Aim

The aim of the study was to use functional coupled with sequence-based techniques in order to discover novel nucleic acid manipulating enzymes, in particular, restriction endonucleases from Kogelberg Biosphere Reserve soil metagenome and functional characterise them.

#### 1.10.1.2 Objectives

- To extract metagenomic DNA and construct a fosmid-based metagenomic library from soil collected from the Kogelberg Biosphere Reserve.
- To develop a functional metagenomic screening technique for REs and other endonucleases.
- To screen the generated metagenome using the novel function-based approach develop in conjunction with sequence based approach.
- To use bioinformatics techniques for the identification of restriction endonuclease(s).
- To clone and express the gene(s) in a suitable host expression system.
- To purify and characterise the isolated enzyme(s).

# 2 Materials and Methods

## 2.1 Materials

Instruments and equipment used were provided by the CSIR (Biosciences, Pretoria). Analytical and chemical reagents and kits were purchased from commercial suppliers.

All the *Escherichia coli* strains and vectors used and constructed in this study are listed in Table 2.1 and 2.2, respectively. Buffers, media and solutions used can be seen in Table 2.3 below. Table 2.4 shows the inducers and antibiotics used, how they were prepared and concentration used.

**Table 2.1**: ***Escherichia coli* strains used in this study for overexpression.**

| Name | Genotype | Selective Marker | Supplier |
|---|---|---|---|
| **EPI300™-T1R *E. coli*** | F– mcrA Δ(mrr-hsdRMS-mcrBC) φ80dlacZΔM15 ΔlacX74 recA1 endA1 araD139 Δ(ara, leu)7697 galU galK λ– rpsL nupG trfA tonA dhfr. | Chloramphenicol | Epicentre (USA) |
| ***E. coli* DH5α** | fhuA2 Δ(argF-lacZ)U169 phoA glnV44 Φ80 Δ(lacZ)M15 gyrA96 recA1 relA1 endA1 thi-1 hsdR17 | None | Zymo Research (USA) |
| ***E. coli* BL21 (DE3)** | fhuA2 [lon] ompT gal (λ DE3) [dcm] ΔhsdS λ DE3 = λ sBamHIo ΔEcoRI-B int::(lacI::PlacUV5::T7 gene1) i21 Δnin5 | None | Sigma®-Aldrich, Merck, Germany |
| **Rosetta™ (DE3)pLysS** | F- ompT hsdSB(rB- mB-) gal dcm (DE3) pLysSRARE (CamR) | Chloramphenicol | Sigma®-Aldrich, Merck, Germany |

**Table 2.2: Vectors and constructs used in this study**

| Name | Description | Size[a] | Supplier |
|---|---|---|---|
| **pCC2FOS™** | Copy controlled vector, linearized and dephosphorylated at Eco72I restriction site. Requires EPI300™-T1R *E. coli* strain for high copy number induction. (Chlor$^R$) | 8181 bp | Epicentre, Whitehead Scientific (SA) |
| **pET-30b (+)** | High copy number plasmid DNA expression vector that includes the T7 promoter and requires IPTG for induction. Proteins can be expressed. Kan$^R$ | 5422 bp | Novagen (USA) |
| **pETDuet-1** | Vector that's designed for co-expression of two target genes. It includes the T7 promoter and requires IPTG for induction. Proteins can be expressed. Amp$^R$ | 5420 bp | Novagen (USA) |
| **pFos_endo_** | pCC2FOS$^{TM}$ derived constructs identified from the endonuclease metagenomic library. (Chlor$^R$) | ± 22 kb | This study |
| **pET30_Endo8** | pET-30b (+) derived expression vector harbouring *Nde*I and *Xho*I restricted endo8 fragment obtained from sequence data. (Kan$^R$) | 1386 bp | This Study |
| **pET30_Endo20** | pET-30b (+) derived expression vector harbouring *Nde*I and *Xho*I restricted endo20 fragment obtained from sequence data. (Kan$^R$) | 2412 bp | This Study |
| **pETDuet-1_Endo52** | pET-Duet-1 derived expression vector harbouring *Bgl*III and *Xho*I restricted endo52 fragment obtained from sequence data. (Amp$^R$) | 5844 bp | This Study |

[a] = Length of DNA given in base pairs (bp) or kilo base pairs (kb), where the size of the constructs given in this table exclude the vector backbone and only refers to the insert fragment.

Amp$^R$ = Ampicillin resistance. Kan$^R$ =Kanamycin resistance. Chlor$^R$ = Chloramphenicol resistance. IPTG-Isopropyl β-D-1-thiogalactopyranoside

**Table 2.3: Buffers, media, and solutions used in this study.**

| Name | Composition | pH |
|---|---|---|
| **Media** | | |
| **SOC Medium (1 L)** | 2% Tryptone, 0.5% Yeast Extract, 10 mM NaCl, 2.5 mM KCl, 10 mM $MgCl_2$, 20 mM Glucose | |
| **Luria Broth (Agar) (1 L)** | 1% (w/v) Tryptone, 171.11 M NaCl, 0.5% (w/v) Yeast Extract, (1.5 % (w/v) Bacteriological Agar) | 7 |
| **Buffers** | | |
| **1X TAE Buffer (1 L)** | 40 mM Tris, 20 mM acetic acid, 1 mM EDTA | |
| **10x SDS Buffer (1 L)** | 124 mM Tris, 960 mM Glycine, 1 % (w/v) SDS | |
| **2x SDS sample buffer** | 0.125 M of Tris, 20% (v/v) Glycerol, 4% (w/v) SDS, 0.2% (w/v) Bromophenol Blue, 200 mM DTT | |
| **TE Buffer** | 10 mM Tris-HCl, 1 mM EDTA | |
| **Transfer Buffer** | 50 mM Tris, 40 mM Glycine and 20% Methanol | 8.3 |
| **1x TBST with milk** | 100 mM Tris, 154 mM NaCl & 0.1% Tween-20, with 5% Milk | 7.5 |
| **Tris-HCl (1 L)** | 1 M Tris | 7 |
| **1x LEW Buffer** | 50 mM $Na_2H_2PO_4$, 300 mM NaCl | 8 |
| **1x Elution Buffer** | 50 mM $NaH_2PO_4$, 300 mM NaCl, 250 mM imidazole | 8 |

| Solutions | |
|---|---|
| **Coomassie Staining Solution I (2 L)** | 10% (v/v) acetic acid, 0.125% (w/v) Coomassie brilliant blue G, 25% (v/v) isopropanol |
| **Coomassie Staining Solution II (2 L)** | 10% (v/v) acetic acid, 0.003% (w/v) Coomassie brilliant blue G, 10% (v/v) isopropanol |
| **Coomassie Staining Solution III (2 L)** | 10% (v/v) acetic acid, 0.0003% (w/v) Coomassie brilliant blue G |
| **De-staining solution (2L)** | 10% (v/v) acetic acid<br>5% (v/v) isopropanol) |

* Unless stated otherwise, all buffers were prepared in distilled water

**Table 2.4: Antibiotics and inducers used in this study.**

| Antibiotic | Preparation |
|---|---|
| **Ampicillin** | Dissolved in distilled water at final concentration of 100 $mg.ml^{-1}$ and filter sterilised |
| **Chloramphenicol** | Dissolved in 99.9% ethanol at final concentration of 34 $mg.ml^{-1}$ and filter sterilised |
| **Kanamycin** | Dissolved in distilled water final concentration of 50 $mg.ml^{-1}$ and filter sterilised |

| Inducer | Preparation |
|---|---|
| **IPTG** | 0.1 M final concentration and filter sterilised |

## 2.2 Standard techniques

### 2.2.1 Chemical transformation

Commercially acquired hemically competent cells of MAX Efficiency® DH5α™ (Invitrogen, Carlsbad, California, United States), One Shot® BL21(DE3) (Invitrogen, Carlsbad, California, United States) and Rosetta™ (DE3) pLysS (Novagen, Sigma®-Aldrich, Merck, Germany) were allowed to thaw on ice for 5 – 10 min before transformation. A volume of 50 µl of the cells was transferred into 1.5 ml Eppendorf tubes on ice and 2 µl of the plasmid DNA was mixed with the cells gently and was left on ice for 10 min. The transformation mixture was then incubated at 42 °C for 45 sec, followed by placing on ice for 2 min. Following this, 500 µl of SOC medium was added immediately and the tubes were incubated at 37 °C for 60 min with shaking at 150 rpm. After the 60 min incubation period, 100 µl of the cells were plated for overnight growth on LB agar with respective antibiotics specific to that plasmid.

### 2.2.2 DNA purification and plasmid DNA isolation

Purification of the DNA was executed utilising GeneJet$^{TM}$ Gel Extraction Kit (Thermo Fisher Scientific, Waltham, MA USA). Plasmid DNA was isolated using GeneJet MiniPrep (Thermo Fisher Scientific, Waltham, MA USA) following the manufacturer's instructions with slight adjustments. The adjustments included two washes with 600 µl of the washing solution. Elution buffer (30 µl) was used to elute the bound DNA.

### 2.2.3 DNA quantification

A NanoDrop 2000 Spectrophotometer (Thermo Fisher Scientific, Waltham, MA USA) was used to determine the purity and concentration of the DNA. This was achieved through measuring the absorbance at 280 nm, 260 nm and 230 nm, where one absorbance unit at 260 nm is equal to 50 µg dsDNA.ml$^{-1}$. According to Marmur (1961), the DNA solution is considered pure when the A260 nm to A230 nm ratio is between 1.8 and 2.3 and A260 nm to A280 nm ratio between 1.5 and 2.0.

### 2.2.4 Separation of DNA by agarose electrophoresis

DNA fragments and plasmids were separated in 1 % (w/v) agarose gels prepared in 1 × TAE buffer containing 20 µl.100 ml$^{-1}$ of PronoSafe solution (Sambrook and Russell, 2001). Prior to electrophoresis, the samples were mixed with 6 × DNA loading dye. A 1Kb Gene Ruler molecular weight marker from Fermentas (Waltham, MA USA) was used. Electrophoresis was

performed in 1 × TAE buffer at 120 V using a buffer chamber apparatus (PEQLAB Biotechnologie GmbH, Germany). The agarose gels were visualized under ultraviolet light, 300 nm and photographed with a ChemiDoc MP System digital imaging system (Bio-Rad, Universal Hood III, California, USA).

### 2.3 Sample collection and metagenomic DNA extraction from the soil

The sample source used for metagenomic DNA ($_{met}$DNA) extraction was a soil sample from the Kogelberg Biosphere Reserve, obtained under Cape Nature permit no: 0052-AAA008-00017 (Western Cape Province, South Africa) which is a known environment for its richness in biodiversity. Prior to the $_{met}$DNA extraction, the soil sample was kept at 4 °C.

DNA isolation was performed using ZR Soil Microbe MidiPrep™ Kit (Zymo Research, USA), with minor modifications. The modifications included the bead-beating step (0.05 mm diameter) using Gene Disruptor (Scientific Industries, USA) for 40 sec in the presence of lysis buffer. The sample was incubated at room temperature for 2 hr followed by centrifugation (13 000 × *g,* 1 min). The resultant supernatant was transferred to the ZT BashingBead™ lysis/filtration and further centrifuged for 1 min at 10 000 × *g*. The recovered flow through fraction was then transferred into a clean 50 ml tube and Soil DNA Binding Buffer was added at a ratio of 3:1. The sample was then loaded onto Zymo-Spin™ V-E Column/Zymo midi filter™ and centrifuged (10 000 × *g* for 1 min). The bound DNA was pre-washed (300 µl) followed by two washes with soil DNA wash buffer. The bound DNA was then eluted with 150 µl of elution buffer.

### 2.4 Fosmid library construction

The constructed of the metagenomic library was achieved using CopyControl Fosmid Library Production kit (Whitehead Scientific, South Africa). Preparation of the fosmid library was constructed based on the manufacturer's instructions. Briefly, the $_{met}$DNA isolated above (section 2.3) was blunt-ended and 5' phosphorylated according to the kit protocol. The end repaired $_{met}$DNA was ligated into pCC2Fos vector overnight using Fast-Link DNA ligase. The ligation reaction was packaged into lambda extracts which were then used to infect Epi330-T1R *E. coli resistant* strain. The library titre was calculated using the equation 1 below.

$$cfu = \frac{\text{Volume plated (µl)}}{\text{No.of Colonies}}\ x\ Dilution\ Factor \qquad (1)$$

### 2.4.1 Library storage

The constructed metagenomic library was stored by re-suspending the colonies in LB Broth. The re-suspended cells/colonies were then placed in a centrifuge bottle and centrifuged at 9000 × *g* for 10 min. The recovered supernatant was discarded and the pelleted cells were re-suspended with 20 ml of LB broth supplemented with chloramphenicol (12.5 $\mu g.ml^{-1}$) and glycerol (final concentration of 20%, v/v), and aliquoted for cryopreservation and stored at -80 °C.

### 2.4.2 DNA purity quantification and DNA size determination

The NanoDrop 2000 Spectrophotometer (Thermo Fisher Scientific, Waltham, MA USA) was used for the determination of DNA concentration and purity, by measuring the absorbance at 280 nm, 260 nm and 230 nm. DNA was purified using the GeneJet™ gel extraction kit (Thermo Fisher Scientific, Waltham, MA USA) with modifications where necessary. DNA fragments were separated based on their size using 1% (w/v) agarose gel electrophoresis using an appropriately sized ladder (Sambrook & Russell, 2001).

## 2.5 Functional screening of metagenomic library for restriction endonucleases using bacteriophage infection

### 2.5.1 Rehydration of the bacteriophage

An overnight culture containing the host strain (*E. coli* MG1655) (ATCC, Manassas, Virginia) was actively prepared in LB broth. From the overnight culture, 250 µl was added to the freeze-dried vial containing the phage. In parallel, LB agar plates were prepared. Once the plates were dry, 2.5 ml of a 0.5% soft LB agar layer was poured over the solid agar. The soft LB agar overlay contained one or two drops of the host cells (*E. coli* MG1655) grown overnight. The re-hydrated phage was serially diluted (1:10) in LB medium in a 1 ml volume. The prepared dilutions were spotted on the surface of the plates and were incubated overnight at 37 °C. After 24 h incubation, the soft agar was scraped off the surface of the agar plates and centrifuged at about 11,200 × *g* for 30 min. The supernatant was subsequently passed through a 0.22 µm Millipore filter (Sigma®-Aldrich, Merck, Germany) and the filtrate stored at 4°C.

### 2.5.2 Fosmid library screening

The fosmid library constructed as described in Section 2.4 was screened for (restriction) endonucleases and other endonuclease enzymes. This was achieved by plating the library onto LB agar plates supplemented with chloramphenicol (12.5 μg.ml$^{-1}$) containing a lawn of bacteriophage (*E. coli* bacteriophage Lambda W60 (ATCC97537)) (ATCC, Manassas, Virginia). Figure 2.1 shows a schematic of the approach used to screen for restriction endonucleases. Accordingly, colonies that were formed were expected to contain putative restriction-modification systems as a defence against bacteriophage infection. Colonies derived from these plates were replica-plated onto fresh plates. These plates were also preceded with bacteriophage as a confirmation that the colonies were not false positives before the fosmid DNA was extracted. The DNA from these colonies were pooled and sequenced using MiSeq Illumina technique of the next generation sequencing (NGS) platform. The service was provided by Inqaba Biotech (Pretoria, South Africa).

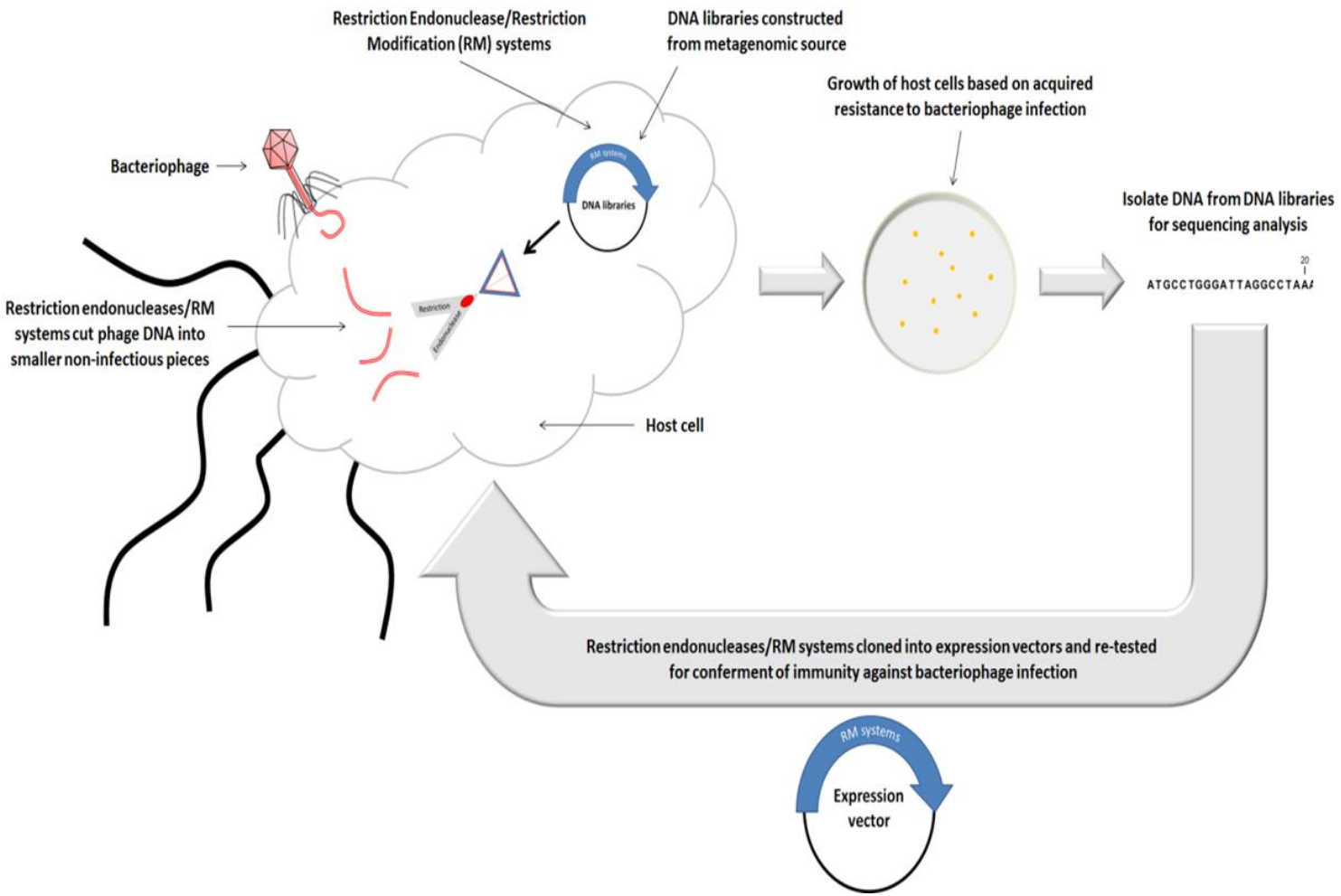



**Figure 2.1: An illustration of the functional screening approach undertaken in this study.**

Endonuclease genes or R-M systems derived from the metagenomic and cloned into a fosmid vector are represented as blue curved arrows. When plated on solid media pre-seeded with a lawn of bacteriophage, E. coli cells which express functional endonuclease(s) produce colonies (yellow dots). Fosmid DNA from these colonies is then isolated and analysed by next generation sequencing.

## 2.6 Sequence analysis and bioinformatics

The nucleotide sequences of the positive clones from the functional screening were obtained using MiSeq Next generation sequencing (NGS) Illumina platform (Voelkerding *et al.*, 2009). Protein sequence identity was used to ascertain the novelty of the enzyme. This technique offers short run times while maintaining long reads and good quality data. The sequences generated were analysed with CLC Bio WorkBench Version 11 full suite (Scholz *et al.*, 2014) and BioEdit (Hall, 1999) software programs. Homology searches were performed using the basic local alignment search tool (BLAST) (http://www.ncbi.nlm.nih.gov/BLAST/), to determine conserved domain of the sequences, the NCBI conserved domain function was used (Altschul *et al.*, 1990). REBASE (http://tools.neb.com/blast/), MG-Rast (http://metagenomics.anl.gov/) and Rast (http://rast.nmpdr.org/) bioinformatics pipelines were also used to identify desired enzyme genes from the sequence data.

## 2.7 Recombinant production of the protein

The gene of interest acquired from the analysis of the NGS data were synthesised and codon optimised into a suitable expression vector in order to recombinantly express the protein for further investigation and characterization. Services for codon optimisation, gene synthesis and cloning were provided by GenScript (Piscataway, USA). The expression system used in this study was the pET expression system which is based on T7 promoters. This promoter is ideal when there is a desire to produce large amounts of protein. The T7 promoter is regulated by the lac operon. *E.coli* (DE3) BL-21 OneShot and Rosetta™ (DE3) pLysS cells were used as expression hosts. The genes were synthesised with a histidine affinity tag to facilitate protein purification via IMAC (immobilized metal affinity chromatography).

### 2.7.1 Protein expression

Three expression constructs were designated the following names: pET30Endo_8, pET30Endo_20 and pETDuet_Endo52 (The numbers are taken from the specific Contigs they were found in and Endo is short for endonuclease, see Results Table 3.3. and Table A2). The constructs were subsequently transformed into commercial aquired chemically-competent BL-21 (DE3) OneShot and Rosetta™ (DE3) pLysS cells followed by plating on LB agar plates supplemented with chloramphenicol (34 $\mu g.ml^{-1}$) and/or kanamycin (50 $\mu g.ml^{-1}$) or ampicillin (100 $\mu g.ml^{-1}$). Fresh colonies were used to inoculate 5 ml of LB broth supplemented with the respective antibiotics and grown overnight. The following day, 1ml of the overnight cultures were

inoculated into 50 ml of fresh LB supplemented with the same antibiotics and was grown at 37 °C with vigorous shaking (200 rpm) until $OD_{600nm}$ reached 0.4 - 0.6. Expression of the recombinant proteins was induced by 0.1 mM IPTG (Isopropyl β-D-1-thiogalactopyranoside) at 17, 25 and 30 $^{o}$C and protein expression was monitored over time by SDS-PAGE. Non-induced controls were also harvested.

### 2.7.2 Immobilised metal ion affinity chromatography (IMAC) purification

Optimal conditions for purification of the protein after expression were taken from expression studies from section 2.7.1. Cells, from an IPTG-induced culture, were pelleted at 10 000 × *g* for 20 min and B-PER (Thermo Fisher Scientific, Waltham, MA USA) was used to lyse the cells. The recovered soluble fraction was purified using immobilised metal ion affinity chromatography (IMAC) purification system. This was done according to the manufacturer's instruction (http://www.mn-net.com/Portals/8/attachments/Redakteure_Bio/Protocols/Protino/UM_ProtinoNi_TED.pdf). The soluble protein fraction was loaded onto the IMAC column packed with Protino Ni-TED resin (Macherey-Nagel, Germany). The bound protein was eluted with elution buffer and Sodium dodecyl sulphate polyacrylamide gel electrophoresis (SDS-PAGE) was conducted to visualise the purified protein.

### 2.7.3 SDS-polyacrylamide gel electrophoresis (SDS-PAGE)

The sodium dodecyl sulphate polyacrylamide gel electrophoresis (SDS-PAGE) was carried out according to the method of Laemmli (1970) using a Bio-Rad Mini-Protean Tetra Cell system (California, USA). The SDS page gels comprise two gels, the resolving gel (12%) and the stacking gel (4%). The gels were prepared with a thickness of 1.0 mm. The gels were prepared freshly before use. Prior to loading the sample on the gel, the samples were mixed with SDS loading dye and subsequently denatured at 95 °C for 5 min. The samples were then loaded onto the gel and electrophoresed under a constant voltage 100 V until the dye migrated into the separating gel; followed by electrophoresis at 140 V until completion. Pre-stained SDS-PAGE protein molecular weight marker (product code: 26619) purchased from Thermo Scientific Life Science (Vilnius, Lithuania) and used.

### 2.7.4 Western blotting

The separated proteins on a 12% polyacrylamide gel(s) were equilibrated for 5 min in transfer buffer, while Immobilon-P PVDF membrane (Millipore, Temecula, CA) was equilibrated in

methanol. After equilibration, the gels were transferred to Immobilon-P PVDF membranes (Millipore, Temecula, CA) using a semi-dry transfer apparatus (Bio-Rad Trans-Blot Turbo, Bio-Rad, California, USA) at 1.3 A, 25 V for 20 min. Membranes were blocked for 60 min with 1 × TBST, followed by a further blocking of 60 min with 1× TBST (5% Milk) with monoclonal anti-poly histidine antibody produced in mouse (Sigma®-Aldrich, Merck, Germany) at 4 °C. All antibody dilutions were 1:2000. The membrane was washed in TBST 5 times for 2 min each time with continuous agitation. Identification of the target proteins was achieved using Bio-Rad Clarity Western Blot ECL Substrate by mixing 1:1 solution enough to cover membrane. The substrate was then added onto the blot, incubated and covered for 5 min at room temperature. The membrane was visualized using digital imager or by exposing it to X-ray film and the ChemiDoc MP System digital imaging system (Bio-Rad, Universal Hood III, Bio-Rad, California, USA) was used.

### 2.7.5 Bradford quantification

The soluble fraction protein concentration was determined using gel quantification. The assay involved loading three different volumes of the eluted protein (diluted/undiluted) on an SDS-PAGE with BSA (20 $mg.ml^{-1}$) standards (10 µl). A standard curve was constructed using six standards 0.0781 µg, 0.156 µg, 0.312 µg, 0.625 µg, 1.25 µg and 2.5 µg. An absolute quantification was conducted using the Image Lab 4.1 Software from BioRad and ChemiDoc MP System digital imaging system (Bio-Rad, Universal Hood III, California, USA). Below in Figure 2.2 is an example of the standard curve that is generated via the ChemiDoc imaging system. Correlation co-efficient values greater than 0.95 were used as indicators of well-prepared BSA standards.

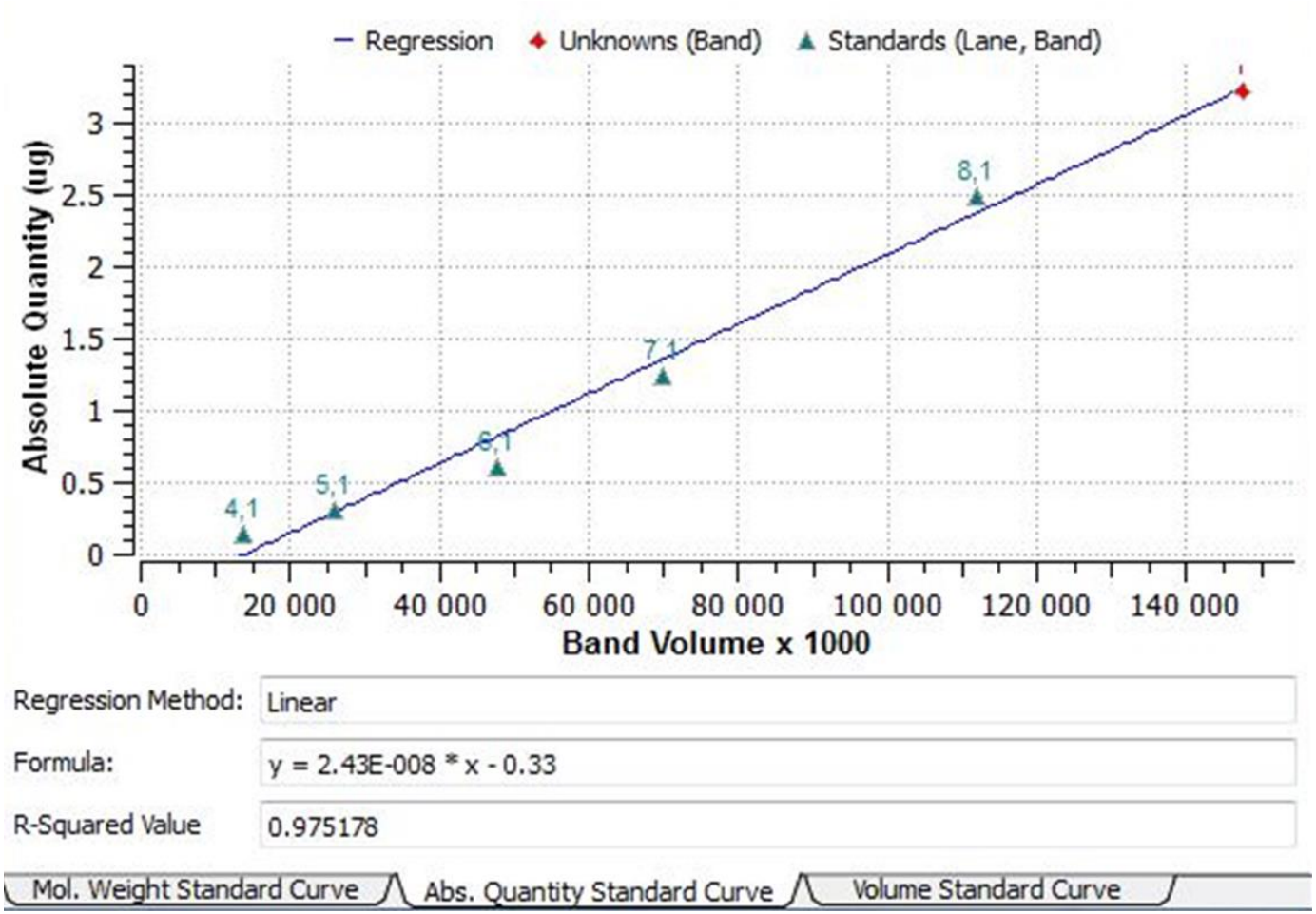



**Figure 2.2 Representation of BSA Standard Curve generated from Chemi-Doc imaging system.**

The known concentrations of the BSA standard were used to determine the concentrations of the expressed and purified proteins. Band volume is the sum of all intensities within the band boundaries.

## 2.8 Characterisation of the enzymes

Functional characterisation of the recombinantly expressed enzymes was conducted. Assays included using the enzyme(s) purified in a restriction endonuclease activity assay. Its functionality was measured against the enzyme's ability to cleave DNA. Various conditions were tested including enzymes stability, temperature and stability profile and buffer preference.

### 2.8.1 Functional characterisation of the restriction enzymes

Functionality of the purified enzymes (restriction endonucleases) was achieved through restriction digestion reactions, using the purified enzyme for activity against λ DNA, pUC19, Human Genomic DNA (Sigma®-Aldrich, Merck, Germany) and Genomic DNA from an Albino Mouse (Sigma®-Aldrich**,** Merck, Germany). Cleavage of the DNA was performed at various time intervals (5, 10, 15, 30, 45, and 60 min) under different buffer compositions (Table 2.5). Reactions were prepared according to the Sambrook and Russell method and details of the reaction mix can be found in Table 2.6. Restriction reactions were stopped and prepared for agarose gel electrophoresis. Where restricted DNA was to be further used directly for enzymatic reactions (cloning or ligation), the digestion was either stopped by heat shock (65 °C, 5 min) or by gel purification.

**Table 2.5: List of buffers used and their compositions.**

| Buffer | Components | pH |
|---|---|---|
| **1** | 50 mM NaCl<br>10 mM Tris-HCl<br>10 mM $MgCl_2$<br>100 $\mu g.ml^{-1}$ BSA | 7.9 |
| **2** | 50 mM NaCl<br>10 mM Tris-HCl<br>10 mM $MgCl_2$<br>1 mM DTT | 7.9 |
| **3** | 100 mM NaCl<br>50 mM Tris-HCl<br>10 mM $MgCl_2$<br>100 $\mu g.ml^{-1}$ BSA | 7.9 |
| **4** | 50 mM Potassium Acetate<br>20 mM Tris-Acetate<br>10 mM Magnesium Acetate<br>100 $\mu g.ml^{-1}$ BSA | 7.9 |
| **5** | 33 mM Tris-Acetate<br>10 mM Magnesium Acetate<br>66 mM Potassium Acetate<br>0.1 $mg.ml^{-1}$ BSA | 7.5 |
| **6** | 10 mM Tris-HCl<br>10 mM $MgCl_2$<br>100 mM KCl<br>0.1 $mg.ml^{-1}$ BSA | 7.5 |
| **7** | 50 mM Tris-HCl<br>10 mM $MgCl_2$<br>100 mM NaCl<br>0.1 $mg.ml^{-1}$ BSA | 7.5 |
| **8** | 10 mM Tris-HCl<br>10 mM $MgCl_2$<br>50 mM NaCl<br>0.1 $mg.ml^{-1}$ BSA | 7.5 |
| **9** | 10 mM Tris-HCl<br>10 mM $MgCl_2$<br>0.1 $mg.ml^{-1}$ BSA | 7.5 |

**Table 2.6: Preparation order of the reaction mixture**

| Component | Volume | | |
|---|---|---|---|
| | Control | Plasmid DNA | Genomic DNA |
| **Nuclease-free Water** | x µl | x µl | x µl |
| **10X Buffer** | 2 µl | 2 µl | 2 µl |
| **SAM** | x µl (80 µM) | x µl (80 µM) | x µl (80 µM) |
| **ATP** | x µl (1-10 mM) | x µl (1-10 mM) | x µl (1-10 mM) |
| **DNA** | 2 µl (1 µg) | x µl (~2 µg) | x µl (~5 µg) |
| **Enzyme** | x µl | x µl | x µl |
| **Total Volume** | 20 µl | 20 µl | 20 µl |

## 2.9 Scale-up of Endo_8 production

### 2.9.1 Fed batch fermentation

Three flasks (1 l) with 100 ml of autoclaved LB media, with 34 µg.ml$^{-1}$ of chloramphenicol (Sigma®-Aldrich, Merck, Germany) and 50 µg.ml$^{-1}$ of kanamycin (Sigma®-Aldrich, Merck, Germany) added to them, were inoculated with 1 ml of cryopreserved cell suspension containing Rosetta™ (DE3) PlyS (Endo_8). The inoculated flasks were incubated overnight (12 to 16 h) at 37 °C, 200 rpm. Post overnight growth, the inoculum was used to inoculate 2 l fermenters containing 900 ml of media. Three 2 l fermenters (Infors HT, Switzland) with 900 ml GMO 20 medium were inoculated with 100 ml inoculums ($OD_{600}$). The batch fermentation parameters and recipes can be found in Table A1. The $OD_{600}$ readings were done to monitor the cell growth and glucose concentration was measured by Accutrend$^{@}$ (Boehringer Mannheim). The fermenters were harvested 3 and 8 hr after induction.

#### 2.9.1.1 Analysis and purification

Dry cell weight was done by quantifying 1 ml of cell suspension that was pelleted at 10 000 × *g* for 10 min and the scientific series 2000 (Scientific, Gauteng, South Africa) oven was used to dry the pellets at 110 $^{\circ}$C overnight (18hours). To quantify the product, three samples (1 ml cell suspension) were pelleted at 10 000 × *g* for 10 min and kept at -80$^{o}$C until analysis. Pellets (1 ml) were lysed using 1 × Lysis-Equilibration-Wash (LEW) buffer. The soluble proteins were then recovered by ultra-centrifugation (30 000 ×g, 10 min 4 $^{o}$C). The SDS-PAGE was carried out according to the section 2.7.3 using a Bio-Rad system (Bio-Rad, California, USA). The harvest (350 ml) done at 3 and 8 h post-induction was centrifuged and the pelleted cells stored at -80 $^{o}$C until further analysis. The pellets were resuspended in 1 × LEW buffer and were lysed using the Cell Breaker (Constant Systems Limited, Daventry, UK) (Pressure and Impact at 4 $^{\circ}$C and 20 kpsi). The soluble fraction was centrifuged at 10 000 × g, 30 min 4 $^{o}$C. Supernatant was kept aside and the same pellet was re-suspended with equal volume of 1 × LEW buffer and centrifuged at 10 000 × g, for 30 min at 4 $^{o}$C. The recovered soluble fractions were combined and centrifuged again to obtain a clarified supernatant, which then was purified.

The recovered soluble fractions were purified using immobilised metal ion affinity chromatography (IMAC) purification system using the Äkta Avant 150 Fast Protein Liquid Chromatography (FPLC) (GE Healthcare Life Sciences, USA). The soluble protein fraction was loaded onto the HiScale 50 column (GE health, USA). The column bed volume was 240 mL and

flow rate was 20 ml.min$^{-1}$. The column was pre-equilibrated with 3 column volumes (CV) of 1 × LEW buffer at a flow rate of 20 ml.min$^{-1}$. The sample was loaded onto the column at a flow rate of 10 ml.min$^{-1}$. Unbound proteins were washed with 4 CV of 1 × LEW buffer at flow rate of 20 ml.min$^{-1}$. Bound proteins were eluted in an isocratic elution with volume of 660 ml of elution buffer at 3 CV. The purity of the collected fraction was analysed by SDS-PAGE (Section 2.7.3).

# 3. Results

## 3.1 Metagenomic DNA extraction and fosmid library construction

Metagenomic DNA ($_{metg}$DNA) in this study was extracted from a soil sample collected from the Kogelberg Nature Reserve, Western Cape using both chemical and physical steps. The metagenomic DNA extracted was further used to construct the fosmid library. Extraction procedures are detailed in section 2.3 of this dissertation. The extraction of DNA from the soil yielded a total of 4 ± 0.7µg.g$^{-1}$. These results are from a triplicate extraction experiments. The extracted $_{met}$DNA was of good quality as evidence by the $A_{260/280}$ ratio of 1.76. Determination of the size of the extracted DNA was achieved through electrophoresis (0.8 % agarose), the band on the gels reflected to be larger than 11 Kb (Figure 3.1). A fosmid library was constructed using the CopyControl pCC2FOS$^{TM}$ vector according to methods described in Section 2.3 of this dissertation. The fosmid library size resulted in ±1.83x10$^{5}$ cfu's (Equation 1). Fosmid DNA was extracted from a selected number of colonies. The fosmid library contained 5,849 independent clones in duplicate with an average insert size of ±27kb.

## 3.2 Endonuclease functional screening

A plate-based functional approach (described in detail in Section 2.5.2) was developed and utilised to screen the Kogelberg Biosphere soil fosmid library. Briefly, the procedure involved introducing lambda bacteriophage into the fosmid library cells by plating on plates pre-seeded the plates with the phage. Clones that survive would grow indicated that they have a mechanism that protects them or confers immunity from phage infection. Figure 3.2 illustrates a set of representative plates from the screen. The control plate did not exhibit any colony growth while colony growth was observed on which the fosmid library was plated; those clones that were immune to bacteriophage infection.

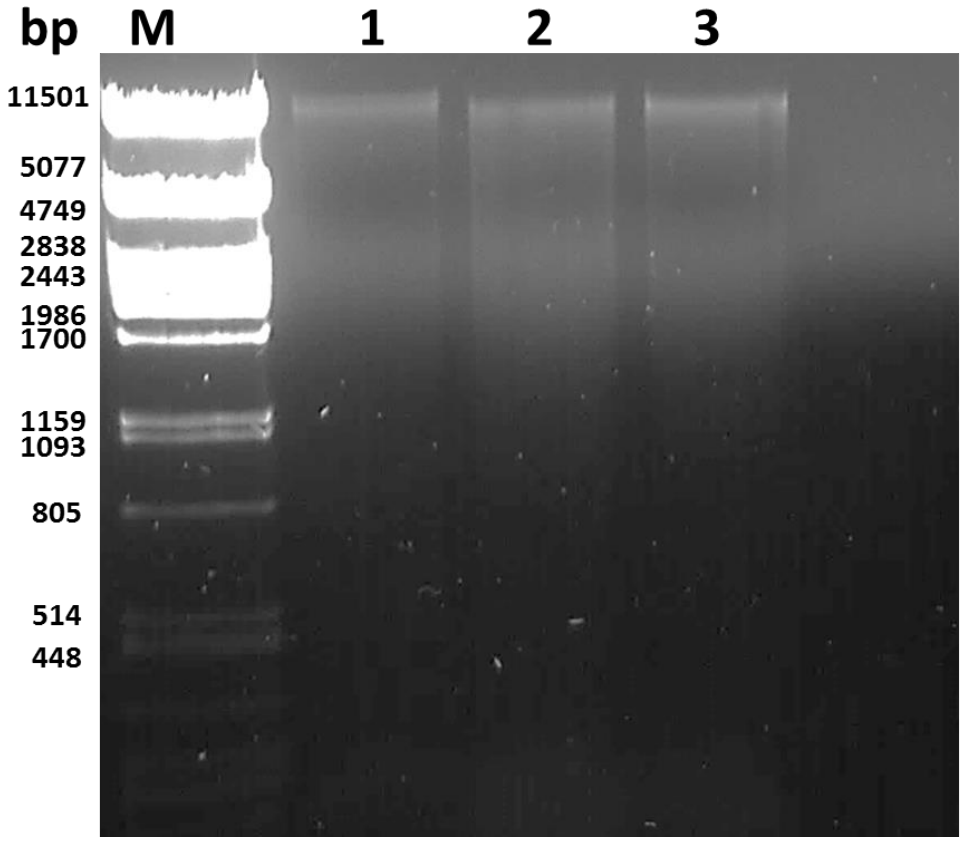



**Figure 3.1: Agarose gel electrophoresis of extracted metagenomic DNA from soil.**

M is a λ *Pst*I DNA marker, lane 1-3 extracted DNA of high molecular weight from the same sample. Extraction of the same sample was conducted in triplicate.

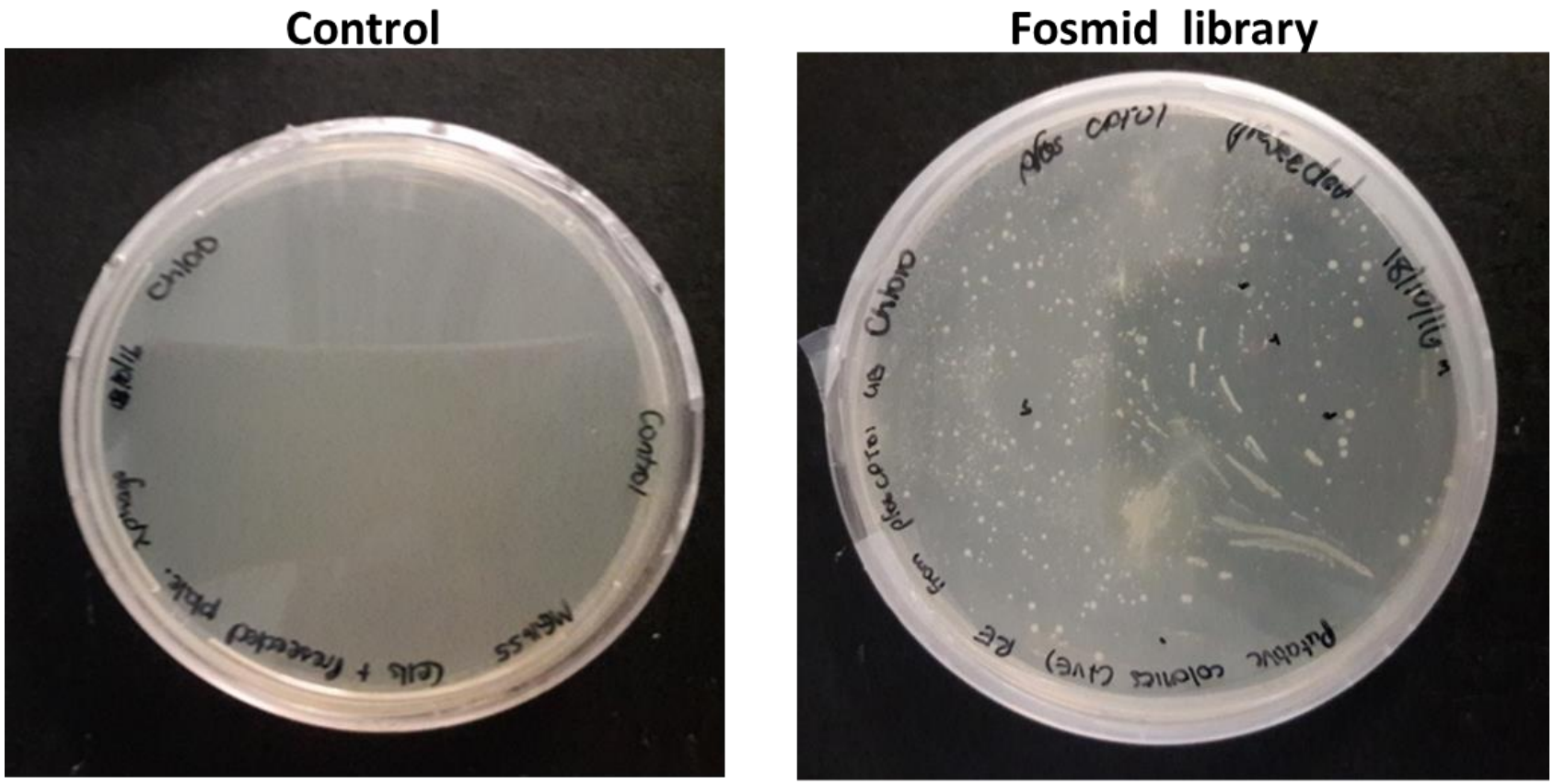



**Figure 3.2: Plate-based screening approach for Restriction endonucleases (REs).**

This technique allows for simple selection of clones that are immune to bacteriophage infection. Plates are pre-seeded with phage and *E. coli* transformed with the fosmid library is then plated on it. Cells that survive were immune to phage infection.

A few selected positive colonies from the screening were replica plated on pre-seeded plates to exclude possibility of false positive appearance. All the selected colonies grew on several cycles of pre-seeded plates and were then inoculated in LB Broth supplemented with chloramphenicol for isolation of the fosmid before sequencing. Fosmid DNA was isolated and analysed on an agarose (1%) gel and fosmid DNA can be seen and of a size bigger than 11 kb (Figure 3.3).

## 3.3 Sequence data analysis

Positive colonies derived from the screening plates were inoculated into fresh LB media supplemented with antibiotics and the fosmid library DNA contained in these colonies was isolated. The DNA from these colonies was sequenced using next generation sequencing (NGS) and the Illumina platform via a service provider as detailed in Section 2.6.

CLC Bio Workbench 6.2 (www.clcbio.com) was used to assemble the data. Table 3.1 shows the assembled sequence statistics obtained from CLC Bio Workbench. A total of 1942 contigs were generated. The average number of reads per contig was 8561 bp, with the smallest being 1002 bp and the largest being 802935 bp. NCBI Blast was used to deduce homologues to inform annotation of the open reading frames (ORFs). Table 3.2 gives an overall summary of the number of sequence that show annotations to the scope of the study, which was to functionally screen a fosmid library for restriction endonucleases and other endonucleases and Table 3.3 provides sequences with similar annotations with sequences from the NCBI database. The results obtained from the NGS sequencing of positive clones show an enriched subset of sequence data encoding a number of restriction-modification systems as well as restriction endonucleases. A comprehensive record of sequences obtained can be found in Table A2. However, more than 200 unknown/uncharacterised and hypothetical sequences obtained are not reported here. These may represent entirely new endonuclease genes or other novel mechanisms for immunity to bacteriophage infection and are viewed as a substantial opportunity for future development.

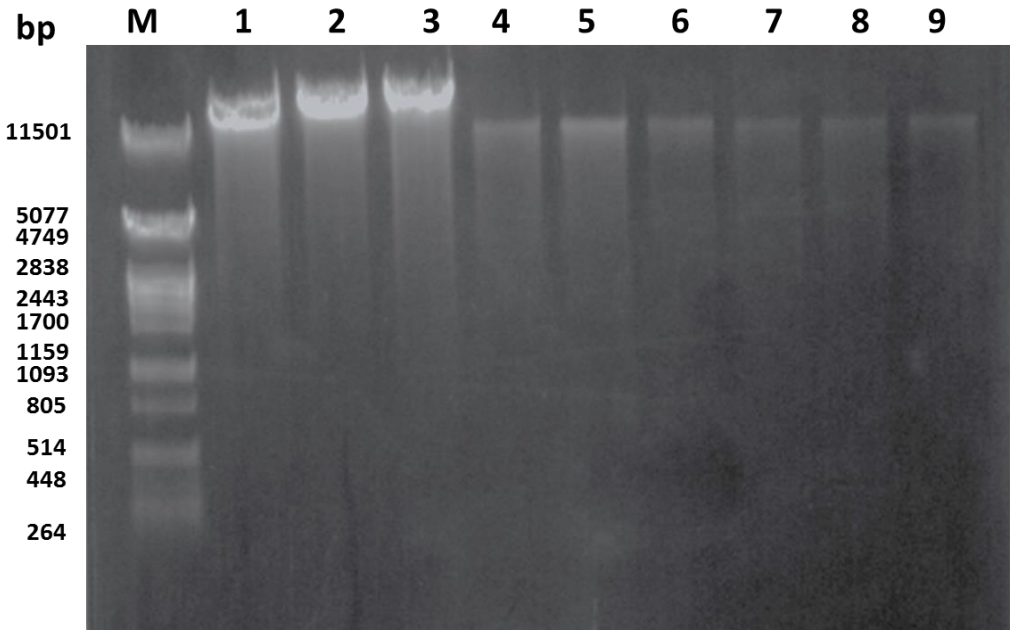



**Figure 3.3: Fosmid DNA extracted from positive clones for further investigation.**

M is a *Pst*I λ DNA Marker. Lanes 1 - 9 are selected positive colonies. Lane 1 - 3 the colonies were taken from a different plate from lane 4 - 9.

**Table 3.1: Summary Statistics of assembled sequences obtained from CLCBio Workbench.**

CLCBio generated statistics of the data assembly and annotation of the sequence data from the positive clones sequenced. The data generated revealed numerous contigs accounting for a total of 16,624,688 bp

| | *Count* | *Average Length* | *Total Bases* |
|---|---|---|---|
| **Reads** | 17 401 926 | 227.94 | 3 966 648 334 |
| **Matched** | 17 178 015 | 228.50 | 3 925 178 941 |
| **Not Matched** | 223 911 | 185.20 | 41 469 393 |
| **Contigs** | 1 942 | 8 560 | 16 624 688 |
| **Reads in pairs** | 11 037 878 | 237.32 | |
| **Broken paired reads** | 6 140 137 | 229.49 | |

**Table 3.2: Summary of the NCBI BLAST results for the positive clones from sequence data.**

The results in the table indicate the success of the screening technique for novel (restriction) endonucleases developed in this study. Positive clones selected for sequenced and assembled using CLCBio workbench, ORF generated revealed a number of positive hits (endonuclease/endonuclease related).

| **Closest hit annotation** | **Total** |
|---|---|
| Endonuclease | 16 |
| Methyltanferase/Methylase | 54 |
| HNH (motif) | 7 |
| DEAD/DEAX | 24 |
| Restriction Endonuclease/ Enzyme | 12 |
| **Total** | **113** |

**Table 3.3: Amino acid of selected positive clones from the functional screening with similar annotation sequences from NCBI database.**

| Closest hit Annotation (NCBI Database) | Amino Acid Sequence | Query Cover (%) | Identity (%) |
|---|---|---|---|
| **RE** | MLAKQLRRDLRFDPEHQFLALLLGLHCLGRELGDIGHETHARRHDKLRRCVEYETNIGTDCDTSSLSR REKKCHVNVGEVDEIEYPPTGGQDLAGLRDAILHTPVAWRSEGAVIDIGNNTFNRRICGNDGSLCIDD LGLRSADRRIGSGKRSPGCSSSRPRPLCRRPVVVERLLGGHVGLDQLLGPGKFSFRGILFSFALSDH RGSRFFFRLPLGEQAFRCIHTDHGALTRGFGLPPLCLQLLGVHASQHLSGGDETPFVDEDFLDPPGR LAETSISVASMRPLPPTIPAGRPVPSYIRQP | **16** | **44** |
| **RE** | MTLTMKESSAIAALAQLLYDFLPGSGNNQTAFPLAAHKAGVGEYWQPGSKLPSLTQLLTLTLEWKRG QFCPLILEIVRQSMTWRGRRDPLKREEVDQLNKLLPGVGFRIPELLDPDFLDTLAGPPAVQPAPVGSG KPVIDTNRLAELSKRLSDLSAISPQERGFAFERFLYDLFDVYGLAPRASFRPRTGEQIDGSFDLDGDTY LLEAKWHSNPTPAADLHVLSSKLNSRPIWSRALFISYSGFSPDGLEAFNRGKNSLICMDGYDLYETISR ELSLGQVIATKARRAVETGLCHVSVRDLF | **100** | **78** |
| **RE** | MPVVIVENDTSQWEDETGAVYHFPKRYQAWLAQGTEVIYYKGRIKDKAFASVRLSTDPHYFGKARIG QVYADRRSDKGDLFALIENFTPFEDAVPSKIDGDYLETIPASRMSNYWRDGVRPISQSDYDAILSHATL LPSRADAFVPDTEDDPLTFESASEGSKTSYFGTRYERRKDLRVKAIAIHGLDCKACGFDFEEAYGEHA KGFIHVHHVVPISDFGGEKAVNPETDLVTLCANCHAVVHRKRDKTLSVDELKGMLRGRWVIESQ | **100** | **91** |
| **RE** | MPGIRPKKRIRQTSNLQAQHTIETIMDYDFKTLSPEDFERLIGDLYYAETKVLPQSFKSGKDGGIDLLVT DGNGHEKVIIQCKRYEPSAIAALKKAMQKEKGNNLDKLRPPRYILATSVKLSPQNKKNLQKDLHPWIR DIGDIWGLDDINARLRLNEDIEKKHIKLWISSTAVLEKILNHNILSITDITIEKIRQSFAKLVIHSAFDECYRK LEHSHSCIITGNPGIGKTTLANLLLCRYIKEGFTPIVATNGIHEIFGLIKSQEKNKAIIYYDDFLGATRYNEL KFSKNEDAELLTLLDHAKRSDHLRFIMTSRDYIIEDAKSNHRHFQDYADQITRHTIEVGDYTKLHRAKIL YNHLFFSDLPKEKIKHLIESKIYAEIINQEYFIPRVIATICKDANSHSLCNSEFIDYIRQEIANPVSIWQHPF ENEISATARLVLLTTWTFGGTTTTSCLKKIITELQPEHERYDSNLKLNKALKELSANFITLTNMLPKWEG DDPQVIVKFQNPSIEDYINGLIQASPDLLTPKTIKFFKQFENLSINLPSTRHHTSNLTRLIVDILDRFPEIE RTETGRVITTEDNRQLYNSHDTICIADRTISYLKLLLIKLRTPPNETQTAERITTSTGWLELMGGHTLKEF DSYGVERLVQWLSNNLTAIPGVLEKKITQSLSEAISIIANNINAWCTSLRAVSCIATCISHLSLTISESTLS SLVNAALKHGRTAIFSDRSEYPFAYSEMLAISNAISHVEIEKIALSLSHGSVKAEHTETPKTLEKPQHEG DEVNMDLDHYYKTLLHSLV | **62** | **41** |
| **HNH** | MPVVIVENDTSQWEDETGAVYHFPKRYQAWLAQGTEVIYYKGRIKDKAFASVRLSTDPHYFGKARIG QVYADRRSDKGDLFALIENFTPFEDAVPSKIDGDYLETIPASRMSNYWRDGVRPISQSDYDAILSHATL LPSRADAFVPDTEDDPLTFESASEGSKTSYFGTRYERRKDLRVKAIAIHGLDCKACGFDFEEAYGEHA KGFIHVHHVVPISDFGGEKAVNPETDLVTLCANCHAVVHRKRDKTLSVDELKGMLRGRWVIESQ | **100** | **97** |
| **HNH** | MGPDPTVMAFDDATNTGQADPCAFEVFHAVQALEHAEQFAGVGHVEAHAVVANADLGFARVLHGA DADARRGAPTCVFDGVGQQVVQGHVDERRVTDHLGQLGDVPDNFAVLVVGRQFAANGLDQGVEV DLGQAQGRAAHLREVEQVVDQASGKMRGFLDVLQKAPAALAEALALDFAEQFGVAGNMAQGCAQV MGHAVGKRFELLVRATQFASQLRQFLGLAQDNPQHCRAQLLHAFDDQRVPGFAVAAQFFLPAIETVP RVEVALRPDLFVRLPGPVHRAHLLGTEPQQVVRVDLRNGHCQHTACMGRELGQFIDVAAQVVAVEH ALFATLG | **35** | **29** |
| **HNH** | MNKAHTTGYVAPLGRYVPPVLFNPYTGEPRDARDIASDPKGVLIVPPGAQLAATNNPAAAPVVLPEPA YTLRVRGVFQDTTPAANAFGIPDGEHKLFTEPQVRALLAGVSAPAAEGSEIRWVSNGLIGARPTTHDL REAIAAAEDCQRCDCTDCTKAMTLPLKLVEAINTAVGSNEWQGDSVVTDLLEPACKVVANLTHRAQA DQKARMKIAAALGHEGVNFAWSYLTGAIKELVKADGEHLELQPQALAFQQRVQPWMMACFGPEISA DRIERNHRFLEEALELVQSCGSTASEAHQLVDYVFARPVGDPMQESGGVMVTLAALCLASGLDMHA CGEAELARIWTKVEAIRAKQAAKPKHSPLPMYAPQTQDALRLAFVLSEIRRDGMDALGAALHDADGF PLDHKASLEAIDRAANAAASGK | **37** | **57** |
| **Endonuclease** | MTTAKKKKTPARKRGKASRTAGNSLSRIRRFALSAGAAALASFQIASCSFNPSVSAERLLGQALAALHI PALDALGQTLQAFRKNGWPDLDDLRSGSGSGAHTPSVGGKVDKADVTATALPTHFARCPQFFPGGK APALQLQPRERELCFSGFAVLHNGSTKTPVFVAERLNRQMLQQAQGLQRSDKFYADARLPRAERSE LDDYKRSGFSRGHMAPAADMSTPEAMAQSFSLANMVPQNQVHNAGAWSQVEQATRKYALRARGD VFVFTGPVFNKNASTIGESKVAVPDYLFKLVYDASTGKSWVYWQANSADTRMGPPISYEEFTRRTGM PLLSAVHLPQA | **100** | **100** |
| **Endonuclease** | MKFVTAKTPLAAAVGLMVATSVWAKVPANEAEQLGKELTCVGAIKAGNKEGTIPEFTGKWVGAPAGV AHVQSSGKHPVDIYADEKPLFVITAENMEKYGDKLSAGQKAMFQKYSKTMQMPVYKGHRDFRYTDE VCAVLKKNALESEVIDNGMGIKGSFGAINFPIPKTGQEVIWNNLLPTRAYTEAITRDMANVLSDGSMSF GRMQNLNLDMVNKPEMLGKPVEGVMAYTRTRTLAPEREKGGVTHSVEPVNFGKDKRLAWSYDPGT RRVRQVPEYGFDQPMAGTGGKMTIDSDRLFNGSPERYNWKLLGKKEMYIPANNYKIHQPTVKYADL LKPGHANPDFMRYELRRVWAVEGTLKDGFRHVYGKRVLFVDEDTGQAVAADMYDARGQLWQHAFI NYYYSFDIKAWHAGTSFYHDLNSGGYMGYNLFQERPQGPILNKGDLVPAMFTPEAARNAGN | **97** | **65** |
| **Endonuclease** | MNAVERHFMQRDRRAESGRWCGFLAVLDAHEVAVVAEQVGQAAGMGAADRVEQGMQRALAQSG LQLALPVGVVVVEHPAHTEGLERGVVVLAGTGPHAQATLRGQLREVEAHRATGADHQHVAAGTFGY LGQGLPGGQGGAGHGGRGCVAERAGDMHDQACIEQAVLGKTAVTRQRLVVGDAAAQRHVDACAN RDYHTGAIHARDRALLPGRVAPLADFPVHRVEGNGAVGHQHLTGAGLWHGFTQQLQAVEAGVWGP GPGLMIGWHGAGPYETGVGGRLAGIGPGKQVR | **27** | **24** |

## 3.4 Recombinant expression, purification and biochemical characterisation

### 3.4.1 Gene synthesis

Analysis of the data generated from the functional screening and sequence analysis of the positive clones sequenced via NGS was conducted as described in section 2.6. Three ORFs that encoded (restriction) endonucleases were selected for further studies. The amino acid sequences for the selected genes can be found in Table 3.3 and Table A2. Sequences that had homology to restriction endonucleases were also detected (Table A2 and Table A3). The three ORFs, designated endo8, endo20 and endo52 had a length of 1386 bp, 2412 bp and 5880 bp respectively. Two expression vectors were used: pET30b(+) and pETDuet-1 (Novagen, pET system). The pETDuet-1 vector was chosen to co-express one restriction endonuclease gene with its MTase encoding gene. Restriction enzyme digests of the plasmids pET30b(+)Endo8, pET30b(+)Endo20 and pETDuet-1Endo52 harbouring the respective ORF's were conducted to confirm the presence of the insert(s). Figure 3.4 double digest on both DH5α and Rosetta show a plasmid of 6616 bp and an insert of 1386 bp. Figure 3.5 double digest on both DH5α and Rosetta show a plasmid of 7641 bp and an insert of 2412 bp. While Figure 3.6 and Figure 3.7 double digest on both Rosetta and DH5α cells respectively show a plasmid of 11 196 bp and an insert of 2784 bp and 3096 bp respectively. The quadruple digestions show plasmid and the two insert on the different cloning sites of 2784 and 3096 bp. Vector maps illustrating the design of how the genes were cloned into their respective vectors can be found in the Figure A1-A3.

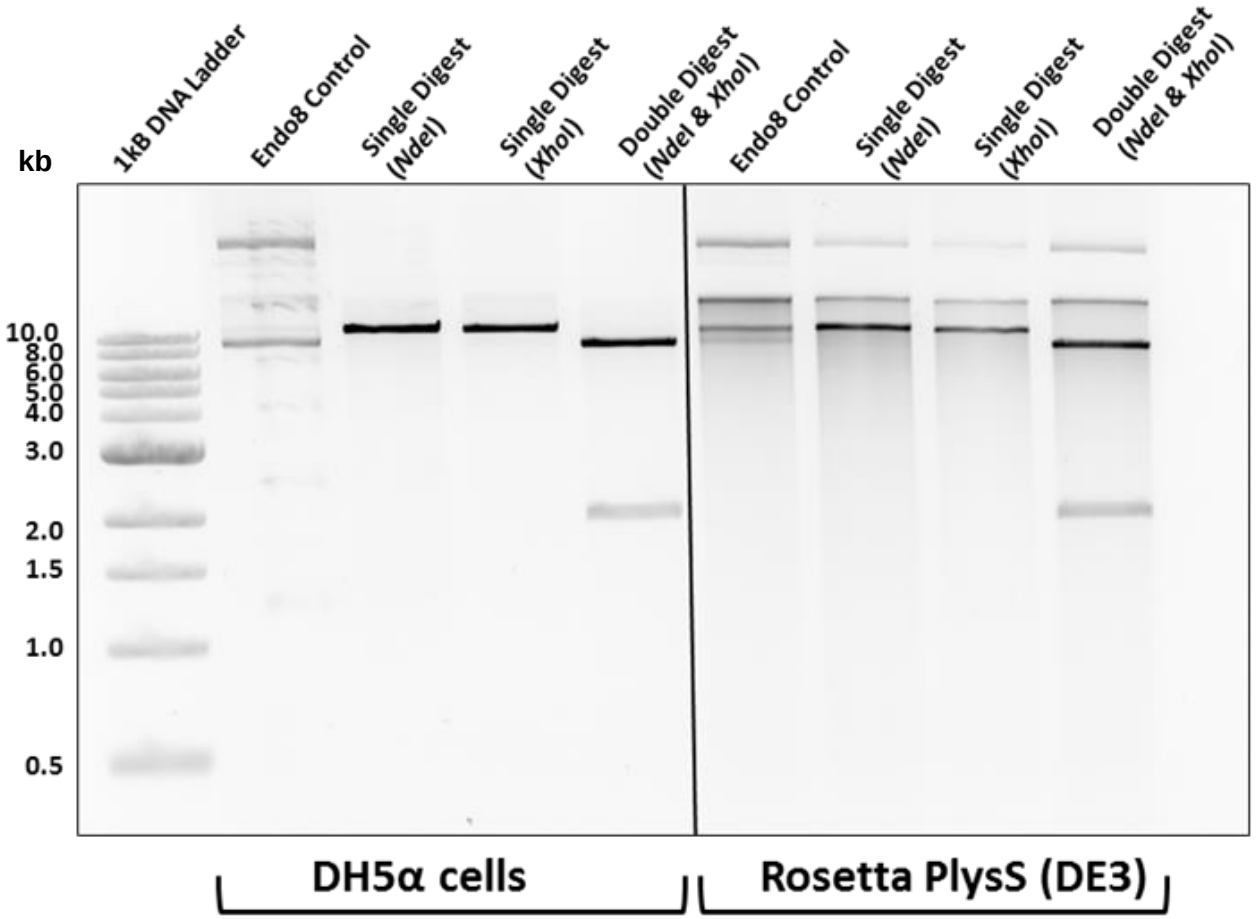



**Figure 3.4: Insert confirmation for Endo8 ORF/gene clones into a pET30b(+) vector.**

1 kB DNA ladder (Cat. #: N3232L). The gene was cloned in with *Nde*I and *Xho*I. Confirmation was conducted on both plasmids in DH5α cells and Rosetta™ (DE3) pLysS cells by restricting the plasmid with the two restriction enzymes and electrophoresed on a 1% agarose gel.

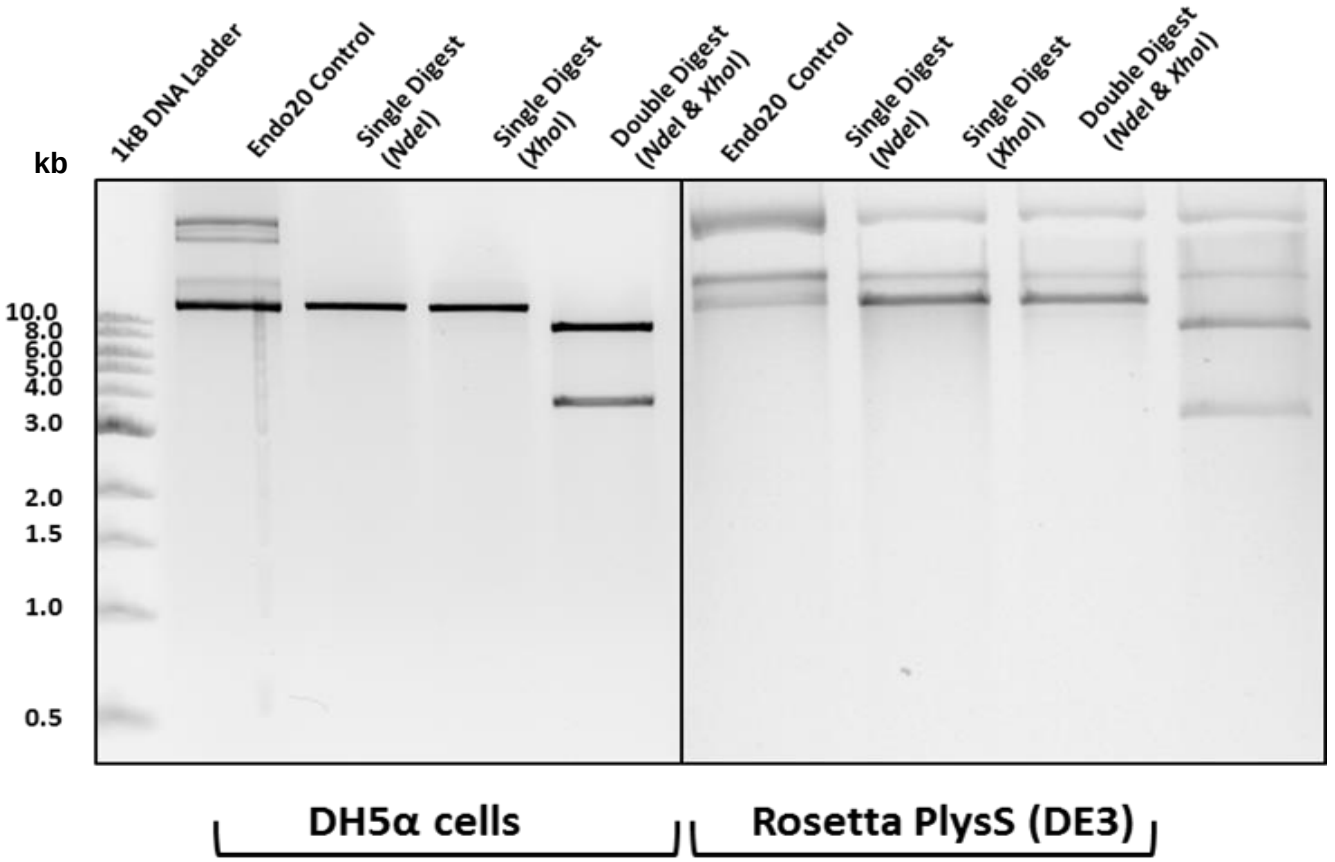



**Figure 3.5: Insert confirmation for Endo20 ORF/gene cloned into a pET30b(+) vector.**

1 kB DNA ladder (Cat. #: N3232L). The gene was cloned in with *Nde*I and *Xho*I. Confirmation was conducted on both plasmids in DH5α cells and Rosetta™ (DE3) pLysS cells by restricting the plasmid with the two restriction enzymes followed by electrophoreses on a 1% agarose gel.

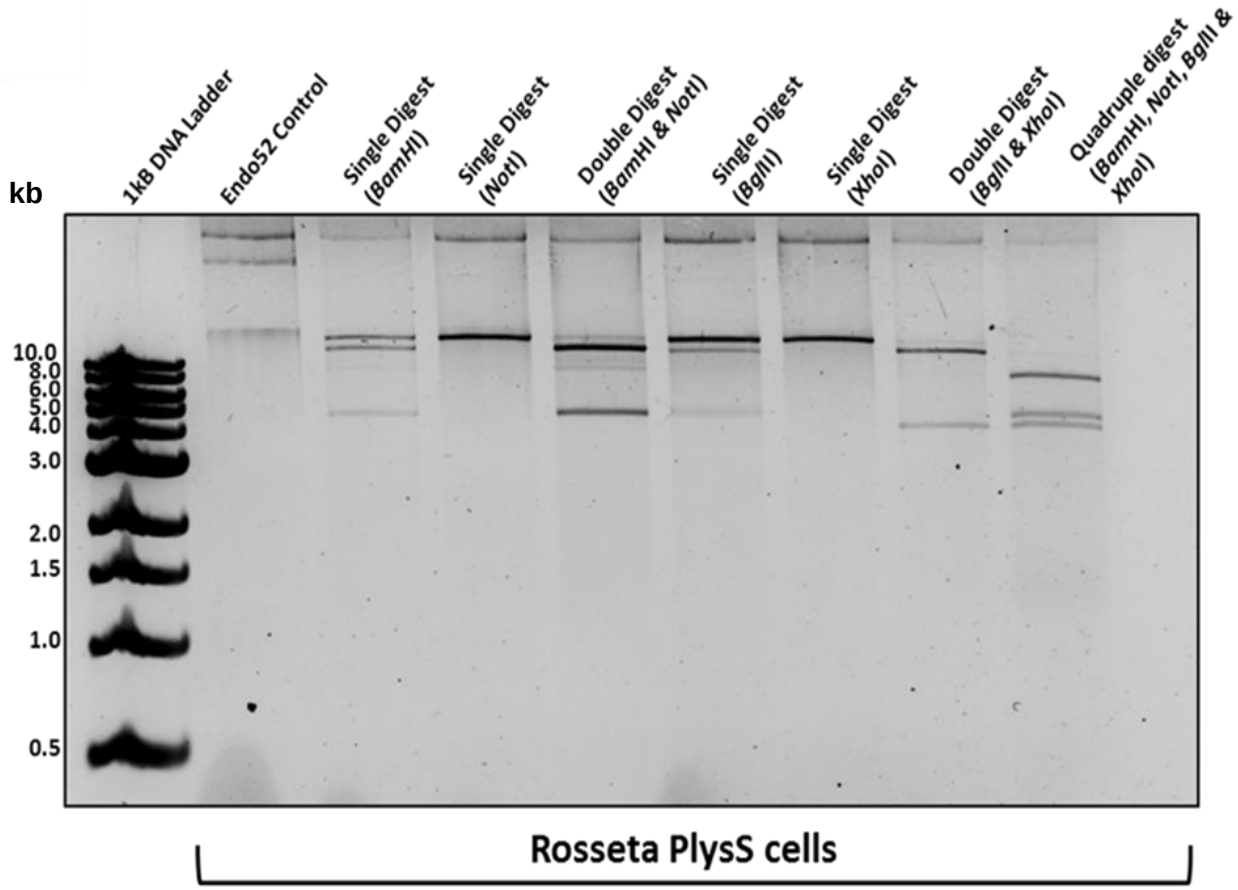



**Figure 3.6: Insert confirmation for Endo52 ORF/gene cloned into a pETduet-1 vector.**

1 kB DNA ladder (Cat. #: N3232L). Two ORFs were cloned in with *Bam*HI and *Not*I and *Bgl*II and *Xho*I respectively. Confirmation of the insert in Rosetta™ (DE3) pLysS cells was conducted the plasmid by restricting the plasmid with the two restriction enzymes and electrophoresed on a 1% agarose gel.

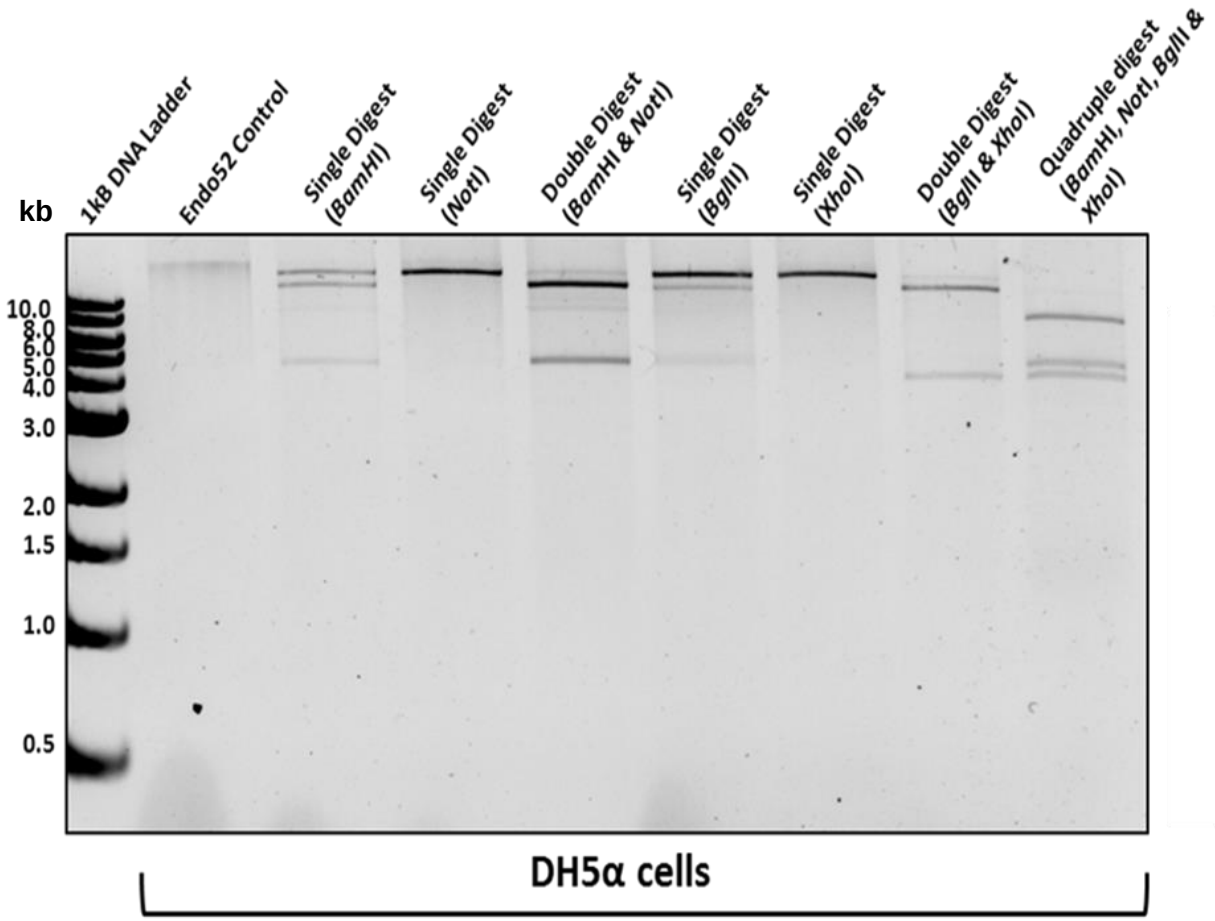



**Figure 3.7: Insert confirmation for Endo52 ORF/gene clones into a pETduet-1 vector.**

1 kB DNA ladder (Cat. #: N3232L). Two ORFs were cloned in with *Bam*HI and *Not*I and *Bgl*II and *Xho*I respectively. Confirmation insert was conducted the plasmid in DH5α cells by restricting the plasmid with the two restriction enzymes and electrophoresed on a 1% agarose gel.

### 3.4.2 Heterologous expression studies of three putative restriction endonucleases in *E. coli*

Recombinant expression of the three selected RE genes was conducted using *E. coli* as an expression host; two different strains were used, Rosetta™ (DE3) pLysS and BL-21 (DE3). Optimization studies were conducted to determine which temperature was best suited for the expression of pET30b(+)Endo8 (Endo8), pET30b(+)Endo20 (Endo20) and pETDuet-1Endo52 (Endo52) using a single IPTG concentration. Three temperatures tested in the study were 17, 25 and 30 °C, with a constant IPTG concentration of 0.1 mM. The molecular masses of the expressed proteins were estimated to be 50.4 kDa, 88.44 kDa and 110 kDa based on the deduced amino acid sequence and determined using the ExPasy Protparam tool (http://web.expasy.org/protparam/). Size determination of the proteins of interest using the Image Lab 4.1 Software from BioRad showed molecular weight that slight differ from the size determined by ExPasy Protparam (Table A4). On average the sizes are estimated to be 53.71 83.89 and 124.16 kDa respectively. Small-scale production and purification of successfully expressed soluble protein was conducted with best optimised conditions.

#### 3.4.2.1 Heterologous expression of endo8 in *E. coli* Rosetta™ (DE3) pLysS small scale purification

The expression trials of endo8 were conducted at three different temperatures. Expression of endo8 at all three temperatures was seen only after induction with IPTG. Figures 3.8 A, 3.9 A and 3.10 A show extracted protein from induced and uninduced cells expressing endo8 over a period of time at the three temperatures used. A protein band that corresponds in size with the estimated molecular weight for endo8 can be seen in post-induction samples. Western blot analysis was conducted to confirm that the band observed in Figures 3.8 B, 3.9 B and 3.10 B was the target protein. Since the protein was fused with a 6x histidine tag on the N-terminal, an anti-His tag antibody was used as primary antibody for blotting and the Western blot confirmed that the expression of endo8 was indeed successful.

Soluble and insoluble expression of the protein were studied under one temperature (25 °C) with the same expression host. As can be seen in figure 3.11, the protein is expressed in the soluble fraction as an intense band can be seen right across the different points; however, a small amount can be seen to be expressed in the insoluble fraction. With optimisation followed by PAGE analysis, expression at 17 °C, with overnight incubation after induction (20 h) was

determined to be the most optimal conditions for expressing endo8. However, the techno-economics of expression at 25 °C results in a more cost-effective production process. Utility cost are expensive therefore, the use of a chiller for lower temperatures is more expensive.

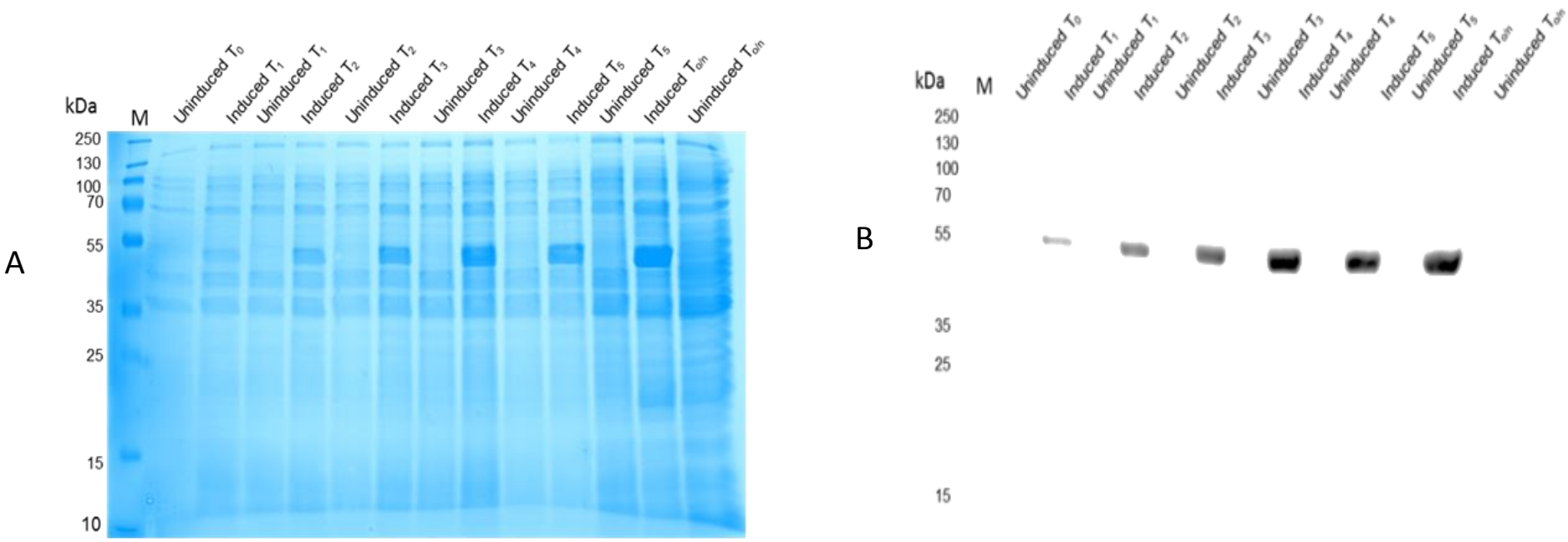



**Figure 3.8: Electrophoretogram of the induced and uninduced protein fractions of Endo8 from Rosetta™ (DE3) pLysS expressed at 17 °C.**

(A) SDS-PAGE and (B) Western blot electrophoretograms of the fraction sampled every hour for five hours post induction and overnight. M-Prestained Protein Ladder. $T_0$ -fraction before induction; $T_1$- 1 hour post induction and similarly other T labels refer to time post induction at an hour interval; $T_{o/n}$ - fraction taken the next day post induction.

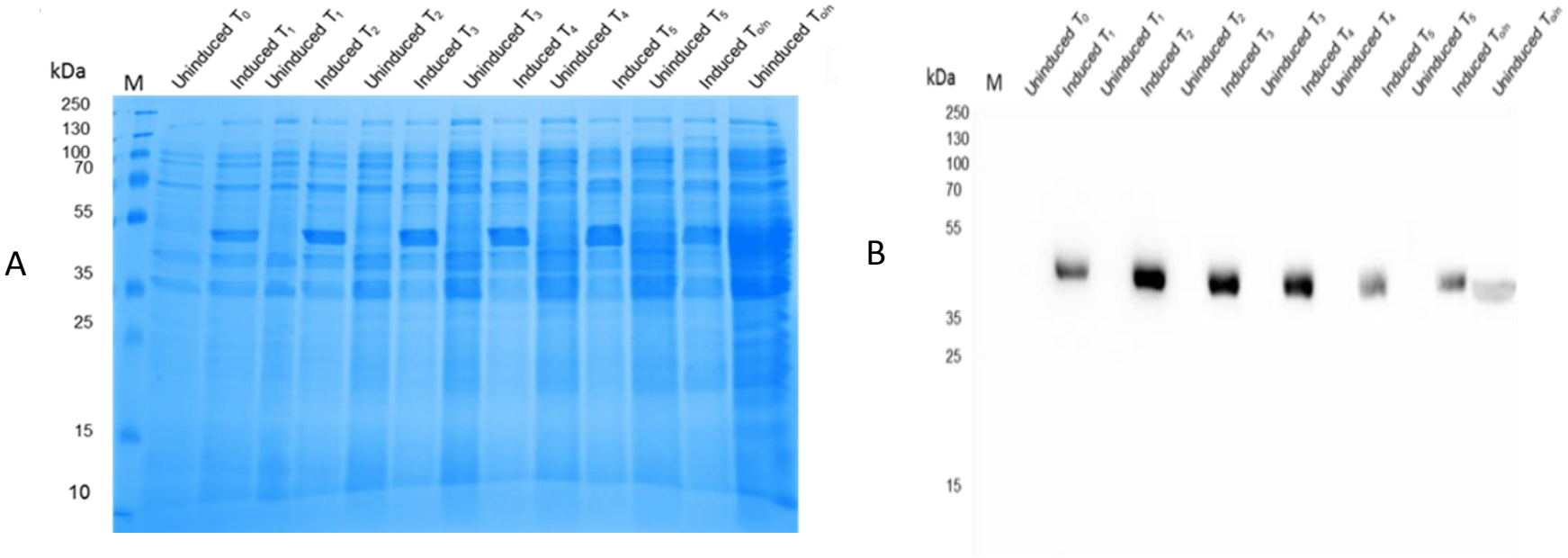
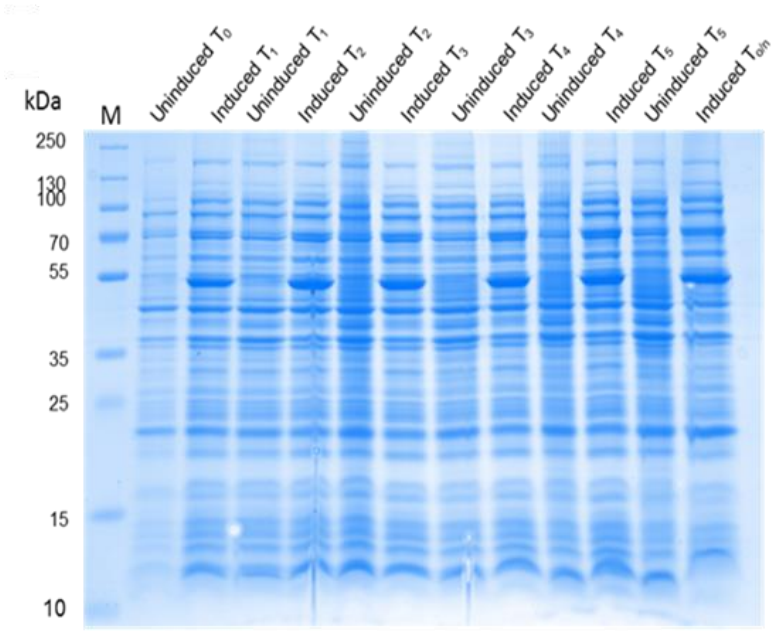



**Figure 3.9: Electrophoretogram of the induced and uninduced protein fractions of Endo8 from Rosetta™ (DE3) pLysS expressed at 25 °C.**

(A) SDS-PAGE and (B) Western blot electrophoretograms of the fraction sampled every hour for five hours post induction and overnight. M-Prestained Protein Ladder. $T_0$-fraction before induction; $T_1$- 1 hour post induction and similarly other T labels refer to time post induction at an hour interval; $T_{o/n}$ - fraction taken the next day post induction.

A




B

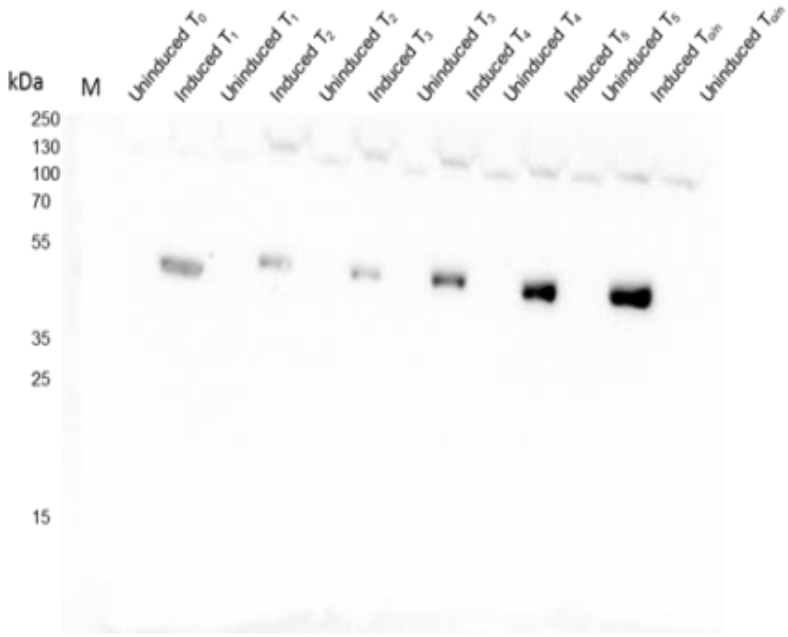



**Figure 3.10: Electrophoretogram of the induced and uninduced protein fractions of Endo8 from Rosetta™ (DE3) pLysS expressed at 30 °C.**

(A) SDS-PAGE and (B) Western blot electrophoretograms of the fraction sampled every hour for five hours post induction and overnight. M-Prestained Protein Ladder. $T_0$ - fraction before induction: $T_1$- 1 hour post induction and similarly other T labels refer to time post induction at an hour interval; $T_{o/n}$ - fraction taken the next day post induction.

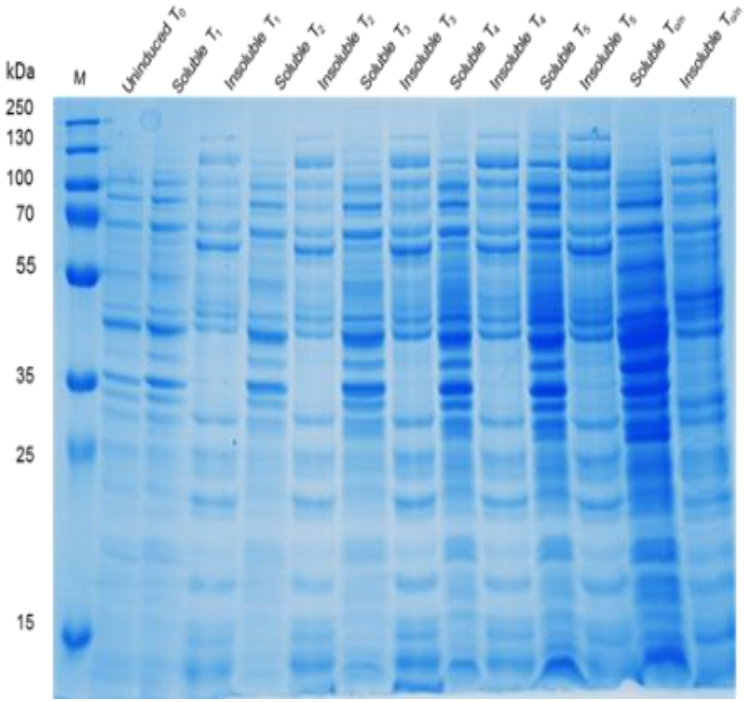



**Figure 3.11: Electrophoretogram of the soluble and insoluble protein fractions of Endo8 from Rosetta™ (DE3) pLysS expressed at 25 °C.**

The sampling was done every hour for five hrs post-induction and overnight. M-Prestained Protein Ladder. $T_0$ - fraction before induction; $T_1$- 1 hour post induction and similarly other T labels refer to time post induction at an hour interval; $T_{o/n}$ - fraction taken the next day post induction.

#### 3.4.2.1.1 Small scale production and purification of Endo8

Endo 8 expression studies yielded soluble protein at all the temperatures studied when induced wit 0.1mM IPTG. The small-scale production of the protein was conducted at 25°C with induction at 0.1 mM IPTG. The endo8 expression construct was designed with an N-terminal His-to enable easy purification. The immobilized metal affinity chromatography (IMAC) technique was used in this study to purify the protein using Ni-TED resin. Figure 3.12 illustrates the purification fractions of Endo8. No visible protein loss can be observed in the flow-through which, indicated that the protein was successfully bound to the resin under the conditions used. The bound protein was eluted with 250 mM imidazole; at this concentration some co- elution of other cellular proteins was observed. The SDS-PAGE (Figure 3.12) analysis shows the different fractions obtained during the purification steps and a band corresponding with the expected size for endo8 can be observed in the eluted fraction (indicated with an arrow). Table A5 illustrates the purity of the purified enzyme.

### 3.4.2.2 Heterologous expression of endo20 in *E. coli* Rosetta™ (DE3) pLysS and BL-21(DE3)

Expression studies of endo20 using Rosetta™ (DE3) pLysS were conducted at three different temperature conditions (17 °C, 25 °C and 30 °C) with induction at 0.1 mM IPTG. Figures 3.13 - 3.15 show the profile of expression at different times post-induction. No indication of expression was observed for two temperatures (17 °C and 30 °C) tested (Figure 3.13 and 3.15) this observation is apparent even on analysis with more sensitive western blots using anti-6x His-tag antibodies. However, on one occasion a band corresponding with the estimated size of endo20 was observed at 4hr post induction and confirmed with western blot (Figure 3.14). SDS-PAGE analysis of the soluble and insoluble fraction was conducted and protein bands corresponding with the estimated molecular mass of endo20 can be seen on both fractions from 3 hr post induction (Figure 3.16). However, these results were not reproducible (results not shown), as the protein of interest was expressed in the insoluble fraction in subsequent trails. A challenges was experienced when separating the pellet (insoluble fraction) against the supernatant (soluble fraction), which might have led to the inconsistency of the results observed. *E.coli* BL-21 (DE3) OneShot cells were then tested as an alternative host. The same expression conditions were used with new expression host to study the expression of endo20. However, all the expression studies in *E.coli* BL-21 (DE3) OneShot showed no expression both in the induced and

uninduced fractions in all the different time points and all the different temperatures (Figure A4 and Figure A5).

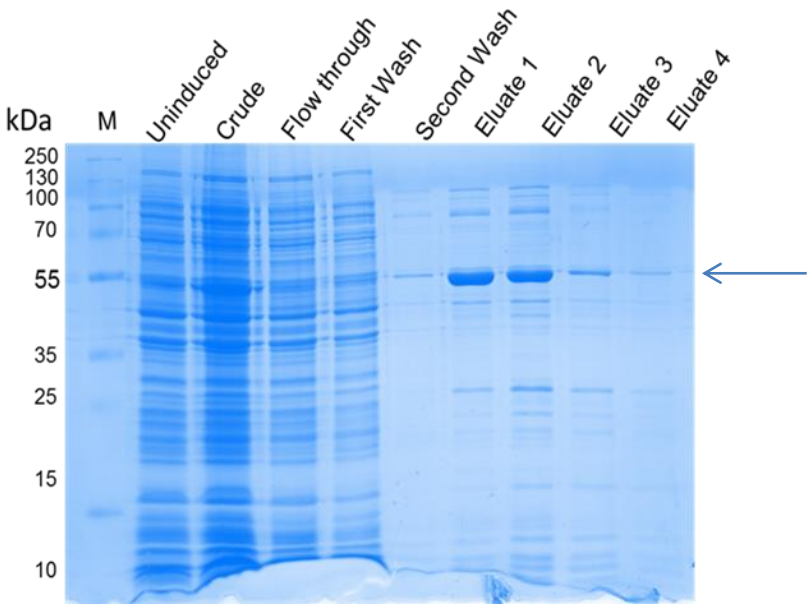



**Figure 3.12 : SDS-PAGE analysis of the purification of endo8 at 25°C with the different elution fractions at 3 hr post induction.**

M-Prestained Protein Ladder. Crude- obtained from B-Per treatment of the pelleted cells; flow through- collection after protein binds to the column; Wash (first & second)-washing of the column with lysis buffer, Eluate-bound protein eluted with elution buffer with imidazole.

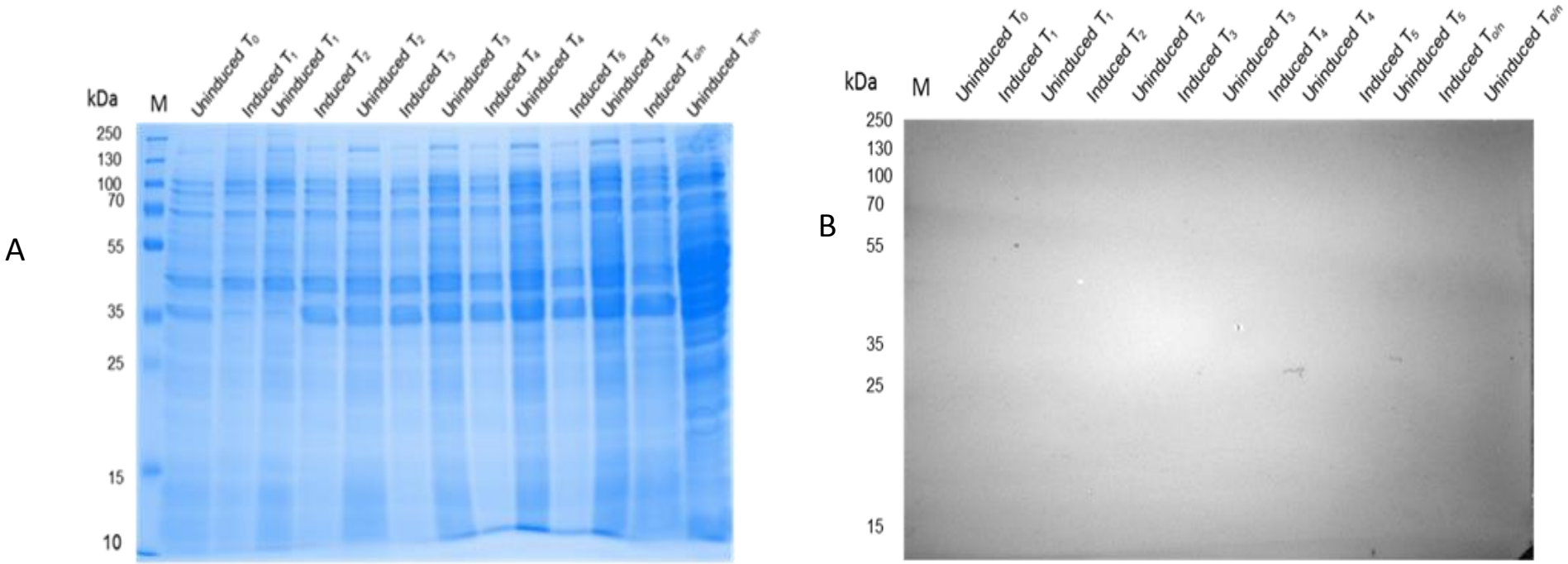



**Figure 3.13: Electrophoretogram of the induced and uninduced protein fractions of Endo20 from Rosetta™ (DE3) pLysS expressed at 17 °C.**

(A)SDS-PAGE and (B) Western blot electrophoretograms of the fraction sampled every hour for five hours post-induction and overnight. M-Prestained Protein Ladder. $T_0$ - fraction before induction; $T_1$- 1 hour post induction and similarly other T labels refer to time post induction at an hour interval; $T_{o/n}$ - fraction taken the next day post induction.

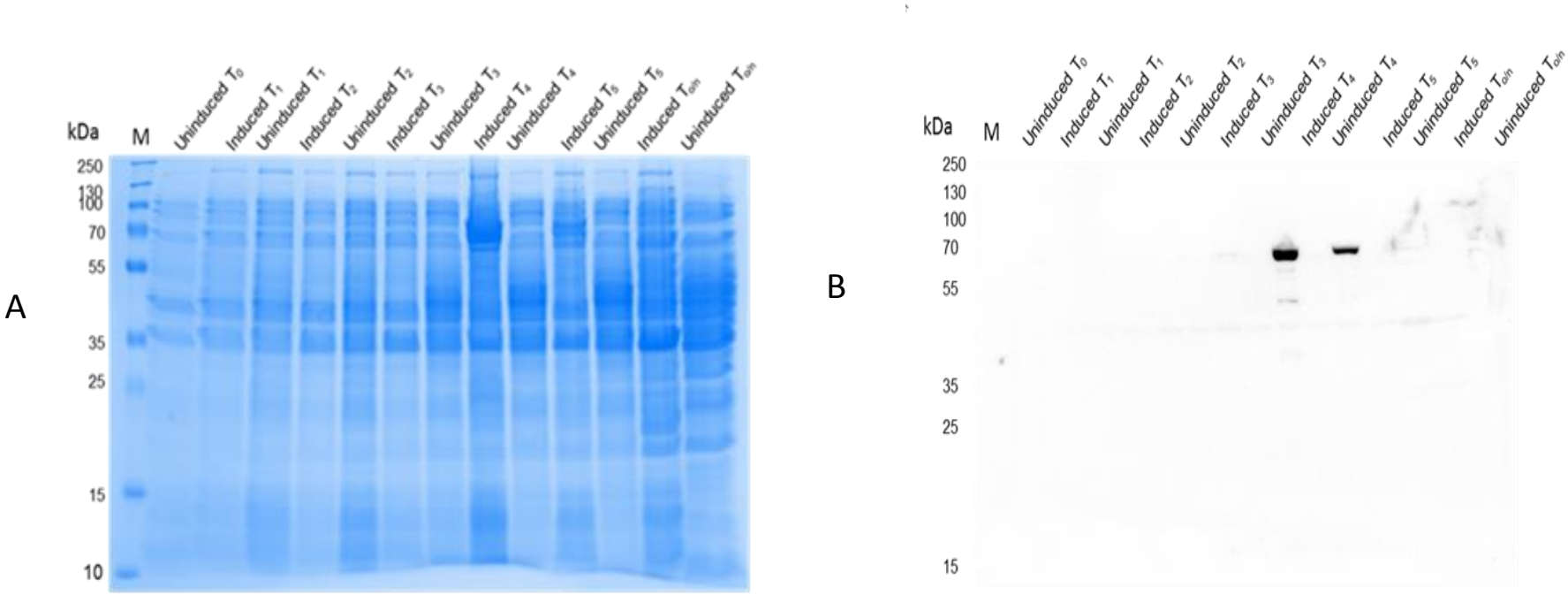



**Figure 3.14: Electrophoretogram of the induced and uninduced protein fractions of Endo20 from Rosetta™ (DE3) pLysS expressed at 25 °C.**

(A) SDS-PAGE and (B) Western blot electrophoretograms of the fraction sampled every hour for five hours post induction and overnight. M-Prestained Protein Ladder. $T_0$ - fraction before induction; $T_1$- 1 hour post induction and similarly other T labels refer to time post induction at an hour interval; $T_{o/n}$ - fraction taken the next day post induction.

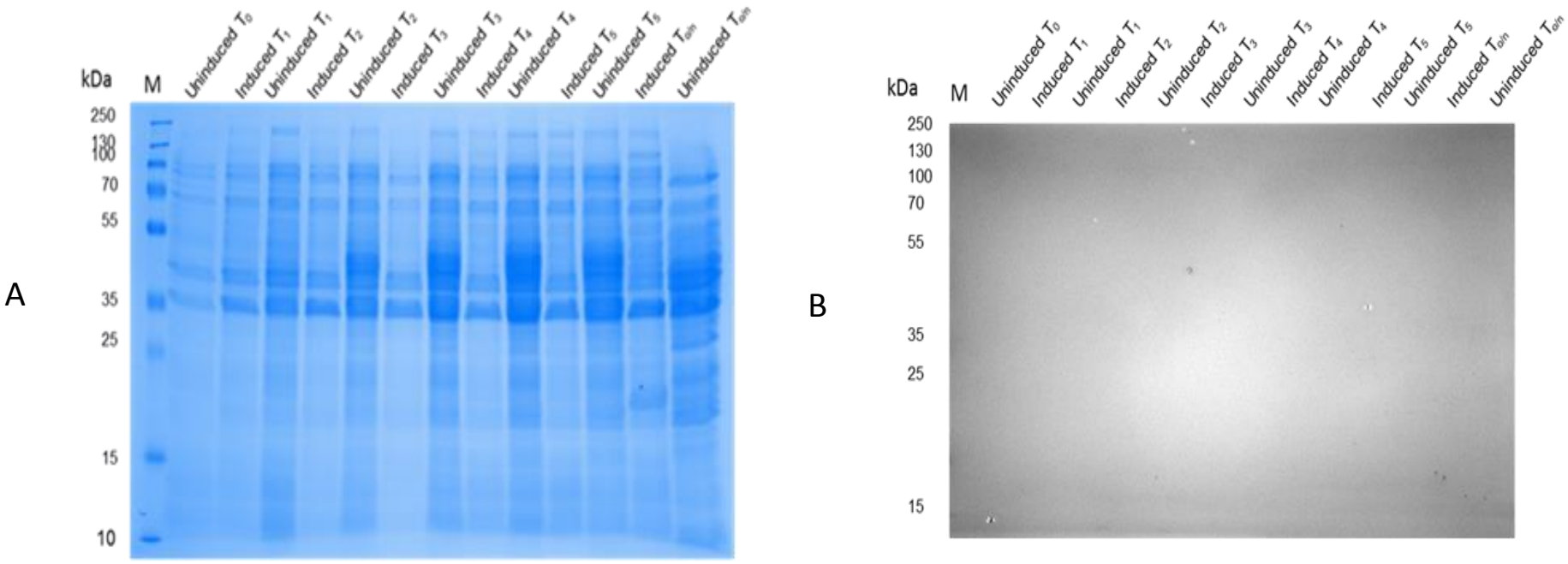



**Figure 3.15: Electrophoretogram of the induced and uninduced protein fractions of Endo20 from Rosetta™ (DE3) pLysS expressed at 30 °C.**

(A) SDS-PAGE and (B) Western blot electrophoretograms of the fraction sampled every hour for five hours post induction and overnight. M-Prestained Protein Ladder. $T_0$ - fraction before induction; $T_1$- 1 hour post induction and similarly other T labels refer to time post induction at an hour interval; $T_{o/n}$ - fraction taken the next day post induction.

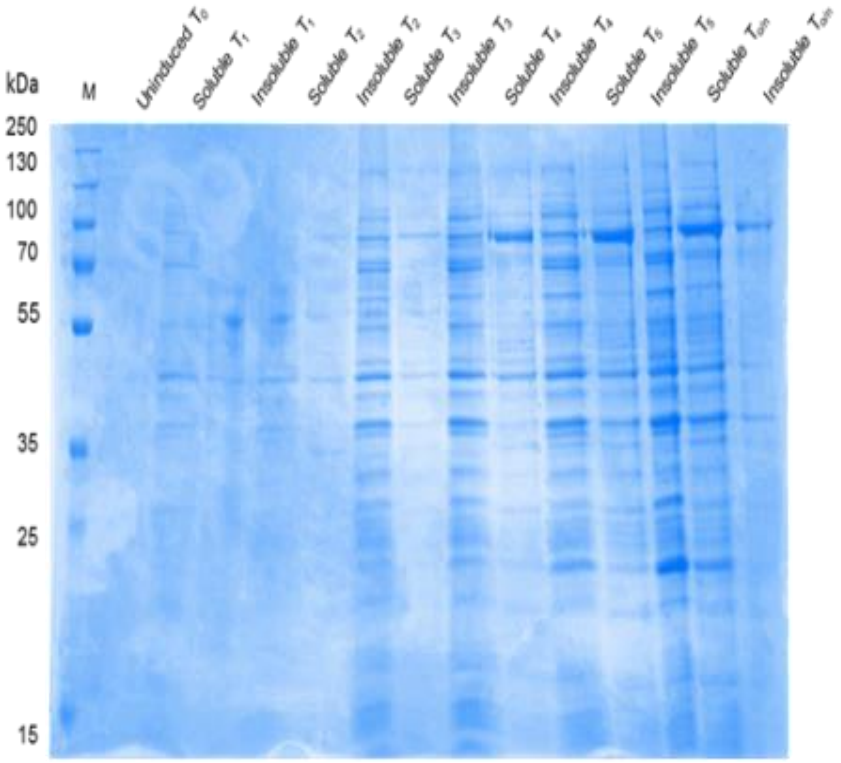



**Figure 3.16: Electrophoretogram of the soluble and insoluble protein fractions of Endo20 from Rosetta™ (DE3) pLysS expressed at 25 °C.**

The sampling was done every hour for 5 h post induction and overnight. M-Prestained Protein Ladder. $T_0$ - fraction before induction; $T_1$- 1 hour post induction and similarly other T labels refer to time post induction at an hour interval; $T_{o/n}$ - fraction taken the next day post induction.

Figure A6 shows the studies of the protein expression in the soluble and insoluble fraction and no expression is observed in both fractions. With the protein resulting in inclusion bodies in Rosetta™ (DE3) pLysS post induction and no expression at all in BL21 (DE3) OneShot, no further studies were conducted on this protein within this study.

#### 3.4.2.3 Heterologous expression of endo52 in Rosetta™ (DE3) pLysS and BL-21

Endo52 was also successfully expressed in Rosetta™ (DE3) pLysS. However, the expression yields differed under the three temperatures and the different time points post induction studied (Figure 3.17 – 3.19). An intense band with an estimated molecular weight was observed in the induced fractions for endo52. Analysis by SDS-PAGE showed that at 17 °C, expression was only detected 20 hrs post-induction (Figure 3.17 A). However the western blot (Figure 3.17 B) detects a 6x His-tagged protein with the molecular weight similar to that of endo52 across all the time points whilst there is evidence of non-specific binding. SDS-PAGE analysis at 25 °C and 30 °C show expression of the protein from two and three hours post induction respectively (Figure 3.18 A – 3.19 A) and the more sensitive analysis with western blot detects expression from one hour post induction (Figure 3.18 B and 3.19 B). However, as seen in the different expression studies analysed with the aid of western blot, expression at 25 °C yielded better results when compared to the others (Figure 3.18 B). The blot detects a single band of a His-tagged protein as opposed to the western blots of expression studies conducted at 17 °C and 30 °C (Figure 3.17 B - 3.19 B).

Analysis of the two fractions (soluble and insoluble) was conducted on expressed cells at 25 °C. The temperature was selected based on the results of the expression observed from the temperature profile studies, using the same expression host as well. The expression of the protein over time can be observed both in the soluble and insoluble fraction. Figure 3.20 shows that the protein also gets expressed in the insoluble fraction, however, the concentration of the protein expressed in the soluble fraction is higher than that of the one expressed in the insoluble fraction. Expression of endo52 was also studied in *E.coli* BL-21 (DE3) OneShot cells. The same conditions used in the expression studies in Rosetta™ (DE3) pLysS expression host were used, however expression was achieved insoluble fraction (Figure A7). Therefore, the standard expression temperature for endo52 was determined to be 25 °C with an IPTG concentration of 0.1 mM using Rosetta™ (DE3) pLysS as the expression host.

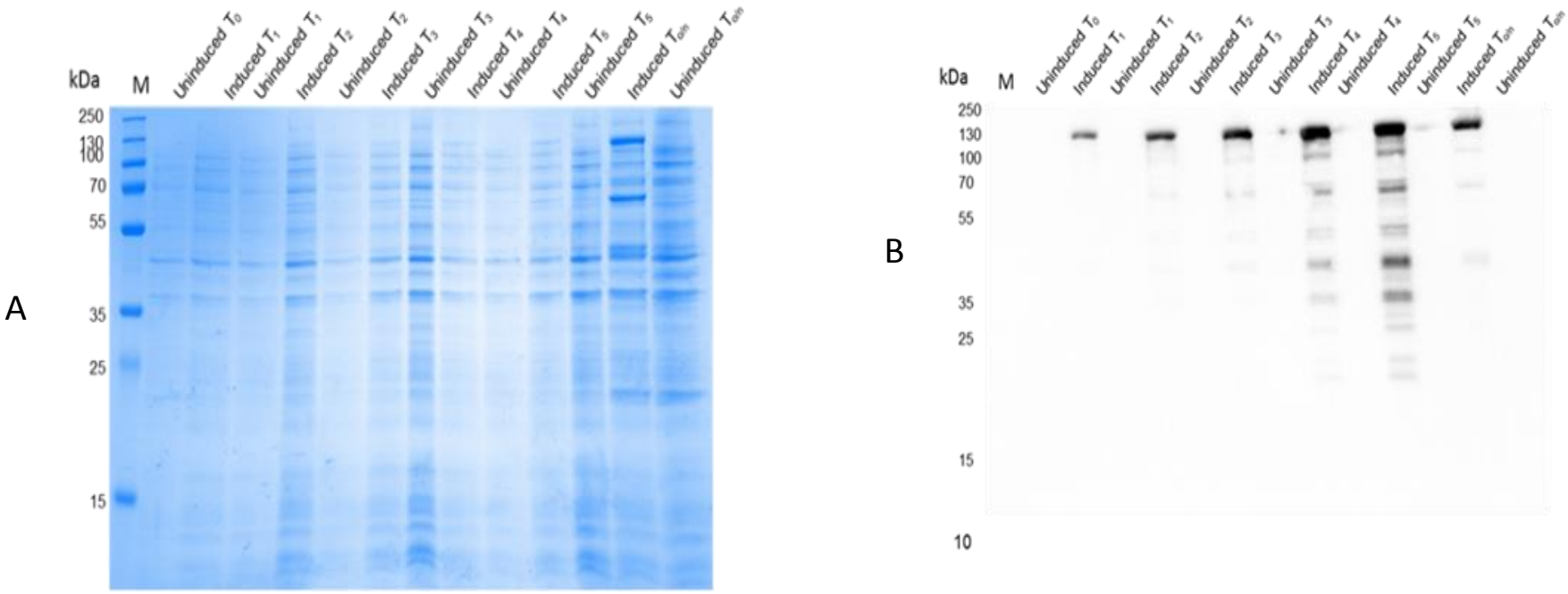



**Figure 3.17: Electrophoretogram of the induced and uninduced protein fractions of Endo52 from Rosetta™ (DE3) pLysS expressed at 17 °C.**

(A) SDS-PAGE and (B) Western blot electrophoretograms of the fraction sampled every hour for five hours post induction and overnight. M-Prestained Protein Ladder. $T_0$ - fraction before induction; $T_1$- 1 hour post induction and similarly other T labels refer to time post induction at an hour interval; $T_{o/n}$ - fraction taken the next day post induction.

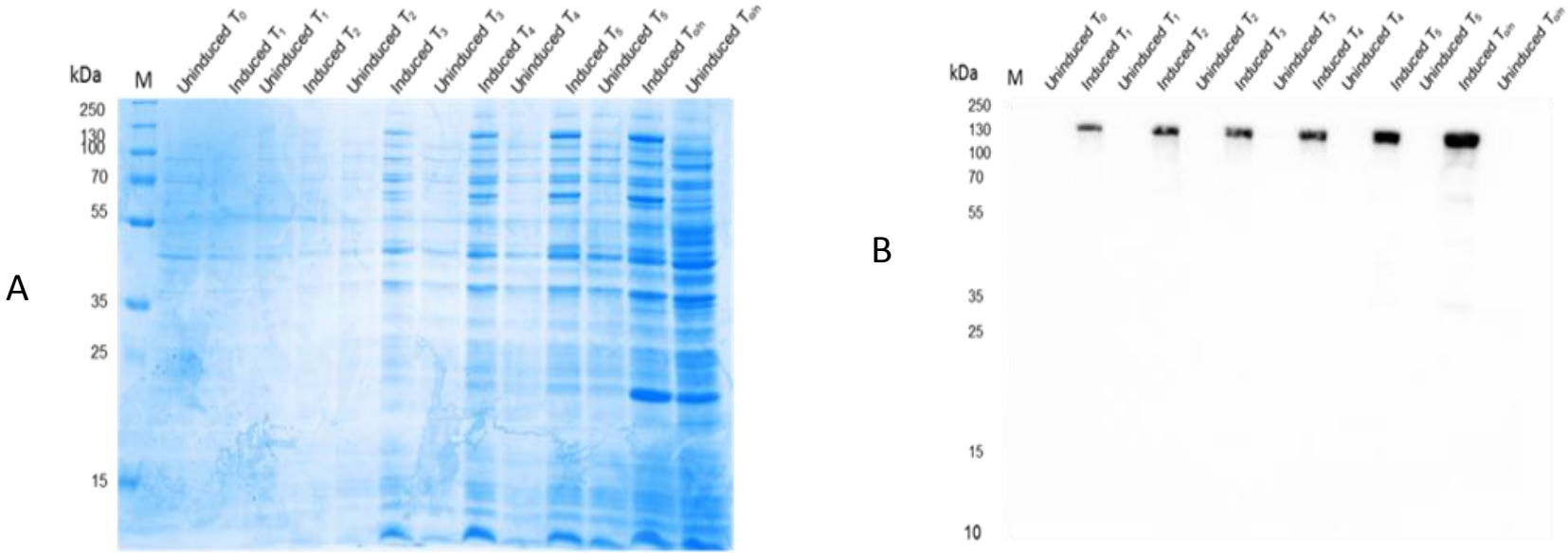



**Figure 3.18: Electrophoretogram of the induced and uninduced protein fractions of Endo52 from Rosetta™ (DE3) pLysS expressed at 25 °C.**

(A) SDS-PAGE and (B) Western blot electrophoretograms of the fraction sampled every hour for five hours post induction and overnight. M-Prestained Protein Ladder. $T_0$ - fraction before induction; $T_1$- 1 hour post induction and similarly other T labels refer to time post induction at an hour interval; $T_{o/n}$ - fraction taken the next day post induction.

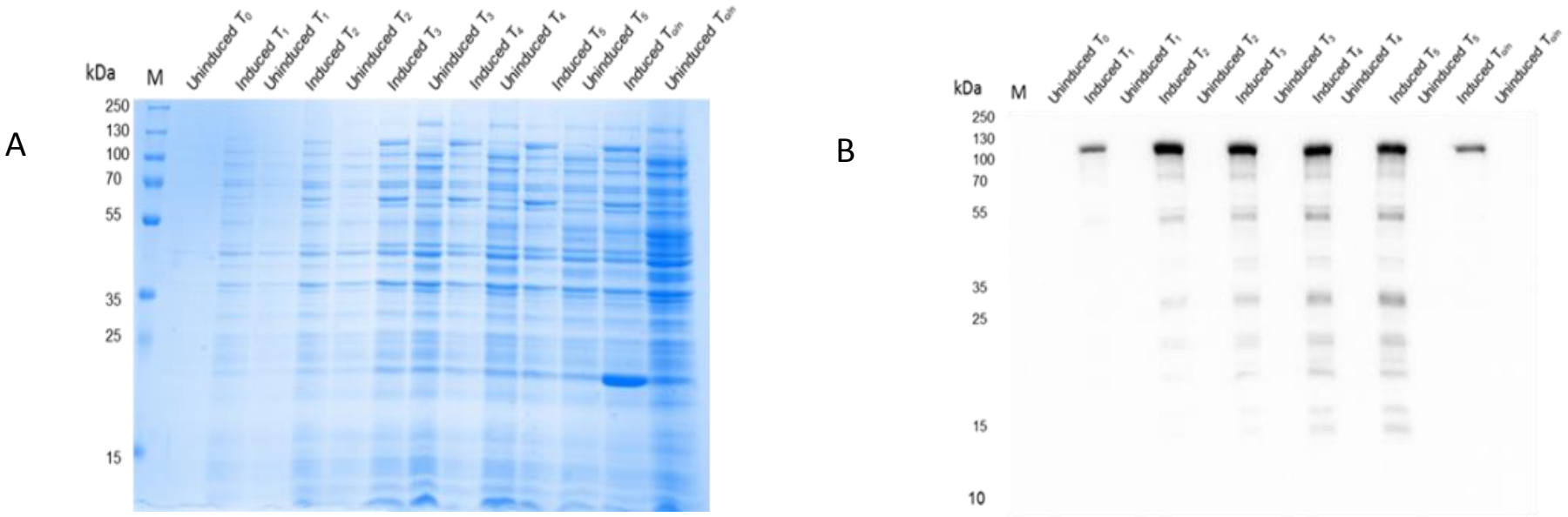



**Figure 3.19: Electrophoretogram of the induced and uninduced protein fractions of Endo52 from Rosetta™ (DE3) pLysS expressed at 30 °C.**

(A) SDS-PAGE and (B) Western blot electrophoretograms of the fraction sampled every hour for five hours post induction and overnight. M-Prestained Protein Ladder. $T_0$ - fraction before induction; $T_1$- 1 hour post induction and similarly other T labels refer to time post induction at an hour interval; $T_{o/n}$ - fraction taken the next day post induction

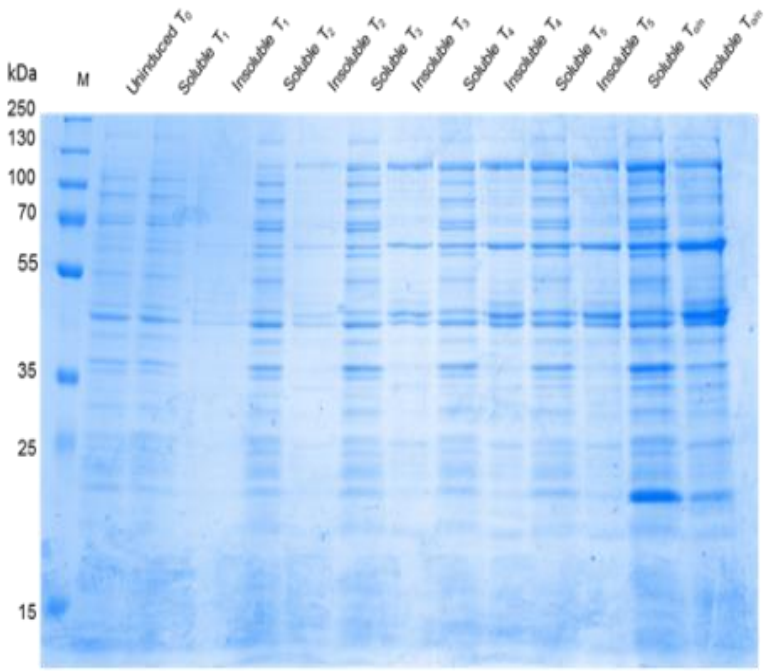



**Figure 3.20: Electrophoretogram of the soluble and insoluble protein fractions of Endo52 from Rosetta™ (DE3) pLysS expressed at 25 °C.**

The sampling was done every hour for five hours post induction and overnight. M-Prestained Protein Ladder. $T_0$-fraction before induction; $T_1$- 1 hour post induction and similarly other T labels refer to time post induction at an hour interval; $T_{o/n}$ - fraction taken the next day post induction

#### 3.4.2.3.1 Small scale production and purification of Endo52

Soluble endo52 protein was observed at various time points in the crude extract from the expression study. Optimal expression was achieved at 25 °C; five hours post induction 0.1mM IPTG. Since the ORF encoding Endo52 was cloned into a pETDuet-1 fused with a 6x His-tag at the N-termini, purification using IMAC was achieved successfully (Figure 3.21). The SDS-PAGE analysis shows that Endo52 bound the resin and was eluted with buffer containing imidazole, an intense band was observed with molecular mass of 110 kDa as expected of Endo52 based on the sequence length post purification, with average purity of 93%.

### 3.4.3 Endonuclease activity assay

After successfully purifying Endo8 and Endo52, the two enzymes were functionally characterised. The purified enzymes were used in restriction digest reaction under various conditions to observe the most optimal conditions for enzymes functionality. The reaction was set-up according to Section 2.8.1 with the necessary co-factors required. All reactions were conducted with a negative control. In each reaction, the concentration of DNA was fixed throughout the different tests conducted. The reaction was set up using different buffers to observe which buffer would show optimal activity.

Both enzymes showed no specific cleavage; however, in both cases either faint bands or entire degradation of substrate DNA due to incomplete linearization or extensive cleavage were observed, respectively. Figure 3.22 represents restriction endonuclease activity of endo8 using different buffers on HgDNA, MgDNA and λDNA. Endo8 activity with buffer 3 has the most activity across the different DNA templates used, while with buffer 7, no activity is observed (Figure 3.22 A and 3.22 B). The activity in other buffers shows less activity if any at all. Restriction endonuclease activity of Endo52's is illustrated in Figure 3.23. In addition to other variables indicated earlier, the effect of different concentrations of ATP and enzyme concentration were also studied. Optimal activity for Endo52 was observed when 1.6 µl of a 10 mM ATP is used in the reaction (Figure 3.23 B). While in Figure 3.23 C, the optimal enzyme concentration is 0.013 $\mu g.\mu l^{-1}$.

Temperature studies of both enzymes' activity were conducted using buffer 9 in pUC19 and HgDNA. Endo8 showed incomplete digestion activity at 25 – 35 °C for (Figure 3.24 A) while Endo52 successfully linearized pUC DNA at 25 °C (Figure 3.25 A). Both enzymes activity on

HgDNA at 45 - 65 °C resulted in a smear, while at 35° C a slight degradation of the DNA can be seen, while at 25 °C the enzymes' activeness was poor (Figure 3.24 B and 3.25 B).

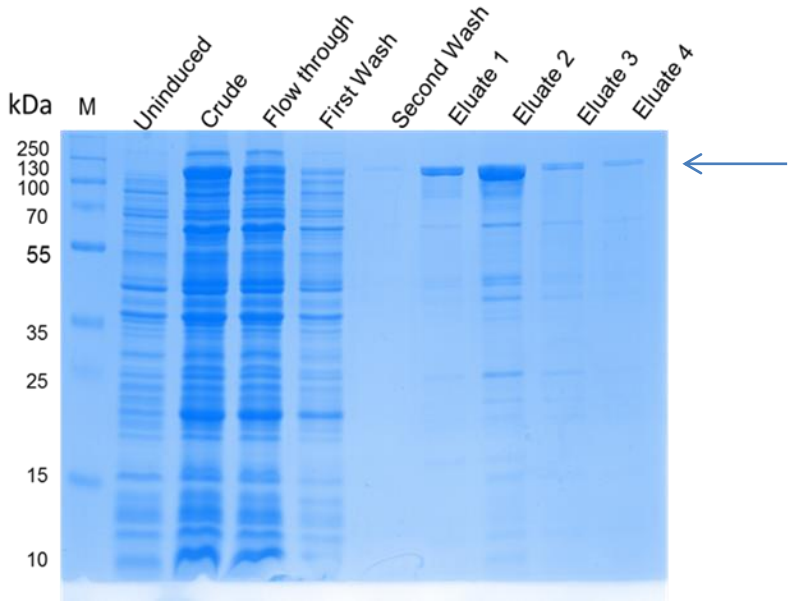



**Figure 3.21: SDS-PAGE analysis of the purification of endo52 at 25°C with the different elution fractions at 5 hr post induction.**

M-Prestained Protein Ladder; crude: obtained from B-Per treatment of the pelleted cells; flow through: collection after protein binds to the column; wash (first & second): washing of the column with lysis buffer, eluate:bound protein eluted with elution buffer with imidazole.

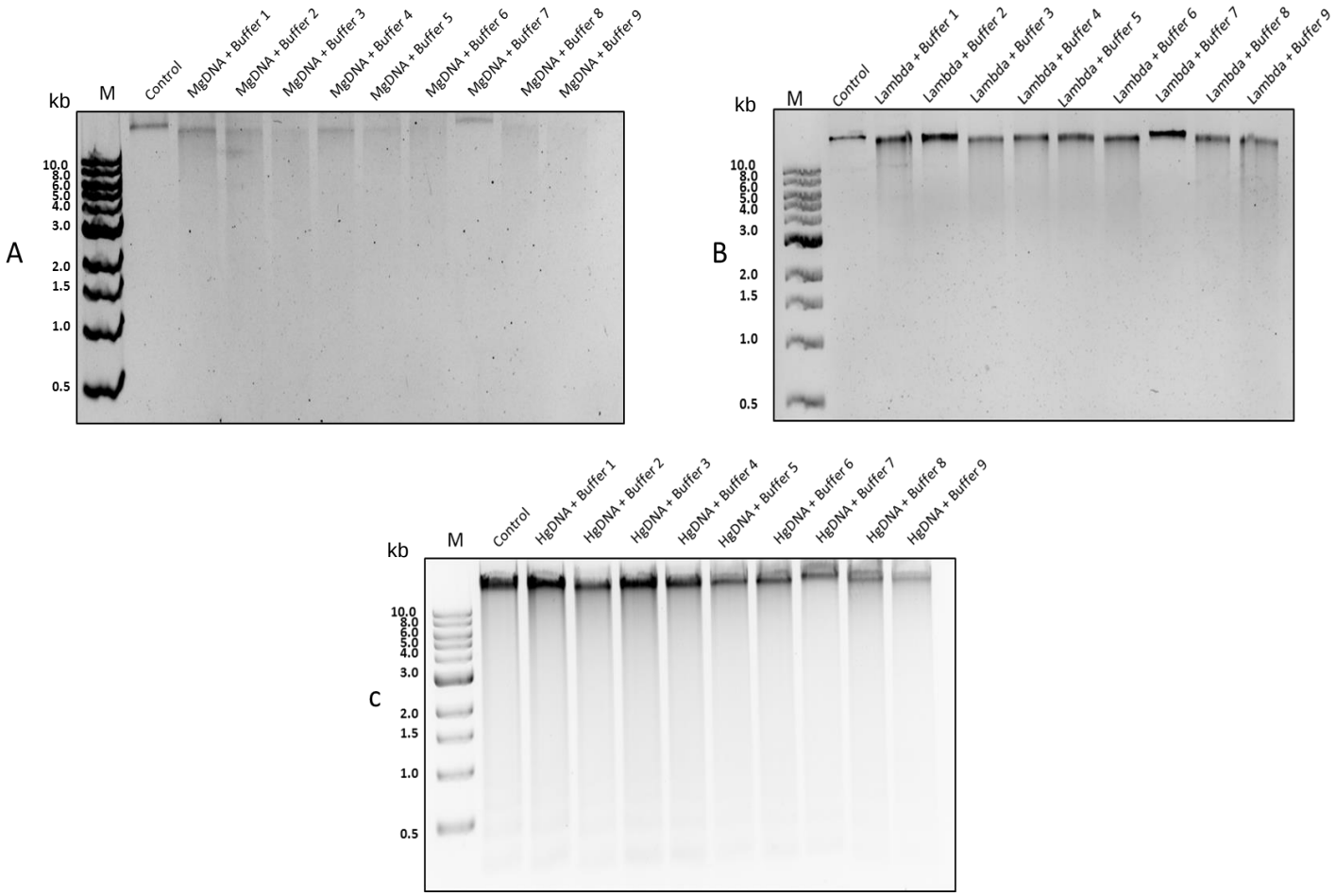



**Figure 3.22: Analysis of endo8 restriction activity on different DNA templates for 60 min.**

M: 1 kB DNA ladder (Cat. #: N3232L). (A): MgDNA, (B): λDNA and (C): HgDNA is cleaved using endo8 in 9 different buffers. The concentration of the DNA preparations was reduced by a 1:10 dilution. List of the different buffer used per reaction are found in Table 2.5. All reactions conducted at 37 °C for duration of 60 min.

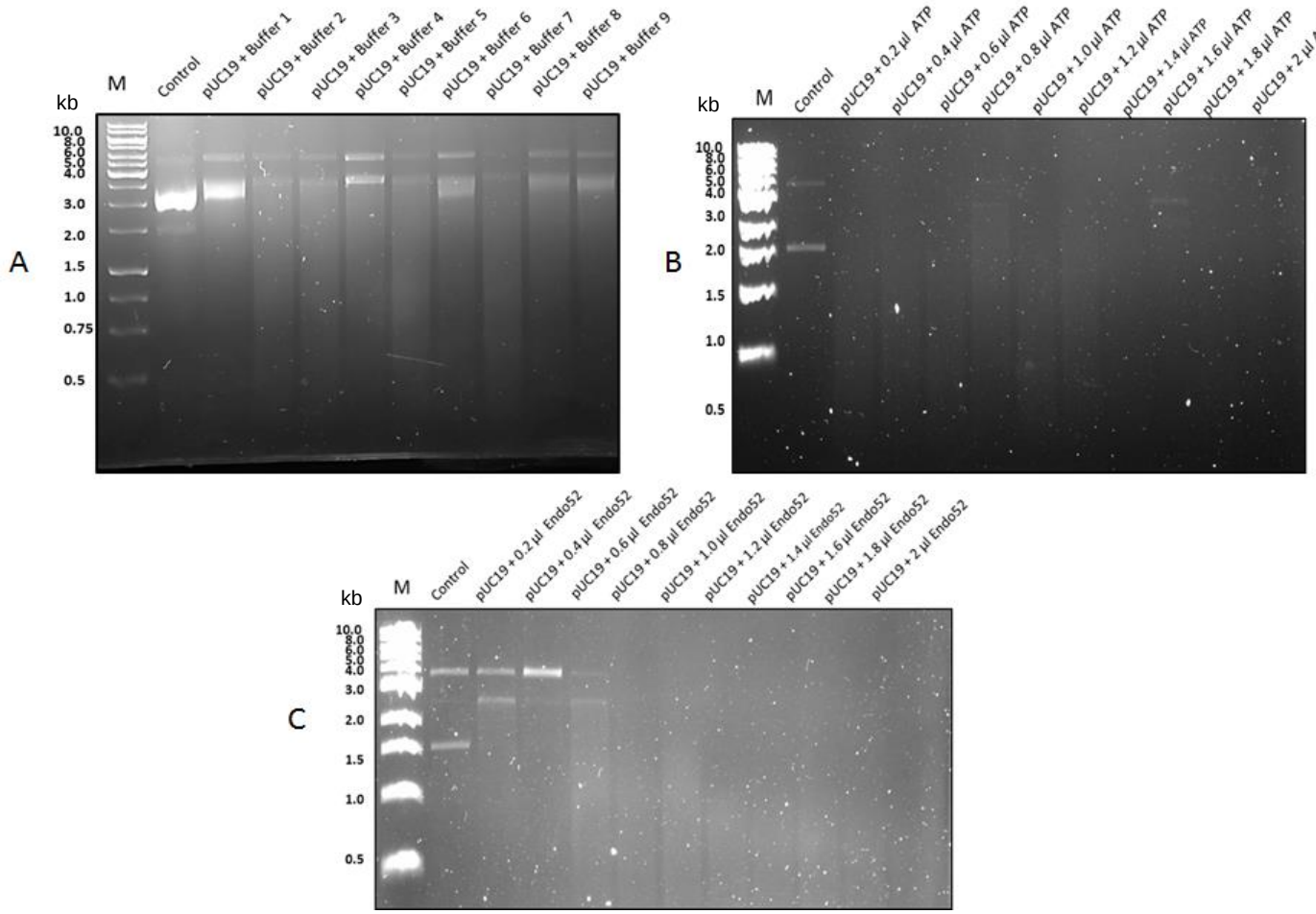



**Figure 3.23: Analysis of endo52 restriction activity on pUC19 plasmid for 60 min with different.**

M: 1 kB DNA ladder (Cat. #: N3232L). (A): pUC19 cleaved with endo52 in 9 different buffers; (B): pUC19 cleaved with endo52 with different ATP concentrations; and (C): pUC19 cleaved with endo52 with different enzyme concentrations. The concentration of the pUC19 was reduced by a 1:10 dilution. List of the different buffer used per reaction are found in Table 2.5. All reactions conducted at 37 °C for duration of 60 min.

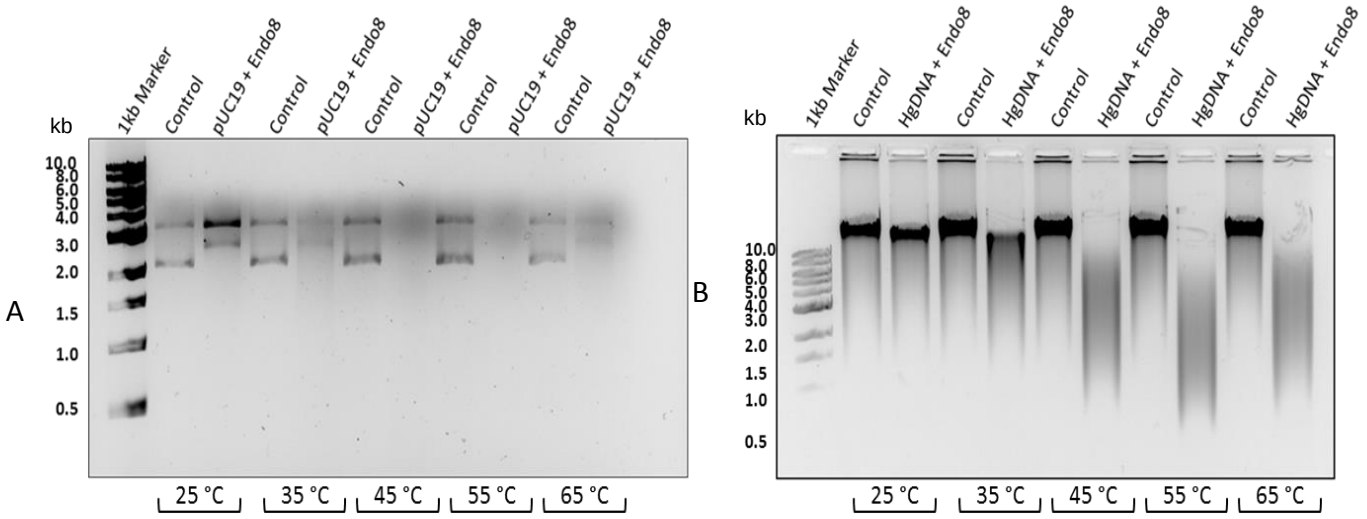



**Figure 3.24: Temperature analysis of endo8 restriction activity on different plasmid for 30 min.**

The enzymes ability to cleave DNA was observed under different temperatures, the reactions were all set for 30 min. 1 kB DNA ladder (Cat. #: N3232L). (A): pUC19 and (B): HgDNA is cleaved using endo8 under 5 different temperatures (25 - 65°). The concentration of the DNA preparations was reduced by a 1:10 dilution.

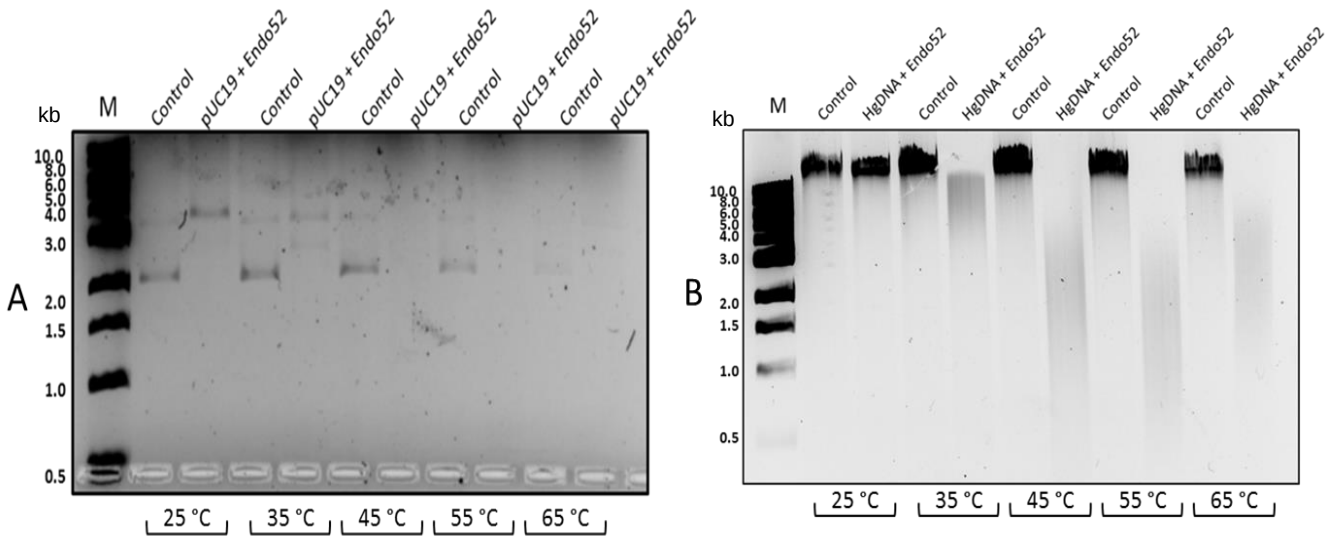



**Figure 3.25: Temperature analysis of endo52 restriction activity on different plasmid for 30 min**

The enzymes ability to cleave DNA was observed under different temperatures, the reactions were all set for 30 min. M: 1 kB DNA ladder (Cat. #: N3232L). (A): pUC19 and (B): HgDNA is cleaved using endo52 under 5 different temperatures (25 - 65°). The concentration of the DNA preparations was reduced by a 1:10 dilution.

A control reaction was set up to monitor the reaction will in the absence of the enzyme over time. The plasmid DNA remained intact without the enzyme and with the enzyme, as the time incubation period increased, the activity resulted in a smear for Endo 8 (Figure 3.26 B and 3.26 C) and Endo52 (Figure 3.27 A). Both enzymes' activity resulted in incomplete linearization (Figure 3.26 A and 3.27 B).

### 3.4.4 Scale up of Endo_8 production

The cells were grown at 37 °C for 9 hours, until the glucose in the media was consumed in triplicate fermenters of 2 l scale with a working volume of 1 l. The consumption of glucose (53 $g.l^{-1}$ ) which was fed at a rate of 10.6 $g.l^{-1}$ was monitored by measuring the glucose every two hours. During this phase of monitoring the glucose consumption, the OD of the cells growth was measured at 600 nm. Once the glucose was consumed, the temperature was lowered to 25 °C and induced with 0.1 mM of IPTG and the booster feed was initiated. Booster feed was at a rate of 11.14 $g.l^{-1}$. The fermenters were harvested at 3 and 8 hr post induction. Figure 3.28 shows glucose utilisation and cell density during the fermentation.

After harvesting, steps were taken to re-suspend the pellet from section 2.9.3 using 1 × LEW buffer. The volume used to re-suspend each pellet was 20% (w/v) of 1 × LEW buffer. The HiScale 50 column packed with Ni-TED resin was used to purify the all the protein from the three fermenters. Harvesting was done at two time points (3 hr and 8 hr). The bed volume of the column was 240 ml. The crude fractions were loaded onto the column on separate runs. The eluted protein resulted in a single peak and was eluted with isocratic gradient (100% elution of the 250 mM imidazole). Figure A8 illustrates the purification chromatogram of the cell bench from the fermentation at 8 hr post induction.

The SDS-PAGE analysis showed a protein band corresponding with the estimated size of Endo8 in the crude lysate and a clear intense band can be seen in the eluted fraction after purification (Figure 3.29). A numeric value is attached to the protein through gel quantification and BSA was used as a standard marker. Protein quantification was conducted for all three fermenters and the two time points and data shown is only for one fermenter (fermenter 2) with both time points. Figure 3.30 illustrates the quantification of the protein post purification. Table 3.4 shows the average protein concentrations from wet and the protein recovered. The elution revealed a single band which corresponded with the molecular mass of endo8.

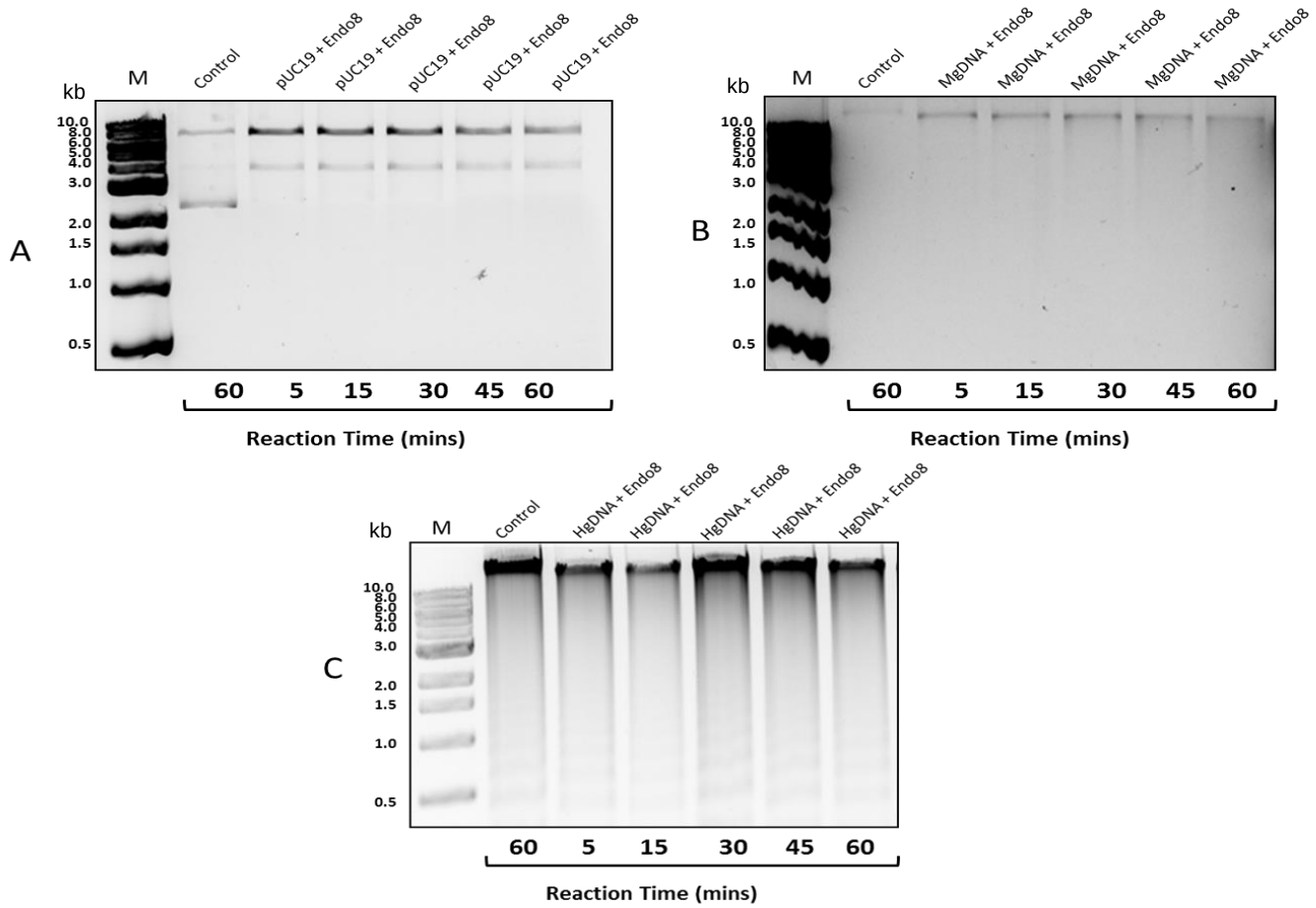



**Figure 3.26: Analysis of endo8 restriction activity on different plasmid for 60 min at various time intervals at the same temperature (37°C)**

The restriction digest reactions were all set of 60 min. M: 1 kB DNA ladder (Cat. #: N3232L). (A): pUC19; (B): MgDNA and (C): HgDNA is cleaved using endo8 at 37 °C. Reactions were set for different time intervals to observe the time the enzyme cleaves DNA and if difference can be observed with longer incubation period.

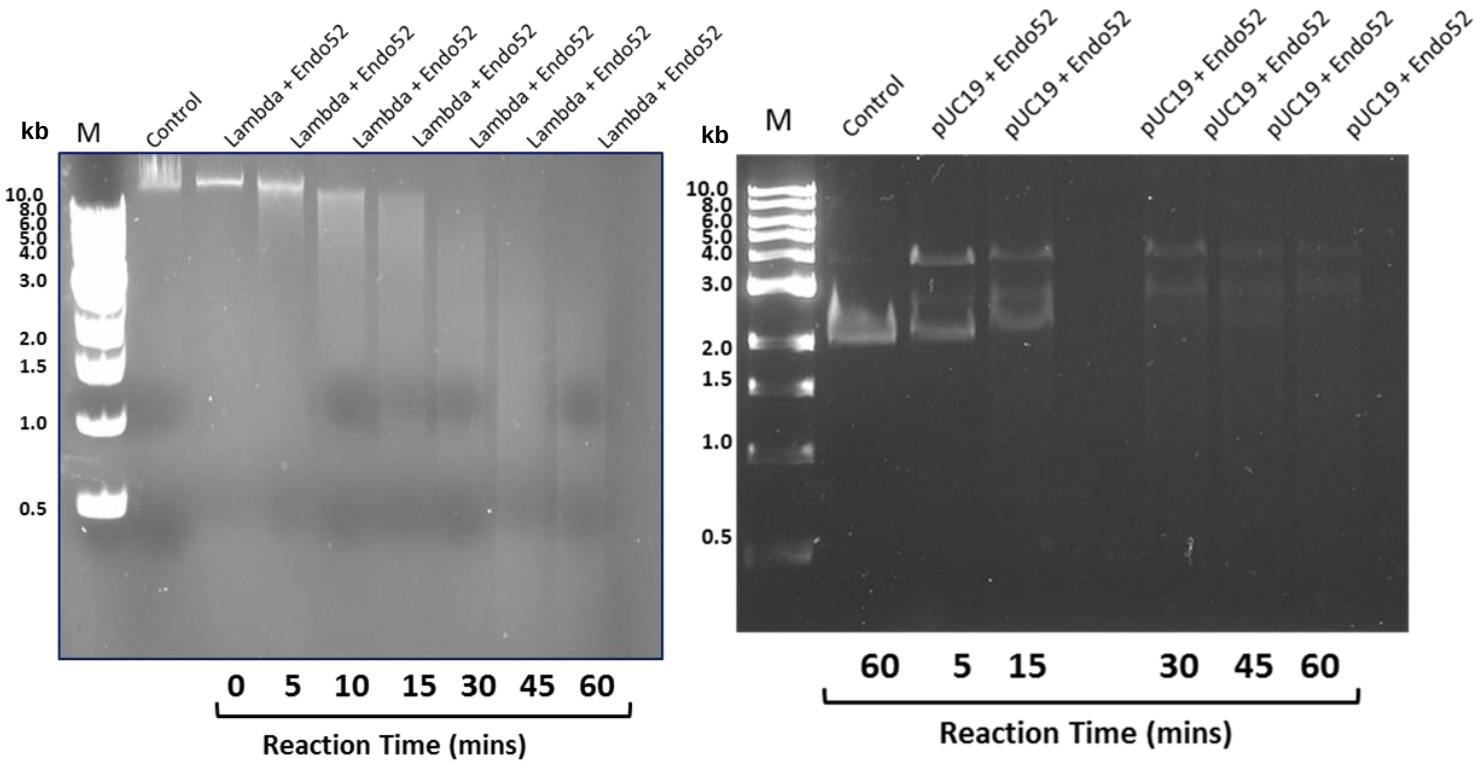



**Figure 3.27: Analysis of endo52 restriction activity on different plasmid for 60min at various time intervals at the same temperature (37°C).**

The restriction digest reactions were all set of 60 min. M: 1 kB DNA ladder (Cat. #: N3232L). (A): λDNA and (B): pUC19 is cleaved using endo52 at 37 °C. Reactions were set for different time intervals to observe the time the enzyme cleaves DNA and if difference can be observed with longer incubation period

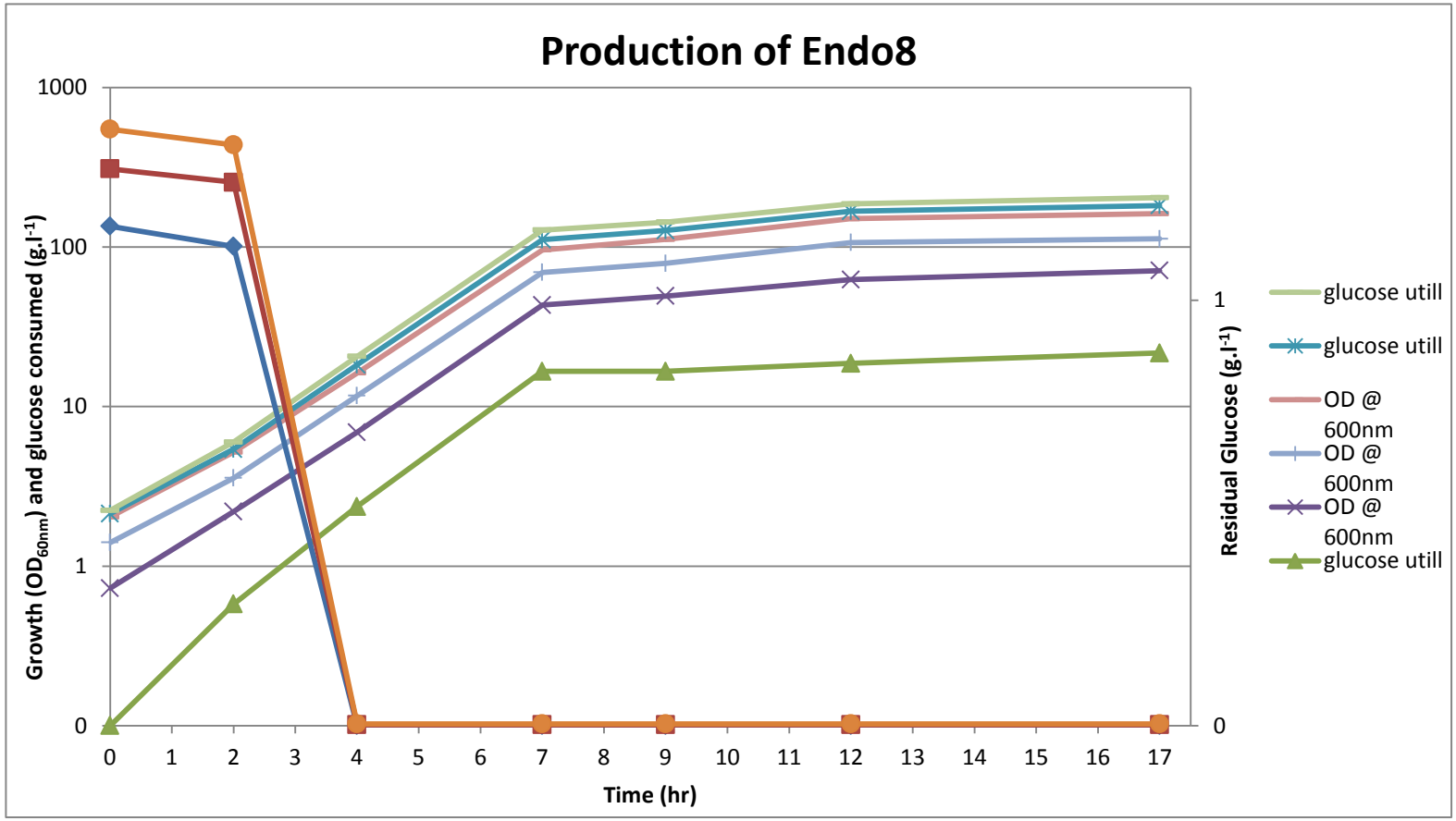



**Figure 3.28: The production of endo8, Rosetta™ (DE3) pLysS in triplicate in 2 l fed-batch fermentations.**

Cell growth is monitored in observing the OD over time and the increase in cell growth also shows an increase in the utilisation of glucose by the cells. The utilisation of glucose is also noted that as time passes, no glucose is detected. (■) Optical density, (●) Residual glucose and (▲) Glucose utilised. The scale on the primary and secondary axis is a logarithmic scale.

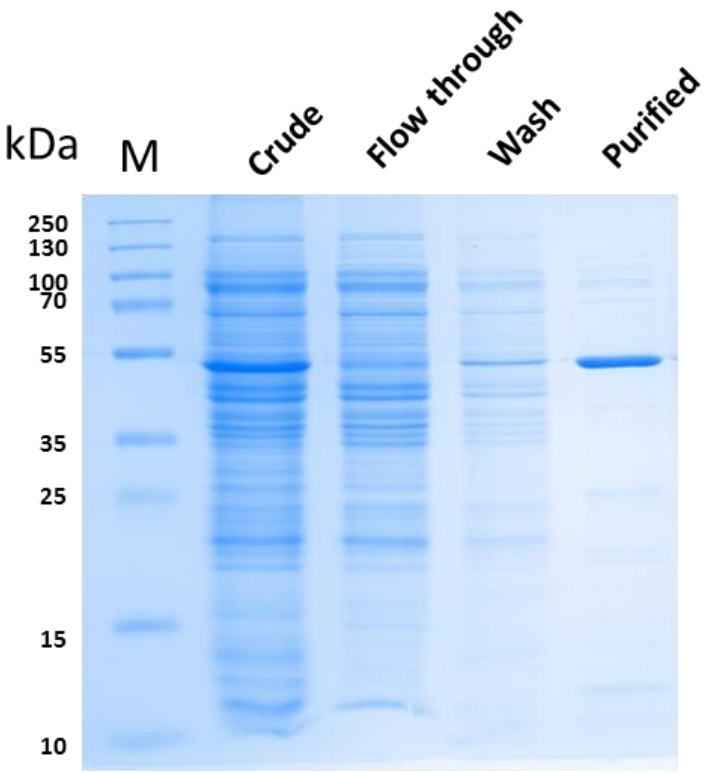



**Figure 3.29: Electrophoretogram of obtained fractions from automated protein purification of Endo8.**

M-Molecular weight marker; crude:obtained from lysis treatment of the pelleted cells; flow through: collection after protein binds to the column; wash:washing of the column with lysis buffer, purified: bound protein eluted with elution buffer with imidazole.

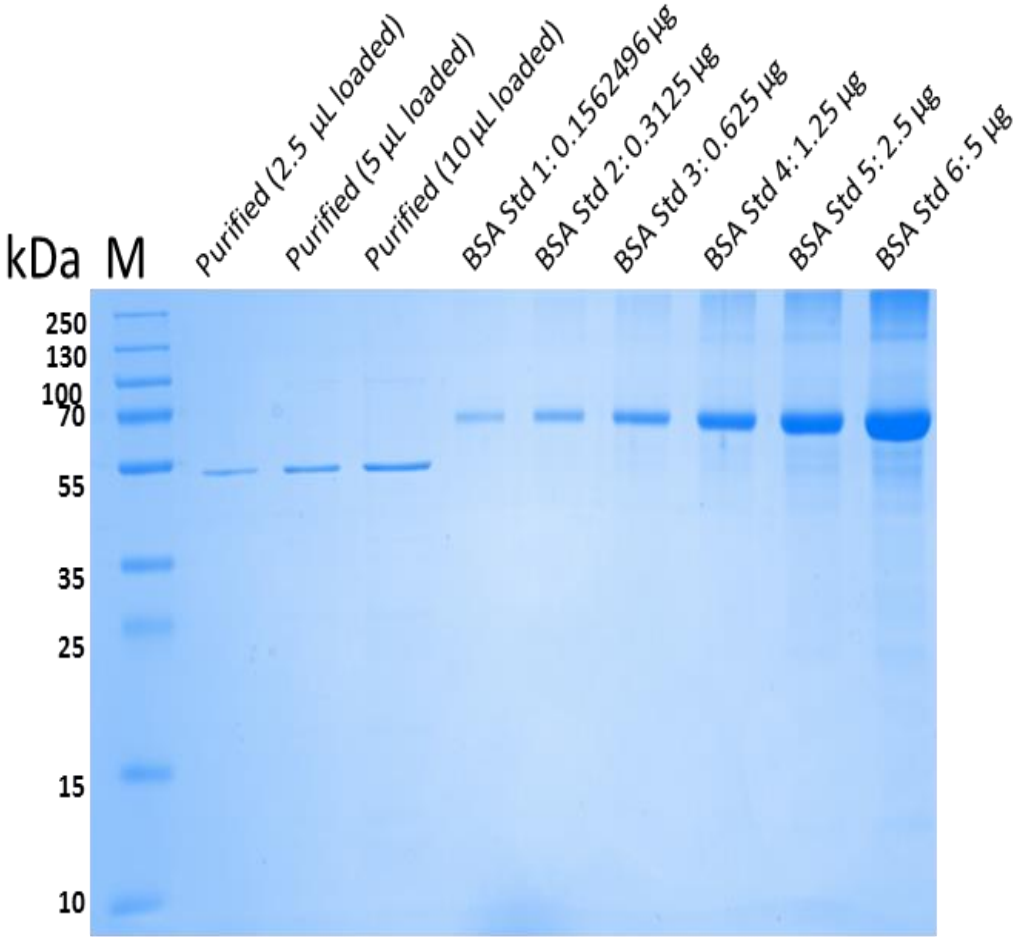



**Figure 3.30: SDS-PAGE gel quantification of purified fraction of Endo8.**

M-Molecular weight marker; different volumes loaded of the same purification fraction of Endo8 from fermenter 2 harvested at 8 hr post induction. BSA was used as a standard for densitometry-based quantitation.

Table 3.4 presents details of the different cell batch fraction (crude and eluate) and summaries the concentration yields attained from the recombinant production of the enzyme per time points post induction. It also shows the percentage recovery of the protein from wet cells until final step of purification and getting the enzyme in solution. Whilst Table 3.5 summarises the parameters of the fermentation at both harvesting using the feed rates describe in section 3.4.4. The final biomass achieved was 68.2 (±2.22) $g.l^{-1,}$ and a total protein of 446.35 $mg.l^{-1}$. The specific titre per biomass of endo8 was 83.33 $mg.g^{-1}$ (12 hr) and 58.82 $mg.g^{-1}$ (17 hr). However, the yield achieved was 2 fold greater at 17 hours.

**Table 3.4: Purification of Endo 8 from wet cells to enzyme solution.**

| | *Crude Volume (ml)* | *Mean Crude (mg.ml$^{-1}$)* | *Total Protein (mg)* | *Eluate Volume (ml)* | *Mean Eluate (mg.ml$^{-1}$)* | *Total Protein (mg)* | *Percent recovery* |
|---|---|---|---|---|---|---|---|
| **3 hrs** | 570 | 0.26 (±0.064) | 148.2 | 660 | 0.21 (±0.029) | 138.6 | 94% |
| **8 hrs** | 565 | 0.79 (±0.392) | 446.35 | 660 | 0.64 (±0.084) | 422.4 | 95% |

**Table 3.5 Performance parameters for Endo8 production in triplicate 2 l fed-batch fermentation**

| | **3 Hr** | | **8 Hr** | |
|---|---|---|---|---|
| **Parameter** | **Average value** | **StdError** | **Average value** | **StdError** |
| **Titer (g.l$^{-1}$)** | 0.26 | ±0.04 | 0.79 | ±0.23 |
| **Specific titer (mg.g$^{-1}$)** | 83.33 | 0 | 58.82 | 0 |
| **Yield on glucose consumed (mg.g$^{-1}$)** | 14.46 | ±2.39 | 37.95 | ±11.54 |
| **Productivity (mg.l$^{-1}$.h$^{-1}$)** | 21.70 | ±3.09 | 46.56 | ±13.33 |
| **Specific productivity (mg.g$^{-1}$.h$^{-1}$)** | 1.50 | ±0.03 | 1.23 | ±0.02 |

# 4 Discussion

## 4.1 Fosmid library construction and functional screening

The diversity of soil microbial communities is a potential “treasure-trove” of genes that can be used for various applications including the development of new drugs, improvement of existing ones as well as the isolation of new and improved reagent proteins for R&D. This soil environment of the Cape Floral Kingdom’s KBR was especially selected for its vast biodiversity. Studies by Stafford *et al*., (2005) and Segobola *et al*., (2018) show that the area has a lot of potential for the discovery of novel enzymes with various applications. The recent advances made in metagenomics have allowed the study of microorganisms from soil without the inherent biases of cultivation (Kakirde *et al.*, 2010).

According to Daniel (2005), Lakay *et al.,* (2007) and Wang *et al.,* (2009), the extraction of DNA directly from the soil can lead to the co-extraction of humic acid which can influence the quality and amount of DNA extracted from the soil sample (Lakay *et al.*, 2007; Zielińska *et al.*, 2017). These effects were minimized in this study through the use of commercial kits designed to counter co-extraction of humic acids. DNA of high molecular weight is crucial when constructing a metagenomic library. The commercial available kits have been designed to counter possibilities of obstructing the efficacy of cloning and transformation of DNA extracted directly from a soil sample. The advantage of achieving DNA of higher molecular weight when constructing a large insert metagenomic library is that genes that are specifically involved in the same biosynthetic pathways are normally clustered; large inserts thus make it feasible to clone entire pathways. Achieving DNA of high molecular weight is very crucial when constructing a metagenomic library. Comparative studies in Verma *et al*., (2017) show DNA yields acquired using different methods and the yields acquired in this study are within the ranges acquired in some of the techniques used and are of good quality. A very good yield was report by Tsai and Olson (1991) to be 746.46 $\mu g.g^{-1}$ of soil. The $_{met}$DNA was extracted and used to construct fosmid library using the pCC1FOS vector which resulted in library size of $\pm 1.83 \times 10^5$ cfu. Results from other studies also demonstrate the capability of cloning environmental genomic DNA into a fosmid as an approach that is independent of culturing Rondon *et al*., (2000). Jin *et al.*, (2012) constructed a fosmid library from a soil sample and achieved a library of ±23-25 kb with approximately 50,000 clones.

## 4.2 Sequence analysis, metagenome screening and recombinant production

A plate-based screening technique was developed and utilised to screen for potential restriction endonucleases. This technique was based on acquired phage immunity. Studies by Mann *et al*., (1978) and Walder *et al*., (1981) used known markers of restriction enzymes to screen for restriction enzymes using electrophoresis. Screening techniques previously used included the use of a crude lysate against (bacteriophage) lambda DNA in a restriction reaction. The results were visualised on an agarose gel and where there are patterns of DNA fragments, those clones would be selected for further study (Meyertons *et al.*, 1987). Sequence analysis of the positive clones revealed over 100 putative endonucleases and over 200 uncharacterised or hypothetical and unknown sequences (not reported). This confirms that the biosphere is a rich source of novel enzymes. A large number of restriction enzymes have been cloned and sequenced; however, there are little sequence similarities amongst them (Neely and Roberts, 2008). NCBI BlastP database sequence comparison of the endo20 enzyme revealed less than 50% sequence identity. Two subunits (M & R) for endo52 revealed high sequence identity to known sequences on the NCBI database, however the S-subunit had less than 50% sequence identity. Endo8's sequence identity was at 65% against known sequences on the NCBI-BlastP international database. Further analysis of the sequence data did not reveal any of the known motifs that restriction endonucleases are known to have.

The three selected genes were cloned into pET expression vectors and expressed in *E. coli* Rosetta™ (DE3) pLysS and BL21 (DE3) expression hosts. One of the newly isolated restriction enzymes from this study, designated 'endo8', demonstrated homology to two types of restriction enzymes in REBASE (type II and type IV restriction endonucleases). The reason that could explain why the sequence for endo8 has hits close to different types of endonucleases is mainly because of how diverse restriction endonucleases are. The second enzyme, designated endo20, had no similarity to any known restriction enzymes in the REBASE database. Endo52 was identified as a type I restriction enzyme, with all the respective subunits (R, M & S sub-units) required for its functionality (Wilson and Murray, 1991; Loenen *et al.*, 2014). The pETDuet vector was selected for cloning endo52; the vector has two multiple cloning sites(MCSs) which allows the cloning of multiple genes and have them expressed simultaneously to enable assembly into a single unit where necessary (Scheich *et al.*, 2007). This is a very useful system for the co-expression of more than one protein chain simultaneously.

All three genes were cloned with a His-tag fused into the N-terminal for easy downstream purification by IMAC charged nickel ions (Strickler, 2008). The expression of Endo8 and Endo20 was conducted without their corresponding MTase, while Endo52 was expressed with all its subunits. There were no cognate MTase genes/ORFs found in the contigs in which Endo8 and Endo20 were located. According to Zylicz-Stachula *et. al.,* (2009), not all restriction enzymes require co-expression with a MTase; they can be successfully expressed without it and retain full functionality. Having a cognate methyl transferase gene or using a strain with a methyl transferase during expression enables protection against host self-cleavage by the restriction endonucleases being expressed. Cell growth studies also showed a normal growth pattern (Figure 3.29), indicating that the lack of MTase in this case, did not affect the cell growth.

In this study, the proteins were expressed successfully in Rosetta™ (DE3) pLysS (Endo8 and Endo52-soluble, and Endo20-insoluble) (Figure 3.11, 3.16 and 3.20) and no expression was observed in BL-21 (DE3) cells for Endo20 (Figure A4-A5) and expression was observed in the insoluble fraction for Endo52 (Figure A74). A protein band migrating at 50.4, 88.0 and 100 kDa for Endo8, Endo20 and Endo52 respectively, were present in cells induced with IPTG (0.1 mM final concentration). Induction of the protein showed improved yields and the phenomenon of "leaky" expression was not observed in this study because if the vector selection (Briand *et al.*, 2016). The Rosetta (Merck) strains carry a chloramphenicol-resistant plasmid, pRARE, that contributes tRNAs for codons rarely used in *E. coli* (Berrow *et al.*, 2006). No further work was done with endo20 after expression optimisation resulted in inclusion bodies although the protein could have been solubilised using urea and purified and refolded for further studies; however, it has been shown that proteins that have been solubilised using harsh chaotropes tend to lose their functionality and to recover the enzymes' bioactivity from that step, can be tedious (Singh *et al.*, 2015; Fathi-Roudsari *et al.*, 2016). The process leading to formation of inclusion bodies is not totally understood (Ramón *et al.*, 2014). It is not uncommon for proteins expressed in bacteria to be misfolded or to be expressed in the inclusion bodies (Fathi-Roudsari *et al.*, 2016).

The two soluble proteins (Endo8 and Endo52) were purified successfully using IMAC purification. Endo8 was selected for upscaling using the fermenter. Fractions were taken at 3 and 8 hr post induction and were purified using the ÄKTA fast protein liquid chromatography (FPLC). (Camper and Viola, 2009). As the fermentation time is increased, the productivity increased. Therefore, should the protein be taken for market use, production will be conducted at 8 hours post induction. The average protein concentration as seen on Table 3, shows increased protein concentration yields. The approach of using fermenters for production is to

also to get improved yields (Botterman *et al.*, 1985). The fermentation of endo8 also yielded a greater concentration of the enzyme, compared to bench-top expression in a shake flask. The significantly higher fermentation yields was due to bioreactor's capacity to provide high dissolved oxygen and continuous nutrient feed in the growth medium, which increases the growth rate of the microorganism being cultivated (Ukkonen *et al.*, 2011). Studies from Czarnotta *et al*., (2017) using fed batch fermentation yielded 182 $mg.l^{-1}$ and 854 $mg.l^{-1}$. Another study for recombinant proteins hosted in *E. coli* showed a 100 % increase from the scale up via fermentation (Tripathi, 2016). In this study, we achieved a yield of 446.35 $mg.l^{-1}$ which three times greater than the yield achieved at 3 hrs (148.2 $mg.l^{-1}$).

## 4.3 Functional characterisation of the recombinant produced enzymes

Restriction enzymes can be used to linearize cloning vectors and to generate DNA fragments (Robinson *et al.*, 2001). This activity is dependent on the correct buffer compositions, temperature; time of incubation and for some restriction enzymes the presence of co-factors. Optimisation of reaction conditions for endonuclease activity is thus quite important. The initial steps taken on functional characterisation have illustrated that the enzymes isolated here were functional under different incubation conditions, despite having an average purity of 92 %. An average of 92% in terms of protein purification is relatively pure and with the protein in abundance, contaminants should be outweighed by the protein thus not affecting the activity of the enzymes in the assays. The activity displayed by these enzymes shows star activity while showing incomplete or partial linearization of pUC19. Incomplete digestion is not uncommon; a number of factors that can contribute to incomplete digestion including an inactive enzyme, suboptimal conditions and enzymes active sites being blocked (Robinson *et al.*, 2001). Temperature profile studies of the enzymes (Endo8 and Endo52) showed that they were both inactive at temperatures higher than 45 °C. Restriction endonucleases cleave DNA at specific regions and result in fragmented DNA. However, observations made from the study show strong non-specific endonuclease activity for both enzymes purified. The activity of the enzyme with other DNA preparations may have been negatively influenced by concentration of the DNA to be digested. Dilutions were done (1:10) but the results show that even that dilution could've been high. The activity observed in the study for both enzymes endonuclease activity against the various DNA plasmids may also be as a result of the technique used in this study to screen for novel restriction-modification systems. A study by Bao *et al*., (2008) shows similar activity; however, the non-specific endonuclease activity was observed in buffers with lower salt concentration.

The sequences that showed no sequence similarities could indicate the identification of novel genes of interest and other novel mechanisms for immunity to bacteriophage infection. A more robust technique is required to analyse these sequences so that novelty of the sequences can be reliably elucidated and have their functions explored.

# 5 Conclusion

In this study, a culture-independent metagenomic approach together with activity-based screening techniques was successfully used to identify and isolate novel restriction endonuclease enzymes. Metagenomic DNA was isolated from a soil sample collected from the Kogelberg Biosphere Reserve allowing for the construction of a fosmid library. The Kogelberg Biosphere Reserve is rich in plant diversity and comprehensive studies on the plant diversity have been conducted. Until recently, little was known about the microbial diversity that exists within the soil community of the Kogelberg Biosphere Reserve. This study focused on bio-prospecting of the soil in the environment in order to discover novel enzymes that can be used in molecular biology research. The screening technique relied on phage acquired immunity. Positive clones generated from the functional screening were sequenced using MiSeq Illumina NGS Illumina sequencing and sequence alignment was done using NCBI, MG-RAST and REBASE, leading to the discovery of three novel genes (endo8, endo20 and endo52).

Three genes were selected and cloned into a pET expression vector system. Rosetta™ (DE3) pLysS and BL-21 (DE3) expression host cells were used for expression studies. Expression optimization studies enhanced the expression levels of the recombinant proteins for Endo8 and Endo52. Optimisation studies for Endo20 resulted only in insoluble expression. Both Endo8 and Endo52 were selected for further purification and characterisation. Endo8 was further selected for production upscaling. The production of Endo8 at 17 hours (8 h post induction) yielded 58.82 $mg.g^{-1}$, with enzyme concentration of 0.79 $g.l^{-1}$.

The functionality of the enzymes was demonstrated in restriction endonucleases reactions using four different DNA preparations (HgDNA, MgDNA, pUC19 and Lambda DNA). Temperature profiling of the enzymes showed inactivity at temperatures above 45 °C. When used to digest pUC19 DNA, both enzymes characterised revealed incomplete linearization, whilst with the other DNA preparations non-specific digestion was observed. Another fact that might be considered is that the concentration of the DNA preparations might have been too high, which had a negative impact on the enzymes activity. A more in-depth investigation should be done to understand the complexity involved in the expression and activity of these restriction endonucleases. This study has successfully demonstrated that metagenomics together with functional screening can be successfully used in isolating novel enzymes from the soil sample metagenome, including endonuclease enzymes. Furthermore, characterization studies revealed that these enzymes could be potentially employed in molecular biology industry.

## 5.1 Future work

### 5.1.1 Functional screening for rare cutting restriction endonucleases

Future work will entail screening the fosmid library constructed in this study for rare cutters using modifications of the same principles for screening for restriction endonucleases. The technique used in this study shows success in finding frequent cutters. As such, for rare cutters, the colonies that would have to be picked for further investigation will be colonies that are taken later instead of colonies that grew overnight. Colonies that appear after three days of incubation will be chosen and sent for sequencing and further studied.

### 5.1.2 Endonucleases gene expression and endonuclease activity

More genes will be selected from the sequence annotations with similarities and will be recombinantly produced and characterised for endonuclease activity. Further investigation into endo20 will be carried out in order to optimise expression conditions. A more in-depth investigation is still needed to understand the complexity involved in the expression of restriction endonucleases. Additionally, more work needs to be done in order to identify recognition sites of the cloned and expressed endonucleases and to determine the most optimal conditions for their useful application in restriction digests in molecular biology. The continued development of novel restriction enzymes is crucial for the further development of improved or new molecular biology techniques.

### 5.1.3 Structural characterisation

The isolated genes will be structurally characterised for low and high resolutions. The low resolutions structural characterization will be done using a range of spectroscopic techniques. X-Ray crystallography will also be used for structural characterisation.

# 6 References


Adrio, J.L. and Demain, A.L. (2014) Microbial enzymes: tools for biotechnological processes. *Biomolecules* **4**: 117–39.

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., and Walter, P. (2002) Analyzing Protein Structure and Function. In, *Molecular Biology of the Cell*. Garland Science, New York, USA.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.* **215**: 403–10.

Ansai, S., Sakuma, T., Yamamoto, T., Ariga, H., Uemura, N., Takahashi, R., and Kinoshita, M. (2013) Efficient targeted mutagenesis in medaka using custom-designed transcription activator-like effector nucleases. *Genetics* **193**: 739–49.

Arber, W. (1978) Restriction Endonucleases. *Angew. Chemie Int. Ed. English* **17**: 73–79.

Bao, Y., Higgins, L., Zhang, P., Chan, S.-H., Laget, S., Sweeney, S., et al. (2008) Expression and purification of BmrI restriction endonuclease and its N-terminal cleavage domain variants. *Protein Expr. Purif.* **58**: 42–52.

Bashir, Y., Pradeep Singh, S., Kumar Konwar, B., Bashir, Y., Pradeep Singh, S., and Kumar Konwar, B. (2014) Metagenomics: An Application Based Perspective. *Chinese J. Biol.* **2014**: 1–7.

Belfort, M. and Roberts, R. (1997) Homing endonucleases: keeping the house in order. *Nucleic Acids Res.*

Bergmann, J.C., Costa, O.Y.A., Gladden, J.M., Singer, S., Heins, R., D'haeseleer, P., et al. (2014) Discovery of two novel β-glucosidases from an Amazon soil metagenomic library. *FEMS Microbiol. Lett.* **351**: 147–155.

Berrow, N.S., Büssow, K., Coutard, B., Diprose, J., Ekberg, M., Folkers, G.E., et al. (2006) Recombinant protein expression and solubility screening in *Escherichia coli* : a comparative study. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **62**: 1218–1226.

Bickle, T.A. and Kruger, D.H. (1993) Biology of DNA restriction. *Microbiol Rev* **57**: 434–450.

Blumenthal, R.M. and Cheng, X. (2002) Restriction-Modification Systems 2nd ed. John Wiley & Sons, Inc., New York, USA.

Botterman, J.H., De Buyser, D.R., Spriet, J.A., Zabeau, M., and Vansteenkiste, G.C. (1985) Fermentation and recovery of the EcoRI restriction enzyme with a genetically modifiedEscherichia coli strain. *Biotechnol. Bioeng.* **27**: 1320–1327.

Briand, L., Marcion, G., Kriznik, A., Heydel, J.M., Artur, Y., Garrido, C., et al. (2016) A self-inducible heterologous protein expression system in Escherichia coli. *Sci. Rep.* **6**: 33037.

Brown, T.A. (2002) Studying DNA. In, *Genomes*. Wiley-Liss, Oxford.

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

Camper, D. V and Viola, R.E. (2009) Fully automated protein purification. *Anal. Biochem.* **393**: 176–81.

Casali, N. (2003) Escherichia coli host strains. Clifton, N.J.

Certo, M.T. and Morgan, R.A. (2016) Salient Features of Endonuclease Platforms for Therapeutic Genome Editing. *Mol. Ther.* **24**: 422–9.

Česnavičiene, E., Mitkaite, G., Stankevičius, K., Janulaitis, A., and Lubys, A. (2003) Esp13961 restriction-modification system: Structural organization and mode of regulation. *Nucleic Acids Res.* **31**: 743–749.

Chen, C.-C., Wu, P.-H., Huang, C.-T., and Cheng, K.-J. (2004) A Pichia pastoris fermentation strategy for enhancing the heterologous expression of an Escherichia coli phytase. *Enzyme Microb. Technol.* **35**: 315–320.

Cheng, N.G., Hasan, M., Chahyo Kumoro, A., Ling, C.F., and Tham, M. (2009) Production of Ethanol by Fed-Batch Fermentation. *Pertanika J. Sci. Technol* **17**: 399–408.

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

Chung, D.-H., Huddleston, J.R., Farkas, J., and Westpheling, J. (2011) Identification and

characterization of CbeI, a novel thermostable restriction enzyme from Caldicellulosiruptor bescii DSM 6725 and a member of a new subfamily of HaeIII-like enzymes. *J. Ind. Microbiol. Biotechnol.* **38**: 1867–77.

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

Czarnotta, E., Dianat, M., Korf, M., Granica, F., Merz, J., Maury, J., et al. (2017) Fermentation and purification strategies for the production of betulinic acid and its lupane-type precursors in *Saccharomyces cerevisiae*. *Biotechnol. Bioeng.* **114**: 2528–2538.

Daniel, R. (2005) The metagenomics of soil. *Nat. Rev. Microbiol.* **3**: 470–478.

Davies, G.P., Martin, I., Sturrock, S.S., Cronshaw, a, Murray, N.E., and Dryden, D.T. (1999) On the structure and operation of type I DNA restriction enzymes. *J. Mol. Biol.* **290**: 565–79.

Delmont, T.O., Robe, P., Cecillon, S., Clark, I.M., Constancias, F., Simonet, P., et al. (2011) Accessing the soil metagenome for studies of microbial diversity. *Appl. Environ. Microbiol.* **77**: 1315–24.

Dewan, S.S. (2014) Life Science Tools and Reagents: Global Markets - BIO083B.

van Dijk, E.L., Auger, H., Jaszczyszyn, Y., and Thermes, C. (2014) Ten years of next-generation sequencing technology. *Trends Genet.* **30**: 418–426.

Divan, A. and Royds, J. (2013) Tools and techniques in biomolecular science Oxford University Press.

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

Engelberg-Kulka, H., Amitai, S., Kolodkin-Gal, I., and Hazan, R. (2006) Bacterial Programmed Cell Death and Multicellular Behavior in Bacteria. *PLoS Genet.* **2**: e135.

Espinoza-miranda, S.S., Gómez-rodríguez, J.A., and Jorge, A. (2012) Mining for Restriction Endonucleases in Nicaragua. *Encuentro Rev. Académica la Univ. Centroam.* **44**: 49–62.

Fathi-Roudsari, M., Akhavian-Tehrani, A., and Maghsoudi, N. (2016) Comparison of Three

Escherichia coli Strains in Recombinant Production of Reteplase. *Avicenna J. Med. Biotechnol.* **8**: 16–22.

Gaj, T., Gersbach, C.A., Barbas, C.F., and III (2013) ZFN, TALEN, and CRISPR/Cas-based methods for genome engineering. *Trends Biotechnol.* **31**: 397–405.

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

Hafez, M., Hausner, G., and Bonen, L. (2012) Homing endonucleases: DNA scissors on a mission. *Genome* **55**: 553–569.

Hall, T.A. (1999) BioEdit: A user-friendly biological sequence alignment program for Windows 95/98/NT. *Nucleic Acids Symp. Ser.* **41**: 95–98.

Handelsman, J. (2004) Metagenomics: application of genomics to uncultured microorganisms. *Microbiol. Mol. Biol. Rev.* **68**: 669–85.

Ichige, A. and Kobayashi, I. (2005) Stability of EcoRI restriction-modification enzymes in vivo differentiates the EcoRI restriction-modification system from other postsegregational cell killing systems. *J. Bacteriol.* **187**: 6612–21.

Jin, P., Pei, X., Du, P., Yin, X., Xiong, X., Wu, H., et al. (2012) Overexpression and characterization of a new organic solvent-tolerant esterase derived from soil metagenomic DNA. *Bioresour. Technol.* **116**: 234–240.

Jurica, M.S. and Stoddard, B.L. (1999) Homing endonucleases: structure, function and evolution. *Cell. Mol. Life Sci.* **55**: 1304–1326.

Jutur, P.P. and Reddy, A.R. (2007) Isolation, purification and properties of new restriction endonucleases from Bacillus badius and Bacillus lentus. *Microbiol. Res.* **162**: 378–383.

Kakirde, K.S., Parsley, L.C., and Liles, M.R. (2010) Size Does Matter: Application-driven Approaches for Soil Metagenomics. *Soil Biol. Biochem.* **42**: 1911–1923.

Karcher, S.J. (1995) Molecular Biology: A Project Approach 2nd ed. Academic Press, London.

Kobayashi, I. (2001) Behavior of restriction-modification systems as selfish mobile elements and

their impact on genome evolution. *Nucleic Acids Res.* **29**: 3742–3756.

Kobayashi, I. (2004) Restriction-Modification Systems as Minimal Forms of Life. *Nucleic Acids Mol. Biol.* **14**:.

Laemmli, U.K. (1970) Cleavage of Structural Proteins during the Assembly of the Head of Bacteriophage T4. *Nature* **227**: 680–685.

Lakay, F.M., Botha, A., and Prior, B.A. (2007) Comparative analysis of environmental DNA extraction and purification methods from different humic acid-rich soils. *J. Appl. Microbiol.* **102**: 265–273.

Lao, W.D. and Chen, S.Y. (1986) HsaI: a restriction enzyme from human being. *Sci. Sin. Ser. B, Chem. Biol. Agric. Med. earth Sci. / Chung-kuo k'o hsüeh yüan, chu pan.* **29**: 947–53.

Lepikhov, K., Tchernov, A., Zheleznaja, L., Matvienko, N., Walter, J., and Trautner, T.A. (2001) Characterization of the type IV restriction modification system BspLU11III from Bacillus sp . LU11. *Nucleic Acids Res.* **29**: 4691–4698.

Li, L.L., McCorkle, S.R., Monchy, S., Taghavi, S., and van der Lelie, D. (2009) Bioprospecting metagenomes: Glycosyl hydrolases for converting biomass. *Biotechnol. Biofuels* **2**:.

Lim, H.C. and Shin, S.H. (2013) Fed-Batch Cultures: Principles and Applications of Semi-Batch Bioreactors Cambridge University Press.

Liu, L., Li, Y., Li, S., Hu, N., He, Y., Pong, R., et al. (2012) Comparison of Next-Generation Sequencing Systems. *J. Biomed. Biotechnol.* **2012**: 1–11.

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

Longo, Ma.C., Smith, M.D., and Chatterjee, D.K. (2002) Cloning and expressing restriction endonucleases and modification methylases from caryophanon (93120847.4 (22)). 20.

Mann, M.B., Rao, R.N., and Smith, H.O. (1978) Cloning of restriction and modification genes in E. coli: the HbaII system from Haemophilus haemolyticus. *Gene* **3**: 97–112.

Mathur, E.J., Toledo, G., Green, B.D., Podar, M., Richardson, T.H., Kulwiec, M., and Chang, H.W. (2005) A biodiversity-based approach to development of performance enzymes: Applied metagenomics and directed evolution. *Ind. Biotechnol.* **1**: 283–287.

Meyertons, J.L., Tilley, B.C., Lechevalier, M.P., and Lechevalier, H.A. (1987) Actinophages and restriction enzymes from Micromonospora species (Actinomycetales). *J. Ind. Microbiol.* **2**: 293–303.

Moon, W.J., Cho, J.Y., and Chae, Y.K. (2008) Recombinant expression, purification, and characterization of XorKII: A restriction endonuclease from Xanthomonas oryzae pv. oryzae. *Protein Expr. Purif.* **62**: 230–234.

Mruk, I. and Kobayashi, I. (2014) To be or not to be: regulation of restriction-modification systems and other toxin-antitoxin systems. *Nucleic Acids Res.* **42**: 70–86.

Naito, T., Kusano, K., and Kobayashi, I. (1995) Selfish behavior of restriction-modification systems. *Science (80-. ).* **267**: 897–899.

Neely, R.K. and Roberts, R.J. (2008) The BsaHI restriction-modification system: cloning, sequencing and analysis of conserved motifs. *BMC Mol. Biol.* **9**: 48.

Oliveira, P.H., Touchon, M., and Rocha, E.P.C. (2014) The interplay of restriction-modification systems with mobile genetic elements and their prokaryotic hosts. *Nucleic Acids Res.* **42**: 10618–31.

Onstein, R.E., Carter, R.J., Xing, Y., and Linder, H.P. (2014) Diversification rate shifts in the Cape Floristic Region: The right traits in the right place at the right time. *Perspect. Plant Ecol. Evol. Syst.* **16**: 331–340.

Pascal, J.M. (2008) DNA and RNA ligases : structural variations and shared mechanisms. *Curr. Opin. Struct. Biol.* **18**: 96–105.

Pingoud, A., Fuxreiter, M., Pingoud, V., and Wende, W. (2005) Type II restriction endonucleases: structure and mechanism. *Cell. Mol. Life Sci.* **62**: 685–707.

Pingoud, V., Sudina, A., Geyer, H., Bujnicki, J.M., Lurz, R., Lüder, G., et al. (2005) Specificity changes in the evolution of type II restriction endonucleases: A biochemical and bioinformatic analysis of restriction enzymes that recognize unrelated sequences. *J. Biol.*

*Chem.* **280**: 4289–4298.

Pool-Stanvliet, R. (2013) A history of the UNESCO Man and the Biosphere Programme in South Africa. *S. Afr. J. Sci.* **109**: 01–06.

Ramón, A., Señorale-Pose, M., and Marín, M. (2014) Inclusion bodies: not that bad…. *Front. Microbiol.* **5**: 56.

Rimšeliené, R., Vaišvila, R., and Janulaitis, A. (1995) The eco72IC gene specifies a trans-acting factor which influences expression of both DNA methyltransferase and endonuclease from the Eco72I restriction-modification system. *Gene* **157**: 217–219.

Roberts, R.J., Belfort, M., Bestor, T., Bhagwat, A.S., Bickle, T.A., Bitinaite, J., et al. (2003) SURVEY AND SUMMARY A nomenclature for restriction enzymes , DNA methyltransferases , homing endonucleases and their genes. *Nucleic Acids Res.* **31**: 1805–1812.

Robinson, D., Walsh, P.R., and Bonventre, J.A. (2001) Restriction Endonucleases. In, *Molecular Biology Problem Solver: A Laboratory Guide*. Wiley-Liss, pp. 225–266.

Rondon, M.R., August, P.R., Bettermann, A.D., Brady, S.F., Grossman, T.H., Liles, M.R., et al. (2000) Cloning the soil metagenome: a strategy for accessing the genetic and functional diversity of uncultured microorganisms. *Appl. Environ. Microbiol.* **66**: 2541–7.

Rosano, G.L. and Ceccarelli, E.A. (2014) Recombinant protein expression in Escherichia coli: advances and challenges. *Front. Microbiol.* **5**: 172.

Rothwell, P.J. and Waksman, G. (2005) Structure and mechanism of DNA polymerases. *Adv. Protein Chem.* **71**: 401–440.

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

Scalley-Kim, M., McConnell-Smith, A., and Stoddard, B.L. (2007) Coevolution of a homing endonuclease and its host target sequence. *J. Mol. Biol.* **372**: 1305–19.

Scheich, C., Kümmel, D., Soumailakakis, D., Heinemann, U., and Büssow, K. (2007) Vectors for co-expression of an unrestricted number of proteins. *Nucleic Acids Res.* **35**: e43.

Schoenfeld, T., Liles, M., Wommack, K.E., Polson, S.W., Godiska, R., and Mead, D. (2010) Functional viral metagenomics and the next generation of molecular tools. *Trends Microbiol.* **18**: 1–19.

Scholz, M., Lo, C.-C., Chain, P.S.G., Scholz, M.B., Miller, J.R., Earl, D., et al. (2014) Improved Assemblies Using a Source-Agnostic Pipeline for MetaGenomic Assembly by Merging (MeGAMerge) of Contigs.

Segobola, J., Adriaenssens, E., Tsekoa, T., Rashamuse, K., and Cowan, D. (2018) Exploring Viral Diversity in a Unique South African Soil Habitat. *Sci. Rep.* **8**: 111.

Singh, A., Upadhyay, V., Upadhyay, A.K., Singh, S.M., and Panda, A.K. (2015) Protein recovery from inclusion bodies of Escherichia coli using mild solubilization process. *Microb. Cell Fact.* **14**: 41.

Singh, J., Kaushik, N., and Biswas, S. (2014) Bioreactors – Technology & Design Analysis. **01**:.

Smith, H.O. and Welcox, K.W. (1970) A Restriction enzyme from Hemophilus influenzae: I. Purification and general properties. *J. Mol. Biol.* **51**: 379–391.

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

Stoddard, B.L., Gwiazda, K., Humbert, O., Mandt, T., Pangallo, J., Brault, M., et al. (2014) Homing endonucleases from mobile group I introns: discovery to genome engineering. *Mob. DNA* **5**: 7.

Strickler, O.U.M.S.D.C.J. (2008) Fusion Tags for Protein Expression and Purification. *Biopharm Int.*

Szalay, A.A., Mackey, C.J., and Langridge, W.H.R. (1979) Restriction endonucleases and their applications. *Enzyme Microb. Technol.* **1**: 154–164.

Tao, T., Bourne, J.C., and Blumenthal, R.M. (1991) A family of regulatory genes associated with type II restriction-modification systems. *J. Bacteriol.* **173**: 1367–75.

Taylor, G.K. and Stoddard, B.L. (2012) Structural, functional and evolutionary relationships between homing endonucleases and proteins from their host organisms. *Nucleic Acids*

*Res.* **40**: 5189–200.

Tock, M.R. and Dryden, D.T.F. (2005) The biology of restriction and anti-restriction. *Curr. Opin. Microbiol.* **8**: 466–472.

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

Tripathi, N.K. (2016) Production and Purification of Recombinant Proteins from *Escherichia coli*. *ChemBioEng Rev.* **3**: 116–133.

Tsai, Y.L. and Olson, B.H. (1991) Rapid method for direct extraction of DNA from soil and sediments. *Appl. Environ. Microbiol.* **57**: 1070–4.

Uchiyama, T. and Miyazaki, K. (2009) Functional metagenomics for enzyme discovery: challenges to efficient screening. *Curr. Opin. Biotechnol.* **20**: 616–622.

Ukkonen, K., Vasala, A., Ojamo, H., and Neubauer, P. (2011) High-yield production of biologically active recombinant protein in shake flask culture by combination of enzyme-based glucose delivery and increased oxygen transfer. *Microb. Cell Fact.* **10**: 107.

Vasu, K. and Nagaraja, V. (2013) Diverse functions of restriction-modification systems in addition to cellular defense. *Microbiol. Mol. Biol. Rev.* **77**: 53–72.

Verma, S.K., Singh, H., and Prakash, C.S. (2017) An improved method suitable for isolation of high-quality metagenomic DNA from diverse soils. *J. Biotechnol.* **7**:.

Voelkerding, K. V, Dames, S.A., and Durtschi, J.D. (2009) Next-Generation Sequencing: From Basic Research to Diagnostics. *Clin. Chem.* **55**: 641–658.

Wahler, D. and Reymond, J.-L. (2001) Novel methods for biocatalyst screening. *Curr. Opin. Chem. Biol.* **5**: 152–158.

Walder, R.Y., Hartley, J.L., Donelson, J.E., and Walder, J.A. (1981) Cloning and expression of the Pst I restriction-modification system in Escherichia coli. *Proc. Natl. Acad. Sci. U. S. A.* **78**: 1503–7.

Wang, Y., Morimoto, S., Ogawa, N., Oomori, T., and Fujii, T. (2009) An improved method to

extract RNA from soil with efficient removal of humic acids. *J. Appl. Microbiol.* **107**: 1168–1177.

Webster, G.R., Teh, A.Y.-H., and Ma, J.K.-C. (2017) Synthetic gene design-The rationale for codon optimization and implications for molecular pharming in plants. *Biotechnol. Bioeng.* **114**: 492–502.

Wei, H., Therrien, C., Blanchard, A., Guan, S., and Zhu, Z. (2008) The Fidelity Index provides a systematic quantitation of star activity of DNA restriction endonucleases. *Nucleic Acids Res.* **36**: e50–e50.

Weiserová, M. and Ryu, J. (2008) Characterization of a restriction modification system from the commensal Escherichia coli strain A0 34/86 (O83:K24:H31). *BMC Microbiol.* **8**: 106.

Wilkinson, R. and Wiedenheft, B. (2014) A CRISPR method for genome engineering.

Williams, R.J. (2003) Restriction endonucleases: classification, properties, and applications. *Mol. Biotechnol.* **23**: 225–243.

Wilson, G.G. and Murray, N.E. (1991) Restriction and modification systems. *Annu. Rev. Genet.* **25**: 585–627.

Wilson, G.G., Wang, H., Heiter, D.F., and Lunnen, K.D. (2012) Restriction Enzymes in Microbiology , Biotechnology and Biochemistry. *Encuentro Rev. Académica la Univ. Centroam.* 19–48.

Wilson, M.C. and Piel, J. (2013) Metagenomic Approaches for Exploiting Uncultivated Bacteria as a Resource for Novel Biosynthetic Enzymology. *Chem. Biol.* **20**: 636–647.

Yadava, A. and Ockenhouse, C.F. (2003) Effect of codon optimization on expression levels of a functionally folded malaria vaccine candidate in prokaryotic and eukaryotic expression systems. *Infect. Immun.* **71**: 4961–9.

Zhao, L., Pellenz, S., and Stoddard, B.L. (2009) Activity and specificity of the bacterial PD-(D/E)XK homing endonuclease I-Ssp6803I. *J. Mol. Biol.* **385**: 1498–510.

Zielińska, S., Radkowski, P., Blendowska, A., Ludwig-Gałęzowska, A., Łoś, J.M., and Łoś, M. (2017) The choice of the DNA extraction method may influence the outcome of the soil

microbial community structure analysis. *Microbiologyopen* **6**:.

Zylicz-Stachula, A., Bujnicki, J.M., and Skowron, P.M. (2009) Cloning and analysis of a bifunctional methyltransferase/restriction endonuclease TspGWI, the prototype of a Thermus sp. enzyme family. *BMC Mol. Biol.* **10**: 52.

# 7 Appendix

Table A1: Fermentation Batch Recipe

| 1 - IC SOLUTION A (Salts and Other) | | | |
|---|---|---|---|
| **Descriptor** | **Unit** | **[Unit/IC] U/L** | **Total U** |
| Citric acid | g | **2.5** | 2.75 |
| $NH_4NO_3$ | g | **5** | 5.5 |
| $(NH_4)_2SO_4$ | g | **2** | 2.2 |
| $Na_2HPO_4.2H_20$ | g | **4.5** | 4.95 |
| $K_2HPO_4$ | g | **14.6** | 16.06 |
| Antifoam | ml | 1 | 1.1 |
| **-** | | | |
| Make up to Volume (tap water) | ml | | 981 |

| 2 - IC SOLUTION B (Protein Source) | | | |
|---|---|---|---|
| **Descriptor** | **Unit** | **[Unit/IC] U/L** | **Total U** |
| | | | |
| | | | |
| Yeast Extract (Biolab) | g | 20 | 22 |
| | | | |
| | | | |
| | | | |
| | | | |
| - | | | |
| Make up to Volume (tap water) | ml | 100 | 110 |

| 3 - FEED SOLUTION CARBON SOURCE | | | |
|---|---|---|---|
| **Descriptor** | **Unit** | **Value** | **Total U** |
| ***Feed*** | | | |
| Target feed Mass | **g** | | 393 |
| ***Solution 1*** | | | *99* |
| Glucose Monohydrate | %m/m | 45 | 48 |
| $MgSO_4$ | g | 1.43 | 1.57 |
| Make up water | g | 41 | **48.8** |
| ***Solution 2*** | | | *294.1* |
| Yeast extract (NEW) | g | 17.51 | 21.0 |
| Tryptone (Vegetable ) | g | 8.75 | 10.5 |
| | | | |
| Make up to Mass (tap water) | | 218.84 | **263** |

| 4 - FEED SOLUTION pH CONTROL | | | |
|---|---|---|---|
| **Descriptor** | **Unit** | **Value** | **Total U** |
| ***Base*** | | | |
| Target Feed Mass | g | | 100 |
| $NH_4OH$ Solution (25%N) | %m/m as is | 100 | 100 |
| Make up to Mass (tap water) | g | 0 | 0 |
| | | | |
| | | | |
| ***Acid*** | | | |
| Target Feed Mass | g | | 100 |
| $H_2SO_4$ | %m/m | 20 | 21.7 |
| | | | |
| Make up to Mass (tap water) | g | | 78.3 |

| 5 - FEED SOLUTION OTHER 1 | | | |
|---|---|---|---|
| **Descriptor** | **Unit** | **Value** | **Total U** |
| ***Antifoam*** | | | |
| Target Feed Mass | g | 250 | 250 |
| Durapol 3000 | %m/m | 50 | 125 |
| | | | |
| | | | |
| Make up to Volume (tap water) | g | 125 | 125 |

| 6 - INDUCTION SOLUTION | | | |
|---|---|---|---|
| **Descriptor** | **Unit** | **Value** | **Total U** |
| ***Inducer*** | | | |
| Target Inducer Volume | | | |
| | | | |
| 1 M IPTG | ml | 1 | 1.2 |
| | | | |
| | | | |

| 7 - TRACE ELEMENT STOCK SOLUTIONS | | | |
|---|---|---|---|
| **Descriptor** | **Unit** | **Value** | **Total U** |
| Target Volume | ml | 100 | 1000 |
| $CaCl_2.2H_2O$ | g | 0.04 | 0.4 |
| $FeCl_3.6H_2O$ | g | 1.67 | 16.7 |
| $MnCl_2.4H_2O$ | g | 0.015 | 0.15 |
| $ZnSO_4.7H_2O$ | g | 0.018 | 0.18 |
| $CuCl_2.2H_2O$ | g | 0.0125 | 0.125 |
| $CoCL_2.6H_2O$ | g | 0.018 | 0.18 |
| $Na_2EDTA$ | g | 2.01 | 20.1 |
| - | | | |
| - | | | |
| - | | | |
| - | | | |
| Make up to Volume (tap water) | ml | 100 | 1000 |

| 8 - INOCULUM FLASKS | | | |
|---|---|---|---|
| | | | |
| **Descriptor** | **Unit** | **Value** | **Total U** |
| ***Flask Ingredients*** | | | |
| Target Volume | ml | 1000 | 100 |
| NaCl | g | 5 | 0.5 |
| Yeast Extract (Biolab) | g | 5 | 0.5 |
| Tryptone meat free | g | 10 | 1.0 |
| - | | | |
| Make up to Volume (tap water) | mL | 1000 | 100 |
| ***Antibiotic*** | | | |
| Target Volume | ml | 0 | 0 |
| 50 $mg.ml^{-1}$ Kanamycin | ml | 1 | 0.1 |
| 34$mg.ml^{-1}$ Chloramphenicol | ml | 1 | 0.1 |
| Make up to Volume (tap water) | ml | 0 | 0 |

| BATCH PARAMETERS | | | |
|---|---|---|---|
| **Descriptor** | **Unit** | **Value** | **Actual** |
| **Initial Charge** | l | 1.1 | |
| **Inoc Volume** | l | 0.1 | |
| **Initial Volume** | l | 1.2 | |
| **Make Up Water** | l | | |
| **Target Run Age** | h | 26 | |

| INOC INFO | | | |
|---|---|---|---|
| **Descriptor** | **Unit** | **Value** | **Actual** |
| **Flask Volume** | l | 0.1 | |
| **Cryovial Designation** | # | | |
| **Check Flask Age** | h | 14 | |
| **Target OD** | ABS(600nm) | 7 | |
| **Target Transfer Age** | h | 14.5 | |

| PHYSICAL CONTROL PARAMETERS | | | | |
|---|---|---|---|---|
| **Descriptor** | **Unit** | **Mode** | **Setpoint** | **Instruction** |
| **Temperature** | deg C | auto | 30 - 25 | Decrease temperature to 25 deg C after Induction |
| **Stirrer** | rpm | cascade | 500 | Control $PO_2$ >40%sat |
| **pH** | # | auto | 7.00 | control with $H_2SO4$ and $NH_4OH$ in auto mode |
| **$PO_2$** | %sat | auto | 35 | increase stirrer when $PO_2$ <30%sat |
| **Antifoam** | ml/hr | off | on dem | maually pump when faom detected for 10 sec. |
| **Airflow** | slpm | auto | 1.2 | - |
| **Pressure** | mBar | none | | manually adjust to SP, check 4 hrly |
| **Glucose feed** | g/hr | setpoint | 11.7 | start after IC glucose depletion |
| **Booster feed** | g/hr | setpoint | 11.7 | Start after 53 g of Glucose feed has been fed |

Table A2: List of putative Endonucleases ORFs isolated from the screening procedures.

| | ORF | Position | Query Cover (%) | Identity (%) | Closest Hit |
|---|---|---|---|---|---|
| contig_52 | MTKGTATELYQALWNSADVLRSKMDASEYKNYLLGLIFYK YLSDTMLVHSSEMLEEKTENLNEALDMYREAYADDEFSE EFQSALVYEMSYRIKPELTFSALMEEINNHTFQREHLQQG LRDIEQSSNVFEDLFEDIDLNSKKLGATPQKQNDTISQVM KALDNLNLANYDGDALGDAYEYLIGQFAEDSGKKAGEFYT PSQVSTLMTRIALANKEDKKGLTVYDPTMGSGSLLLNASK YSNEASTIRYFGQELNTSTYNLARMNMFLHNVDPENQTL RNGDTLDADWPQDEPTNFDAVLMNPPYSAKWSAAKGFL DDPRFASYGVLPPKSKADFAFLLHGYFHLKNDGKMAIVLP HGVLFRGAAEGKIRKALLEQGAIDTVIGLPANIFFNTSIPTT VVILKKNRDSRDVLFIDASNEFTKAKSQNKLEEKHLDKIYE TYLKRENVEKYAHVATYEEIEENDFNLNIPRYVDTFEEAEP IDVVALKDEMKQTDQEIEDVSKELLAMVDDLEVTADTKDII DALKEVLG | 32578 to 34164 | 100 | 99 | HsdM (Type I M-Subunit |
| contig_52 | MGIDKASQKAPNLRFKGFTDDWEQRKLSNLGNLNRGKSK HRPRNDPKLYGGEYPFIQTGDVATAGLFLKDFKQTYNTLG ISQSKLWKKGTLLITIAANIAETSILSIDAAFPDSIIGFESNQV DMIFIKSVLDNSNKKIKSKAETSSQSNLNLSKLSELDIWVPT LGEQKKIGFLFQKLDNTIDLHQRKLDQLNQLKEALLQQMF PGKGETVPKLRFAGFEGEWEERKLKDVSDMYDNLRVPV TASDRIAGETPYYGANGIQDYVKDYTHIGEFVLIAEDGAND LVNYPVHYVTGEVWVNNHAHVVSAIEGKLDNLFLVSRLKS MNFIPWLVGGGRAKLNGDVLKKLPIIIPHLEEQQKIGSFFK HLDNTISVQQRKLDQLKNMKQVLLQNMFI | 34167 to 35348 | 97 | 48 | HsdS (Type I S-subunit |

| | | | | | |
|---|---|---|---|---|---|
| **contig_52** | MEDKLIDQLVNGESQWTRRPDLKNEEALWDNFRQILTRN<br>NKDKLNDVPLTDTEFDQVKTQLNFGSFYRAAEWLKGENG<br>IAQVLVQREDAKLGKVSLTVFKSQDISGGISVYEVINQYAS<br>SKRDEQDRNRRFDVTLLINGLPLIQIELKNRSEGYMAAFE<br>QIKKYSNEGKYTGIFSMLQMFVVTNGVDTKYIAAADGQHI<br>NKEFLTSWVDKNNNRVNNYLGFAKEVLSIPQGHKMITEFS<br>VLDADHKAIILLRPYQIHAILAVEQAVRQRQSGYVWHTTGS<br>GKTLTSYKVARNLLRSPALDKTIFIVDRIDLDQQTGTAFKS<br>YAMNDVVEVNDTDNVSDLVRKLTANDRDLVITTIQKLNYV<br>MKRYGDKEDNRIAKKLRNLNIGFVVDECHRAVTPKKQQEI<br>TKFFPKSLWYGFTGTPIFAENARDEFGDMPRTTEEQYGP<br>RLHEYTVKEAIHDKAVLGFQVEYINTLEKNSIVDYFDQSGI<br>DIDGLSEQEIEAKLPREAYENDEHKLKVIDQIVNYSRHKFK<br>LTRGPGNTYSAILTTRSIPDAQRYYELFNEVKEGKSSVKIS<br>EKTKSLVSDFPKVAVTYSLNETEETSFERQSQYKQIIQDY<br>NETFNTHYDLEQVRAFNQDINNRLARKRDIYRTRSEQLDI<br>VIVVDRLLTGFDSPSTAILFMDRPPAKPHHLIQAFSRTNRIY<br>DKDKQYGQIMTFQYPIQYQEAVENAFILYSNGGESAIQAP<br>GWNESYGRFQDAYERLISVAPSPESIDINDDVDTLKLFVK<br>AYQEFDKSLGAIQVYSEFDEDMFEQQYNLRPEVIESYHGK<br>YENALEKIREQIDEDEEDDLTIDFDYQLSEVGKQQIDYEYL<br>MLLMQTIVNEPNSQNRIKRIVDAEAYLTKFKDNNPKLGEIIE<br>DIFSTIKDGNELTKAEGTLNVASEIEKRIDERVSELVHDLSE<br>SLYVQEDDLHYLIENYKPEKAGKQTGESNLLENMDKDRYL<br>EENRKKVQKKFRVKKFAKQAYTDVIESEILPLTNKNF | 35477 to 38497 | 100 | 99 | HsdR (Type I R-subunit) |

| | | | | | |
|---|---|---|---|---|---|
| **Contig_20** | MPGIRPKKRIRQTSNLQAQHTIETIMDYDFKTLSPEDFERLI<br>GDLYYAETKVLPQSFKSGKDGGIDLLVTDGNGHEKVIIQC<br>KRYEPSAIAALKKAMQKEKGNNLDKLRPPRYILATSVKLSP<br>QNKKNLQKDLHPWIRDIGDIWGLDDINARLRLNEDIEKKHI<br>KLWISSTAVLEKILNHNILSITDITIEKIRQSFAKLVIHSAFDE<br>CYRKLEHSHSCIITGNPGIGKTTLANLLLCRYIKEGFTPIVA<br>TNGIHEIFGLIKSQEKNKAIIYYDDFLGATRYNELKFSKNED<br>AELLTLLDHAKRSDHLRFIMTSRDYIIEDAKSNHRHFQDYA<br>DQITRHTIEVGDYTKLHRAKILYNHLFFSDLPKEKIKHLIESK<br>IYAEIINQEYFIPRVIATICKDANSHSLCNSEFIDYIRQEIANP<br>VSIWQHPFENEISATARLVLLTTWTFGGTTTTSCLKKIITEL<br>QPEHERYDSNLKLNKALKELSANFITLTNMLPKWEGDDP<br>QVIVKFQNPSIEDYINGLIQASPDLLTPKTIKFFKQFENLSIN<br>LPSTRHHTSNLTRLIVDILDRFPEIERTETGRVITTEDNRQL<br>YNSHDTICIADRTISYLKLLIKLRTPPNETQTAERITTSTGW<br>LELMGGHTLKEFDSYGVERLVQWLSNNLTAIPGVLEKKIT<br>QSLSEAISIIANNINAWCTSLRAVSCIATCISHLSLTISESTLS<br>SLVNAALKHGRTAIFSDRSEYPFAYSEMLAISNAISHVEIEK<br>IALSLSHGSVKAEHTETPKTLEKPQHEGDEVNMDLDHYYK<br>TLLHSLV | 105311-107713 | 62 | 41 | Restriction Endonuclease |
| **Contig_12** | MSQGSTELPNSYVDSTRLPVGPQLWGTHPGASFQLVSV<br>FEPGRLDIDLIVQTENGKLGVQKAHEQRLATLQQDLDQAL<br>RDQAGSLNGLGEIQATERTIAAINELQALFATRLQATRDQA<br>YSFFGGDPVNRNLHEFLAVARKPNAPADPQQACLDSYRA<br>AFDVKYLEQVSGELAQRQQQLQSTLSTLRMRESHDQDID<br>RTLQQYHQALAQLGTASQAFLQASSALKAMLTAFERHQQ<br>ESVAPQAALEHSRELERETSDLLKAREQLASTYSSMNLS<br>QGHLASQATWIDSALQYHAPEGSADRKAALLNVQAQLQA<br>QIAAHTQTSPAESTEVMRLLAQTDNTVATVFDEQTSLAVL<br>SGRTPGNTPVTFNAWVASIHHPLILATSRGGIAHFEPIWLD<br>FGEALGKAAQRLLAGGLLAVARYVPLMLYSARLGDAERM<br>GVTVPLALMSPGADLTLEANRKAGQTLELPLRMDAVPQG<br>MQTEVYLAATDGSALLRDVRVRQAQWDAAQGAYRFTAE<br>GPGGATLLWHPATPPSTLSATGENGTVIPGFPIVEDLQKH<br>LPGPIIVPPDPDIRTLPELPELQIDDYVIIFPADSGLAPIYVML<br>RNPRNLPGVASGNGVITPDRVLDAATTAAGAPIPAKIAERL | 103853-101664 | 100 | 99 | HNH Endonucelase |

| | | | | | |
|---|---|---|---|---|---|
| | RGRRFARFDRLKEAIWMEIATDEVFSRHIMPAILDDMRRG LAPFAKRDQRVGKRVKLEIHHKHEIAKGGAVYDFDNLMF MTPQVHINHHRGNNQ | | | | |
| **Contig_25** | MLVRLVAYHHQRSQIAVFHLGEVAALGNAPMVNQRHLLA GSKHIHLLVGLAIAHRDGVSDDGHHSAYRRQLPHGDQIVF GDLARRSAAGRRFAAHIENGCSHGAHLLDHFAPCALAYG QHDDHRGHADHDAQQGQAGAHPVHAHTAPGTACSFGN LSHPGYCRRCGVALRGQWRNLCSALNAIGSTVIDDTSIAD FDDAPCLSRNFAVMRDHNDGVSRIGQLAQQSHHLGAAM RVKRPGGLVGQNDVAAIHQRPRNGHALLLTAR | 193597 - 192803 | 20 | 35 | Restriction Endonuclease |
| **Contig_16** | MRRFRTARGIGLHQPQVNEHHLHAPGLPEDGRLRLLSFNI QVGISTERYRHYVTRSWQHLLPHNGRAGNLQKIGQLLGD FDLVALQEADGGSLRSGYVNQVEHLAHLGAFPYWYQQL NRNLGRFAQHSNGVLSRLKPQLLEDHPLPGPAGRGAILV RFGEGEDALIVVMMHLALGAKTRALQLGYIRELIGGYRHQ VLMGDMNTHATDLLEHSPLRDLGLIAPQVEATFPSWRPQ RCLDHILLSPSLTLERVEVLAQPISDHLPVAVEIRLPDALTV DTLPVLS | 79928-80779 | 100 | 100 | Endonuclease |
| **Contig_8** | MLTPQFVLPVHTELVVDLFAGGGGASTGIEQAIGRHVDIA VNHDPEAVSLHTANHPQTRHFCSDVFEVDPLAVTEGQPV GLLWASPDCKHFSKAKGGKPVSKKIRGLAWVVIKWAKLT RPRVICLENVEEFQTWGPLGVDSRPCPERKGQTFQRWV SQLRNLGYKVEWKELRACDFGAPTIRKRLFLVARRDGLPI SWPQPTHAQPDESGKVAKGFKPWRTAAECIDWSIAAPSI FERERPLADATCRRIAKGIDRYVVKTAKPFIVGDSAPFLTE HANASTQRTFRADEPLRTQVAQIKGGHFALAVPTLVQTG YGERAGQAPRVPGLDKPLGTVVGSPKHALVQAFLAKHYT GVVGSDLQDPIGTVTSVDHHSLVTAHMTKFRAGSIGSAAD EPLHTVTAGGTPARPSTGNTMGLVTANLVHLGHGEGKDG TKRFSHGIRDVAAPLNTVTAQGATAGLVTSHMVKLRNNQ FGQSHEEPFPTLTAGGGHAGEVRAFLVKYYSEGGQDAS CADPMHTIPTKDRMGLVMVHGEPYAIVDIGLRMLTPRELY RAQGFPESYIIDRGAAGEAITKTAQVRMCGNSVCPPLSRA IVAANYSEAVQLRKVA | | | | |

| | | | | | |
|---|---|---|---|---|---|
| **Contig_8** | MTQNLPSTAPQRPFSQACENNQAPIFEVLKSAFQHSRHV LEIGSGTGQHSVYFAPRLPQLVWQTSDLADSHAGIQAWH AAHGAPNLLAPLEFDLATDAWPATKDPSGFDAVFTSNTC HIVAWPLVQRMFDLVGSHLPEGGVFAIYGPFNYGGHFTS DSNRAFDAWLRQRDARSGLRDFEAIVALAAEHGMELRLD QAMPANNRTLVFRKR | 535039-535668 | 100 | 98 | Methlyase |
| **Contig_8** | MKFVTAKTPLAAAVGLMVATSVWAKVPANEAEQLGKELT CVGAIKAGNKEGTIPEFTGKWVGAPAGVAHVQSSGKHPV DIYADEKPLFVITAENMEKYGDKLSAGQKAMFQKYSKTM QMPVYKGHRDFRYTDEVCAVLKKNALESEVIDNGMGIKG SFGAINFPIPKTGQEVIWNNLLPTRAYTEAITRDMANVLSD GSMSFGRMQNLNLDMVNKPEMLGKPVEGVMAYTRTRTL APEREKGGVTHSVEPVNFGKDKRLAWSYDPGTRRVRQV PEYGFDQPMAGTGGKMTIDSDRLFNGSPERYNWKLLGK KEMYIPANNYKIHQPTVKYADLLKPGHANPDFMRYELRRV WAVEGTLKDGFRHVYGKRVLFVDEDTGQAVAADMYDAR GQLWQHAFINYYYSFDIKAWHAGTSFYHDLNSGGYMGY NLFQERPQGPILNKGDLVPAMFTPEAARNAGN | 584030-582654 | 97 | 65 | Endonuclease |
| **Contig_8** | MSFAFSAPLSSFSSRVPLWLRHAGIAMLAIGSVACSARMP APSAAVGPASPIQVKNPTRSTQGFEQCPQFFANGKPPVL SDQPKLRALCYDAFAVLHSGQSKTPVFVAQRLNKALVAD ADEKRTNKFYSDARLPRDERAELDDYKRSGYSRGHMAP AGDMPSAQAMAQSFSLANMVPQSIKQNGGPWARIEKDT RSYAQRAQGDVYIITGPVFEAGAASVGHNQVRVPSFLYKL VYDAQSQRAWAHWQANDDAARVTEPISYAELVRRTGIEF LPGAALQTSGKLQQASMLPAQRH | 763290-764177 | 91 | 100 | Endonuclease |
| **contig_8:** | MTETSAYPNYRASGLLWAPRLPDGWQVLRNGRLFSHRV DTGFPNLPILEVSLRTGVRVRDMENLKRKQVMSQKEKYK RAAKGDIAYNMMRMWQGAVGPAPVDGLVSPAYVVVKPY DEANSSYYSYLFRTAAYMHEVNKFSRGIVADRNRLYWES FKQMPSLVPPRPEQDQIVAYLRAQDAHIARFIKAKRDLIKL LTEQKLRIIDHAVTRGLDASVALKPSGIEWLGEVPEHWEV ALIKHVADVRFSGVDKHSHDHETPVRLCNYTDVYKNDRIT DDMDLMRATATAAEIARLTLKAGDVILTKDSETPDDIGVPA WVPEDLPGVVCAYHLGLLRPVPDRVLGEFLFRAIGSARTA QQFHVLATGVTRFALGKHDVKNAVVALPPVEEQQSICRWI | 389030 to 390442 | 100 | 94 | Type I S-subunit |

| | | | | | |
|---|---|---|---|---|---|
| | TNECQPLDDAITRTEEEIKLIREYRDRLIADVVTGQLDVRG<br>WQPGSEDVVDDAALAALGDDPDDVTEEEDGDGED | | | | |
| **contig_8** | MQKKQQQDQSQIKWISDFIWNIADDRLRDVYVRGKYRDVI<br>LPFTVLRRLDAVLEASKDAVLERKKFLDTHKVAEQDGALR<br>MAAGQAFYNVSEFTLTKLKASAAGQRLRDDFIAYLDGFSP<br>NVQEILTKFNFRNQIQKLVDSHVLGYLIDDFLDPEVNLAPL<br>PVKDADGRIKLPALDNHGMGTVFEELIRRFNEDNNEEAGE<br>HFTPRDVVQLMAKLLFLPVADRIESSTYSLYDGSCGTGG<br>MLTVAEEALHELSEQHGKEVSIHLFGQEISDETYAICKADL<br>LLKGEGAEAENIVGGADKSTLSADQFRSREFDFMISNPPY<br>GKSWKTDLERMGGKKEFNDPRFIVSHAGNNEFKLLTRSS<br>DGQLMFQVNKLQKMKDNTPLGSRIALVHNGSALFTGDAG<br>QGESNIRRWVLENDWLEAIIALPLNIFYNTGIATYIWVLANK<br>KAEARRGKVQLIDASQWFQPLRRNLGKKNCELSDGDIQRI<br>LDLYLGEAQETAQSKWFDTQDFGYWKITVERPLRLKSQL<br>SDERIEPLRFASGDEALRAEIYATHGEALYTEFAKRKPAIE<br>AWLKGEDENEDDDSEDSDSGDDGEAPATRKAVPAKRRK<br>KLLDASTWQRDKGLMEVAQRVQKALGSAVFDDHNEFRA<br>RFDAALKAQGDKLGAPEKKAIYKAVSWRDEAAPPVIAKRS<br>KLKAGEHFEPGFDGAYLETVGKDRFMVEYEPDSELRDTE<br>QVPLKEPGGIDAFFAREVLPHAPDAWIATDKTQIGYEISFA<br>RYFYTPVPLRTLAEIRADILTLEQQSEGLLHKIVGGAQ | 386646 to 389030 | 100 | 97 | Type I M-subunit |

| | | | | | |
|---|---|---|---|---|---|
| **contig_8** | MAMAKTDTSERWFEARVVRGLTGVPQPEYSHALAPTDF<br>AATHNGYVQGKPTDYNRDVALDVAQLLAFLQATQPKAVE<br>TLELAADGIKRTQFLHRLQGEITKRGVVDVLRKGVSHGPV<br>HVDLYKLLPTPGNAAAADAFGKNIFSVTRQVRYSNDSGN<br>ELDLAIFINGLPVLTFELKNSLTKQTLADAIVQYQTTRSPQE<br>LLFQLGRCVAHIAVDDAEAAFCTEIKGKASWFLPFNQGW<br>NSGAGNPPNPDGLKTDYLWKQVLTRESLANVIESYAQVV<br>EEEEADASGKKRKKRKQIFPRFHQLRTVRALLRRAHTDG<br>VGKRYLIQHSAGSGKSNTIAWLAHQLVELRRKDDPMTAQ<br>FDSIIVITDRRALDTQIARTIKGYDHVAAIFGHSDNAQELRE<br>YLRRGKKIIVTTVQKFPFILDELGDLSGKSFALLIDEAHSSQ<br>GGKTTARMHEALGGKAAEEEFEEDSTQDAVNAEIEKRIAS<br>RKLLANASYFAFTATPKNKTLELFGEKTLVGDKVQFRSPE<br>ELTYTTKQAIQEKFILDVVENYTTYDSFYQVAKTVADDPEF<br>DKVKALKKIRHYVESHDKAIRRKAEIMVDHFIAQVAGKQKI<br>GGKARAMIVCNGIARAIDYWREVSDYLTQIKSPYKAIVAYS<br>GDFEIGGQKKTEADLNGFPSKDIPANLKQDPYRFLIVANKF<br>VTGFDEPLLHTMYVDKPLAGVLAVQTLSRLNRAHPQKHD<br>TFVLDFADNAEAVKAAFQDYYRATIQTGETDANKLHDLKA<br>ELDGQQVYSWQQVEDLVALYLSGADRDKLDPILDACVAE<br>YTDKLGEDDQVKFKGKAKAFVRSYGFLAAILSYGHPTWE<br>KLSIFLNFLIPKLPAPKEEDLSKGVLETIDMDSYRVEAKAAL<br>KMAMDDADATVEPVPPGGGGGKGEADIDRLSAIIKTFNDL<br>FGNIEWKDEDKIRKVIAEEIPARVAKDKAYRNAQANSDKQ<br>NAKLEHDKALNRVVLELLSDHTELFKQFSDNPNFKRWLTD<br>TVFDATYQQGAVPPKTPPQTWASA | 390426 to 393500 | 100 | 91(87) | Type I R-Subunit, EcoRI |
| **Contig_143** | MTNLLPIEHHGQVLWLLSDKAIYWPARHALLVADLHIGKA<br>ASYRALHQPVPRGTTEATLARLDALLARHDCEQLIILGDFL<br>HARAAQAPATLATLQAWRERHRTLKIVLIRGNHDRNAGD<br>PPASLGIEVVSEPWLLGPFALQHEPRPHPTQPVLAGHVH<br>PVFVLRGKARQRLRLPCFLIDGQVSLLPAFGEFTGGWEIA<br>PTGASRVFLTGADRVWPL | 9652 to 10302 | 100 | 100 | DEAD/DEAX BOX Helicase |

| | | | | | |
|---|---|---|---|---|---|
| **Contig_143** | MYWPACWFWRPWPVVCNWHCVPGCCHWHSPSATIPH CCACNAWPLTTCCPIKGAGACSPATFARCLPHACCLVPA ANGWPTVSACTWPAWPRAPCYCSDCCWPCTIACVRALP NHVIKESSDFMNAAKRLEIFRRLHEDNPDPKTELAYTTPFE LLVAVTLSAQSTDVGVNKATARLFPVANTPEAIYALGVEGL SEYIKTIGLYNSKAKNVIEACRLLIEHHDSQVPQTREALEAL PGVGRKTANVVLNTAFRQPTMAVDTHIFRVSNRTGIAPGK TVLEVEKKLIKFVPKDYLLDAHHWLILHGRYVCQARKPRC GSCRIEDLCEYKHKTSDD | 17656 to 18660 | (85) 100 | (99)87 | Endonuclease III(IV) |
| **Contig_147** | MTSTQHTDVVQRQFGEQASAYLSSAVHAQGSEFALLQAA LAGQGHARVLDLGCGAGHVSFHVAPLVAEVVAYDLSQA MLDVVASAAAERGLANVTTERGAAERLPFADASFDFVFS RYSAHHWSDLGLALREVRRVLKPGGVAAFIDVMSPGSPL LDTYLQTVEVLRDTSHVRDYSAAEWQRQVSEAGLHVRSH TRQPLRLEFSSWVERMRTPEPMRLAIRQLQQAMGEEVR QYYQIEADGSFSTDVLVLWAER | 6705 to 7466 | 100 | 99 | SAM-dependent methyltransferase |
| **Contig_175** | MHTKITAEQIHKVYAHFKQHQEHPGKWPAENPVELILGAIL VQNTTWTNVQKTLDNLRPITGFDAEKILALSTEELQELIRP SGFFKNKSKAIQSSLQWFQDQNWDFDAIATRYGRHLRNE LLNLHGVGQETADVYLVFIFHQVHFVADAYARRLFGYLTG TEFKTYQDLRKVVEIPDDFTTQDAEDLHGYIDDYGKLHKN PEDFDQSFFGGFDL | 21539 to 20895 | 81 | 47 | Endonuclease III |
| **Contig_208** | MLRREQLIGANFSFQHHPFQWVAGQLRQMGFKRMELW GIAPHLDLFHGSPARLAELRSILGDNGISVHCFTPEQVLYP VNIASGDRIYREKSLDCFLRAADISAELGASYLFLTPGRGF ESESRHLAWAHTVESLSKMATHAAGLGIRCLLEPLQRLES NIANNAADLERLWRDLNANNVDLVLDLVAMAAAGDQVGG YMKRFGKRLAHIHIVDGTPSGHLVWGDGNLPLDDYLAEIS DHSFEGTLTFEPFGNGSYALDPAAAWRRCLDAIAPHFDT AE | 2844 to 2008 | 100 | 87 | AP Endonuclease |

| | | | | | |
|---|---|---|---|---|---|
| **Contig_232** | MTIVAASGISTSAAAGLCSSPRPRNRASRSKAGSRIFVAR CPRRPPASSPRCSPCRIFPSVSSLTTLRKATSACAPIRTGI PNRPRSSLQSTDLHKASCLTAGRFFHARRAAVAGGESER RAGLAVTGLKAGERLPPPVPRSFPHEPRPPDHLHVDRRR SRAAGHAGPVSRRTNAAARDLHRRCRLRRLQRSRQLRP YGRHRRSPDAGGRNVPRQPACLALRYSH | 108 to 782 | 9 | 82 | Probably Mtase/Helicase |
| **Contig_251** | MSSFEESLKIIADRVKTHSSTMATEEAVKTAVVLPFLRALG YDVFDPNEVIPEFTADAVGKKGEKVDYAIKIDEEIRILIECK PITFTLEKKHLDQLYRYFSVTNAKFAILTNGRTFNFYTDLD APNKLDARPFFVFDIADFNPAIAVELKKFEKASFNVDSILAT AERLKYASGIKKAINSLIEEPTEDFVRIVAADVYDGRLTAPV KE | 2372 to 3001 | 100(100) | 82(98) | RE (Methylase) |
| **Contig_260** | MQNAAGGKGERGDVSPSQQGRAGLRLQHALPNARIVYV SATGATTVHNLAYAQQLGLWGGEDFPFSTRAEFVEAIEA GGVAAMEVLARDLRALGLYTARSLSFAGVEYELVEHELTP EQRRIYNAYAGAFTVIHNNLDAAMEASNITGSSGTLNRQA KSAARSAFESAKQRFFGHLLCSMKAPTLLRSIAADLEAGH SAVIQIVSTGEALMERRLAELPTEEWNDVRVDITPREYVLD YLAHSFPVQLYEPFTDSEGNLSSRAVTRDGQPVESREAV ARRDALIAKLASLPPVPGALDQLIQHFGTDTVAEVTGRSR RIVRKAGNGVTVDRLVVETRAASANLAETQAFMDDAKRA LVFSDAGGTGRSYHAELSAKNTRLRVHYLLEAGWKADTAI QGLGRTHRTNQKQPPLFRPISTNVKAEKRFLSTIARRLDT LGAITRGQRQTGGQGLFRPEDNLESHYARDALRQLYLLL VRGKVEGCSLERFEAATGLKLMDSTGIKDELPPITTFLNRL LALTIELQGVLFTAFEQLLAAK | 1619 to 6 | 100 | 93 | Methylase |
| **Contig_311** | MVPGALLGKARHLVLTNHYPRYDDLYRNGFVHSRVKHYF KHGVPVDVFRFQPNAALGYYEFEDIDVTTGGAEALDKLIG DRQLESIAVHFLDPAMWQVLEHHIDKIRVNVWIHGAEIQP WHRREFNFRNDMERQAERAKSEQRLTFWRKLLAKQHDN LHIIFVSKNFSEEVMEDLGFRLAEWRYSIIHNPIDTELFAYR KKDISHRQRILSIRTYASKKYANDLSVAAILELQKEPFFSAL EFRLIGDGQLFDEETLPLQGFRNVFLEKRFVSHTEIVALQE NYGVFLTPTRWDSQGVSRDEAMSSGLVPVTTGISAIPEFL DETCGYLAKPEDAAGLAGAIADLYRHPQKFLDMSRTAAK RVRNQSSANEIIRRELALICRLDAEAFT | 2016 to 850 | 97 | 57 | MTase |

| | | | | | |
|---|---|---|---|---|---|
| **Contig_399** | MRSDAAPSAAIISSRHSHPQVGSQPCIIGCGRLYVRRCIS GRLGHLQIEEQIMSRLDSFIRRLCAQRDILNSIADDVREIPG AIMELGLGNGRTYDHIRELFPDRRIIAFDRVARSFESSTPQ GDNLVVGEIKDTIGDYLGIEAALVHADIGTGHEDLDALTLT WVPDAVASLLKRGGIAACGLPLEHPELTPMPLPQSVAAN RYFLYRRR | 1500 to 2132 | 74 | 66 | SAM methyltransferase |
| **Contig_608** | MDRHMPHFQHVAGAHRVPWVHNTIMTEGDINASSHQFR HAGHAATLWIGVGASLQCDIDERVRNSVHLRFRNQRNEL GNIVIVHGMHGCEVRARYATLKPETHGLGGQRFHMTRH WIIGFVAMNIDRKAALGGNPAEFFERCSAIGHGALEMRDA TDNVEAHIERAVDEIDRARRAVIAILRKGDELQIYIRLYLFA HLDHRFRRQQTRIADIDMAANGQQALADGQIAIAQGALDH GFNGQHGLEFTPERDAFQKRAGAVEARQAKRQRCVHME MRIDKRRRYEAVGSVDFLGSLMLALRWNDIGNLAVLDQN VLLLTTVREISVSHNQIEHRQAS | 293 to 1303 | 20 | 35 | SAM-dependent methyltransferase |
| **Contig_615** | MTVLTETVLTADCRDLMPARGPFDLILADPPYGDTSLAWD RRVADWITLARVALTPTGSLWVFGSLRHFMATADQFADA GLQIAQEIVWEKQNGSSFHADRFKRVHELAVQFYPAETA WRDIYNDVQTTPDATARTVRRKMRPPHTGHIDAGHYVSH DGGPRLMRSVIYLRNCHGRAIHPTEKPSALMEILIRTSCPE GGLVGDWFAGSGAAGEACRLTGRRYLGCKIRPDMAEKA RARIASVLPFDGRAAP | 2258 to 1503 | 95 | 89 | Site Specific DNA MTase |
| **Contig_632** | MSMISASAAASATAPLPLALETDTAARVFAAAGLLLPHVE RGQRIDAATLRGAMEAAFGASDAAGAWDWKTAYEACEA ATVLFLRKYGKTLIRKAGSAAHGLPMLTKIAGLLPTHTRRS EEAQTFQQFSTPIPLGLVAATAAAITPTDRVLEPSAGTGLL AILAEIAGGGLVLNELADMRASLLSSLFPALSVTRLDAAQID DHLDPALVPSVVLMNPPFSALANVDTRMADAVYRHVASA LARLAPGGRLVAITGANFASDNPAWTDAFTRMQERGRVV LSAAIDGSVYAKHGTTIATRLTIIDKLPADDPAIFPAAPGVA PDVATLMGWVTEHVPARLPVELPGLPAATPTSAAPRTVR GYVNRNTRLASPITPAEPEAEPPD | 1160 to 6 | 99 | 84 | Methylase |

| | | | | | |
|---|---|---|---|---|---|
| **Contig_659** | MEKANYQFSTPVLANASERVLGPWTPQQMILDGLVALRK DDRATGRAYTIRLRRSSKPSRANAARTLAAYDAAPADSIV FIQLVDDLGGAVVGDIIAHRLKKLNVAGVVVEGPVRDIDGL IQFGPPIWYRNAVASGLELAETEVEVQVALQLGAALVEPG SIISIDRDGVFVFPGTHYEALVSQADTITAKEALVHAALENN QSLLKVLNLDDEITP | 2147 to 1497 | 72 | 30 | MTase |
| **Contig_688** | MIVTLLNQKGGVGKTTLALHLAGELAMRGSRVTLVDADP QGSALDWSQQRSRESLPRLFGVVGLARDTLHREAPELAR DVDHVVIDGPPRVAGLMRSALLAADLVLIPVQPSPLDGWA SAEMLALLAEARIYRPQLVARFVLNRCGARTVIARETAETL ADHDPPLLAAMVGQRVVFADAAQTGRLASDIDRQSPAAR EIAALAAEVGRLCIGRTAP | 919 to 164 | 100 | 97 | Site Specific DNA MTase |
| **Contig_715** | MNPPFSAMANVEGRMADAAYRHVASALARLAPGGRLVTI TGANFAPDNVAWTPAFTRLQERGRVVFSAAIDGSVYAKH GTTIPTRITVIDKLPADDPALFPAAPGVAPDVATLMGWLDT HLPVRLPVDPSVAVPVVTAPAPRTVRGYLNRAAAARPAP SAPLAEPEAVALEYETVDWQAAEDGRISDAIYEEYGLQVI RIPGSTAHPTKLVQSVAMASVAPPKPSYRPHLPTN | 441 to 1139 | 100 | 74 | Methylase |
| **Contig_774** | MRISDWSSGRVLFRSDLPNLDLVIHADLPNNPETMLHRSG RTGRAGRKGTCVLVVPFSRRRSAERLLHMAKLDAQTIPA PGIAAVQAKNNERILNAEAFGQPVEEEHQDVLKALLERYT PEQLAAAYLNRELSLAPAPEEVSDAPVHPVGGKKPRERS DRGDRFERNDRFDRGERGEPGERFDRNAHFDGVWFSV SAGRKHRADPKWLLPLICKAGDVSKRDVGSIKILDGETRF EIAASKADEFRQSVAERGTGEKGLVIRPAVAGAGEDAGR RESKSFKPRGEKSYDEKSWGDKPRGDKPWGDKPRGEK SWGDKPRGEKSWGDKPRGDGFKKGPKKSEGGYSKKPR SE | 23 to 1072 | 77 | 86 | DEAD/DEAX BOX Helicase |
| **Contig_778** | MKTIKGPGIFLGQFAGDAAPFNTWDGITKWAAEKGYLGV QVPTWASQLIDLKKASESKDYCDEFAGVAKENGVVVTEL STHLQGQLVAVHPAYDEAFDGFAVPEVRGNPKARQAWA VEQVKMAIKASRNLGIGAHATFSGALAWPFIYPWPQRPA GLVETAFDELARRWKPILDHAEEHGVDICYEIHPGEDLHD GVTFEMFLERVKNHPRANMLYDPSHYVLQCLDYLDNIDIY KDRIRMFHVKDAEFNPTGRQGVYGGYQGWVNRAGRFRS LGDGQVDFGAVFSKMAANNFDGWAVVEWECALKHPED | 1082 to 2137 | 99 | 97 | AP Endonuclease |

| | | | | | |
|---|---|---|---|---|---|
| | GAREGAEFVKAHIIRVTDKAFDDFASGGTDDAANRRMLG LCD | | | | |
| **Contig _837** | MHPDIIELRSFYDTTLGHLAERAIRMAIAGLWERVPGERLI GMGYSLPYLDRFSADTERTFAFMPAGQGAIAWPSAEKST TALVFDEELPLPDSSIDRILMVHALEYAESAPETLKEMWRV LAPNGRLVIVVPNRRGVWARFEHTPFGSGRPYSRDQLST LLREANFTVNTVSDALHFPPATRRWMMRPCLAIEGLGRR LWPLFSGVLVVEAQKRLYQGLPVAQRASRRVFVPVLGT | 1188 to 478 | 99 | 92 | SAM-dependent methyltransferase |
| **Contig _972** | MQYSCAYFERADMSLDEAQQAKKRHIAAKLLPGRHARVL DIGCGWGGLALYLAEAFDAHVTGITLSSEQLEFARGRAAA ASDPARLEFRSQDYRDVNERFDRIVSVGMFEHVGVNHYD EFFRKCRDLLDDEGVFVLHSIGRSDPPGFTNPWIARYIFP GGYIPALSEVLPAIERAGLVVTDVEILRLHYAETLRHWRER FLARRTEAVDLFDERFARLWEYYLGVSELAFRYQGMMVF QIQIARRQDAVPLTRGYIEREERRLQAVSTTQPECESAQE HSCVAVG | 399 to 1253 | 97 | 66 | SAM-dependent methyltransferase |
| **Contig _1013** | MTIQTPDISRSKSMTQERTGCRLCGKRLKHTFVDLGMSP PCESFVAADQLDRMEPYYPLYALVCDHCYLVQLKEYFSP AEIFTEYAYFSSFATSWVDHARAYCDAITERLELNENSFVV EVASNDGYLLQHFLPKGIPVLGIEPAANVAEAAIAKGVPTR VDFFGAKLALEMLGANQSADLIIGNNVLAQVPDLNDFVRG MQLLLKPEGVITLEFPHLANLIEFNQFDTIYHEHFSYFSLLTI RFMAHRHHLKIIDVEEVPTHGGSLRVYLARTGSKRKASPR VAALLKKEESFGLTDIAT | 357 to 1259 | 100 | 88 | SAM-dependent methyltransferase |
| **Contig _1082** | MRNYRRPYRRPNRRNNSGTRKSGMGGLRSIILTLLFFSAI VIGISYLPNEKRPTETRDGSVYVIDGDTVIINKVHIRLKGIDA PEMTQSCERNGNSYDCGKEARNFLRARIGRATIRCETEG FDRYGRDLARCYLGETDLNGWMVQQGWALAYGDYDRE ETDARRNSRGMWAGRFEKPSSWRKENPREDKDAQTGS KTSSLDRNNIDAFVYYIKERIAAFIDRFQ | 199 to 873 | 73 | 45 | Endonuclease YncB |

| | | | | | |
|---|---|---|---|---|---|
| **Contig_1173** | MAAETGAARTIRTIVETIQPQLNVKLWDGTQIGAFDGPTLA IKDPSVVRQVVLKPNYDTLISLWTSGAVDIENGTIFDLAET KIDGHLKERIKALPKWQLLKGIPSLLFAGKAKDQANIDGKS PFVSGSNKEAITHHYDVSNEFYQLFLDERMVYTCAYFTD WQNSLDQAQYDKLDLICRKLRLKPGDRFLDIGCGWGALLI HAVQNYGVIGTGVSLSEAQTALARQRIKAAGLEDKITIHIKS YTELDQEFDKISSIGMFEHVGIANYDTYFKSVRRLLRPGGL YMHHAITRRMKKNKKSFNRKSAEHLALVKYIFPGGELDHL GMTVENLEGHGFEVHDVENLREHYGRTCRLWAERLHAN FDKAVAEVGYPKARLWILYLAGCALAFERGTVQINQTLAS KRKRGISAVPQTRADIYQS | 1273 to 8 | 99 | 81 | SAM-dependent methyltransferase |
| **Contig_1200** | MRTHQSEYWDQFYTNVDTSSPIYPSQFAAFALGEASFAT KVIEFGCGNGRDAEFFASQGKDVLAFDASEVAINLCQSRN CRDNILFRKFAIGDPLEEGLFGSTETKTLLYARFFLHAITDE QERQFVSLASDILKEGSVLALEYRCSGDEENNKIFGNHYR RYIRHEDLCQYISENGFEILYEIIGKGLAKYKSEDALVGRCI AAKR | 1335 to 715 | 96 | 49 | SAM-dependent methyltransferase |
| **Contig_1236** | MLAKQLRRDLRFDPEHQFLALLLGLHCLGRELGDIGHETH ARRHDKLRRCVEYETNIGTDCDTSSLSRREKKCHVNVGE VDEIEYPPTGGQDLAGLRDAILHTPVAWRSEGAVIDIGNN TFNRRICGNDGSLCIDDLGLRSADRRIGSGKRSPGCSSSR PRPLCRRPVVVERLLGGHVGLDQLLGPGKFSFRGILFSFA LSDHRGSRFFFRLPLGEQAFRCIHTDHGALTRGFGLPPLC LQLLGVHASQHLSGGDETPFVDEDFLDPPGRLAETSISVA SMRPLPPTIPAGRPVPSYIRQP | 605 to 1507 | 16 | 44 | RE |
| **Contig_1251** | MSDAPEHFGGTFLVGDEHRWPQKFHRAQKRFDLFLAAG CQDALRRVLQAFDAGRHDALMHRIIDDTAIGIVEMEQNVA DRTIGAQRLFLLFIDEPHAVGRATGGLDAANAEQHHVEIF RHVQHRVQATALNQGCASLLAVHKGGRGDKVAFALQHL RELPGKFGRVGASWLHNAAEKRNDLEIVHRTRANMLQGA GQSAAPADANAHAVLDFFRQISACRRADRRNRIGLGRGE GRDNGLLQPESLFKRRFQRGQSRKLAFDGDLHLVAGAGI GKKTHNGHTANAQRIRNIALGHLLDIIHPRRTGSEAAGAVF RCNAFGSHQKVVGFSCVPFPLGGLSQIDSPKFLQFASL | 3 to 1058 | 40 | 29 | SAM-dependent methyltransferase |

| | | | | | |
|---|---|---|---|---|---|
| **Contig_1260** | MPPISTTVSDPTSSSVAFFSITACFFRKDMCVRSLFRRLH RWCRLGRGCAGNRRFIAQLSANGLIDVPGHDETAREVEQ AASHADDVVREHRGDRFCEGIDEETLVVVFAPHQALLDT RNPHGGGVEHDAENSEPEVPVDHAHRIEAFAVPQLRSQII DCAEGDHAVPAKRAGMDVPDGPVGIVRQRIDRLDRHQR TFEGRHAVEGDGRHHHADDRIGADLVPCARKRHQAVDH AAPGRHPQHDREHHAQRLRPVGQGGIVQVVRTSPDIEED QRPEVDDRQAVGIDWTVRLLRHEIIHHAEER | 916 to 5 | 39 | 67 | SAM-dependent methyltransferase |
| **Contig_1355** | MRRFRVVLAEHAHDLGQFVHQLGLVLQAPGCIDDQHVAA RVPGFFPGVIGEARRISAQLGRDDRCAGALAPDMQLFDG CGAERITGRQHDFQATCRQLCRELADGRGLAGAVDADH QNDMRLMRKVEFQRPGDGTQHFLDFRAHDRAHFIARNIL AVAPGGQRIGDAHRGLKAEIGLDQHVFQILKRVLIQLALGE NAADGFAER | 1011 to 1625 | 63 | 34 | Putative MTase |
| **Contig_1365** | MDDFTYASYEHEAVLMQLTPDEWREQLDTFQPEILFVES AWRGKDGLWGNQVGHRSNQLVELVEWCRQQHIPTVFW NKEDPVHYETFLNTAVLFDHVFTTDIDCIGRYKAALGHDN VWLLPFACPTARFNPIETYAREQAFSFAGAYYARYPERN GDLASVISALAENHKVAIYDRNFGQDDPRYQFPPEFQPFI VGYLPYDQIERAYKAYTHAINMNSIKQSQSMFARRVFELL GSNTITISNFSRGVRNLFGELVVCSDNGAEIIRRLDKISDGT TERKFRLAGLRKVMSEHTMQDRLAYLASKALKTQQAQLL PDILVVAYISDKEEF | 993 to 1 | 100 | 62 | MTase Type12 |
| **Contig_1405** | MELPPILRQAVDAALEGVALADLKRASDLLSRRYRAETRD GRLHISDDLAAKAYLAARLPATYAAVRASLDSAAEVRPDF APQSMLDVGAGPGTALWAARGCWPSLENATMIEASPAIR AVGSSLAGNSGFASLDWRAGDVVKEKLDFPQADLVTIAY VLDELAPDDRRKLVAQLWASTRQMFVIVEPGTPAGWQRI LDARRALIELGAVIAAPCPHQLECPLVAPDWCHFSRRVAR SRIHRMTKDAEVPWEDEKFIYLAAVRQPSEAVEARVIAPP RVGGGKVSLKLCKGDGTAEERLFTKRDGDFFRWARRAD WGDAYLEEPAR | 2380 to 1403 | 99 | 84 | MTase Type11 |

| | | | | | |
|---|---|---|---|---|---|
| **Contig_1429** | MCCCSACFIVHRRHLWRCFRLGPLNRFGLTHERVLQNFT DARHRNDLKIVLHIVRNIRQILGVFFRDQNLLDAAAQSREE LFLQATDRQNPTAQRHFTRHRDVLANRYAGQHGNDGRD HGDTCGRTVFRRRTFRHVDVNVALVEQWRLDAEINCSAA DIGCSSGNRFLHHVTQVTRDGHATLAGHHDAFDGQQFPA NFRPGQAGNDTDLIFAVDLTKAETLHTQIVGKILVGDLHRL LLRLQDFGDSLTGERHHLTLKATHTGFPGIKADHITQCIVR ERELLRLQSMVLDSLRQQVPLGDLQLFILGVTGDTNDLHTI QQRTGDVQRIGGRHEHDVGQVILNLKVMVHEGCVLFRIQ HLKHCRRRITTEILAHLVDFIEQEKRVGLLCLLHRLDDLAG HRADIGAAVTADFSFVPHAAKRHTDIFTARRLGDRACKRG LAHAGRSDEANDRALDLGRTSLHSQILDDAFLDL | 284 to 1702 | 12 | 32 | SAM-dependent methyltransferase |
| **Contig_1485** | MVAEVTGRSRRIVRKRSSGGIDRLVVESRAASANLSETQA FMDDEKRALVFSDAGGTGRSYHAELSAKNTRLRVHYLLE PGWKADKAIQGLGRTHRTNQAQPPLFRPIATNVKAEKRFL STIARRLDTLGAITRGQRQTGGQGLFRPEDNLESVYARDA LRQLYLLLVRGKVEGCSLDRFESATGLKLMDSTGIKDDLP PITTFLNRLLALTIELQGILFTAFEQLLAAKIEGAVASGTYDV GLETLTAESFVVTDRKTIYVHPGTGAETRLLTITERKRNRP VTLTEALGHLDDPRAMLLINERSGRAAVQIPTTSVMLDDG EIERRVRLIRPMEGHNIPVKLMAETYWLEADHDAFAAAWD AELAEVPEFTDGTIHVVAGLLLPIWKRLPQESTRVYRLQTD EGERIIGRRVSPAWAATAT | 45 to 1313 | 100 | 88 | Methylase |
| **Contig_1561** | MNLIKSKQRVNDHGEVFTPPWLVEKMLDLVKGETERIDS RCLEPACGSGNFLVRVLQRKMAAVEMKYGKSEFEKRHY ALYGLMCTYGVELLDDNIAECCANMLEVLAEYLNIDEGDD CYKAASYVLSQNLVHGDAMTMKDQSGQPIVFAEWGYMQ KGKFHQLDFRFDVLTGSSAYNTEGTLFAHLGAHEIFTPVK VYTPMTMGDLAEQLRG | 67 to 699 | 100 | 91 | RE M-subunit |
| **Contig_1796** | MMDTLGANATTTDWHESAFDILRGVAAYTSSPLIKGLSKV LADHPEADLGNAFNHKQVGCKIWARQSLFETFGGRFNRI AILGGWYGVLAAMFFEDQRFDIEAIDSFDIDPDVGAVAETL NNAWKDRFRALTADMYQLAYPELGADLVINTSCEHIADLP AWLSLLPKGTRVLLQSNDYFSEPTHVNCVASLDEFVAQA ALETTAFAGALPMKKYTRFMLIGTV | 306 to 977 | 100 | 68 | SAM-dependent methyltransferase |

| | | | | | |
|---|---|---|---|---|---|
| **Contig_82** | MSPLPLNRILNGDCGQVMRALPTGSIDLIITDPPYLIRYLDR NGRRVANDDNGRWLQPAFTEMYRLLKPGALCVSFCGWN EIAQFAEAWQQAGFQIVGHMVFYKKYASSVRYLRHHHEQ AYLLAKGKAPRPLRPLPDVMAWNYTGNRLHPTQKPVMPL QALVKAFSQPDDVVLDPFCGSGSTLVAAKAAGRRFIGIEL DERHCFTASMRVQTLSV | 17288 to 16644 | 99 | 70 | DNA MTase |
| **Contig_82** | MTLTMKESSAIAALAQLLYDFLPGSGNNQTAFPLAAHKAG VGEYWQPGSKLPSLTQLLTLTLEWKRGQFCPLILEIVRQS MTWRGRRDPLKREEVDQLNKLLPGVGFRIPELLDPDFLD TLAGPPAVQPAPVGSGKPVIDTNRLAELSKRLSDLSAISP QERGFAFERFLYDLFDVYGLAPRASFRPRTGEQIDGSFDL DGDTYLLEAKWHSNPTPAADLHVLSSKLNSRPIWSRALFI SYSGFSPDGLEAFNRGKNSLICMDGYDLYETISRELSLGQ VIATKARRAVETGLCHVSVRDLF | 28030 to 27125 | 100 | 78 | RE |
| **Contig_83** | MPGGTRVGVALGTTGRPASAVAACPPHSAGRGQPGPQA QGPGRPCQPTRTGRPAPARTAAELAGLSAATFRTGVPVK GVVMSLDPQYFADLYATSEDPWAFRTRWYEKRKRELVM ACLPRQCYQRVFEPACANGELSALLAERCANLVCQDLDP TAVALAGERLAGLRNVSVELARLPADWPGGRFDLIVLSEV GYYLDPTDWLQVIEQSVASLTYHGGLLACHWKHPIAGCP QDGREVHRMLARHLPLYLQFQHDEADFLLEYWSSQPSV VDLDETCP | 16226 to 17062 | 71 | 99 | MTase |
| **Contig_83** | MLLDPQQIQADQAMLQLGRRLRADGYRFTCVTPATHARV NARPEAGQARTVRDVFGWSRPFRSSLVSADELDQMRRA QVLKQHGELLISTVRWSTIDELLLLHSAYPTEETDAVFLGP DSYRFAQVIHDHLQRPPKRVEHAVDIGCGTGVGALLIARA APHAQVSAVDINPLALRYTAINAALAGLSNVSVEPSDVLD GITGLFDLIVANPPYMLDACQRTYRHGGGSLGAQLSLRIV EQACERLGSGGSLLLYTGVAISEGRDALLEAIRLRLAGPE WSWVYREIDPDVFGEQLNEPGYEQVERIAAVVLTVTRNS | 22583 to 21633 | 100 | 99 | SAM-dependant Mtase |
| **Contig_85** | MTDQAFAQADPDWVKLISLAREWFNGPLGQLMLREEEKL LEEELGRFFGGYLVHYGPCAEPPPSAPQVQRSVRLGAPL PGVEIVCEEQAWPLSEHAADVVVLQHGLDFCLSPHGLLR EAASAVRPGGHLLIVGINPWSSWGMRHFFSHGALRKARC ISPSRVGDWLNLLGFALEKRRFGCYRPPLASPAWQQRLA GWERVAGGWQSSGGGVYLLVARKMVVGLRPLRQERRE | 63997 to 64767 | 100 | 99 | SAM-dependant Mtase |

| | | | | | |
|---|---|---|---|---|---|
| | PMGKLLPLPLAKVNRTAANPDTEKH | | | | |
| **Contig_94** | MDEPLNIEISKFQKPISATKSIATYQQGDLVVVQIWQGQTR LYHDALTYEEYLTFKDSVPEYGLPNYPPPKDYKNKDIDGT TASGNTIGRVVGETAKAKIPAPSSNPPPEVETEKESLWKK ASPWVHGTLDGLGFVPGLGAIPDGISAFIYVLEGDIENAGL AAFAAIPVFGDAAKGGVLVGKAAGKVNKSLKNSHATPPKP SPPTERKPNKPKNNGGKSKGEKRNCRLRRYGNGNNCPP GKTGHHIVADRAFRLPNKNKVPGPRLPGGLSHANGYTICV DGGTPTKKGSKANEHGLIHAIYDPAERELGRRGTPQGTA TLGELELIGVMAASAITGCNPTRMLAELHAYHASMGLRSN DRYRAYNKTNLLNPGDITKLGSNTQKGSGGL | 10159 to 8990 | 48(41) | 52(37) | DNA/RNA Non-specific Endonuclease (GHH/ HNH containing Endonuclease VII |
| **Contig_94** | MCPLCRSGLVSRSGRTAAPAILALVLTNGAAARPDRDTRP LLQGSAQTRDRGVTVNPEAGSKGFASVNQAPAVKRLRVL TVNTHKGFTAFNRRFILPELREAVRSTQADIVFLQEVLGSH DRHAARYPGWPQTSQYEFLADSMWSDFAYGRNAVYPD GHHGNALLSKYPIIEHRNLDVSITGPERRGLLHCVLDVPG QHQVHAICVHLSLLESHRQKQLQLLRKLLESLPADAPVIIA GDFNDWKSHGNRTLGLQRDLHEAFERHHGHLARTYPAR LPLLRLDRVYLRNAESHGPRILGHKPWSHLSDHLPLSVEV RLSNHSS | 54633 to 53665 | 83 | 99 | EEP domain containing Protein (Endo/Exonuclease Phosphatase |
| **Contig_94** | MDRSLQLNRASWDERAPLHAASNDYEVERLVQHPEHLS ETVRFDLPLLGNIDGLNVVHLQCHIGTDTLSLARLGAKVC GLDYSAASLAEARALAQRCAAPIGYVESDVYAADKVMPA GTFDLVYTGIGALCWLPRIEPWARTVAALLKPGGRLFLRD GHPMLMAVNEDHQDRLQLEYPYFEHEEPTVWHNDQTYV ETEQRLSHTETHEWNHGLGEVISALLAHGLQLTALVEHQS IPWEALPGQMVKGDDGEYRLREQPARLPLSYTLVAVKA | 131970 to 132788 | 100 | 99 | SAM-dependant Mtase |
| **Contig_98** | MNAVERHFMQRDRRAESGRWCGFLAVLDAHEVAVVAEQ VGQAAGMGAADRVEQGMQRALAQSGLQLALPVGVVVVE HPAHTEGLERGVVVLAGTGPHAQATLRGQLREVEAHRAT GADHQHVAAGTFGYLGQGLPGGQGGAGHGGRGCVAER AGDMHDQACIEQAVLGKTAVTRQRLVVGDAAAQRHVDA CANRDYHTGAIHARDRALLPGRVAPLADFPVHRVEGNGA VGHQHLTGAGLWHGFTQQLQAVEAGVWGPGPGLMIGW HGAGPYETGVGGRLAGIGPGKQVR | 10228 to 9359 | 27 | 24 | DNA Endonuclease |

| | | | | | |
|---|---|---|---|---|---|
| **Contig_98** | MSLQDSVATFIGLSNDNEFFSAHYLAEVFQGDLADTIKEW<br>EIREEQGDGFVTPHRALRNLNQGYFALRHKLKTERSASE<br>RIRLQREFYHDLLVALGIPYQPGNREVAANMELPVLSVLG<br>DQLWVLGALDANSEGEDPLSLQLHRDQFFGSGPHHDKL<br>SNTDWYRILNEVVFRQSSPFNEQPPRWVLLLSDRQGILID<br>RYKWSQNRMLRFDWEEILGRRDDRTLKATAVLLHRESLV<br>PDDGQSRLDSLDENSHKHAFAVSDDLKYALRHAIELLGNE<br>AAAQLVEQARDRKEGIYSGSNALDPDQLSSECLRTMYRIL<br>FLFYIEARPELGYLPHQHDAWRQGYSLESLRDLESVRLTT<br>EESRRGHYFHHSLQRMFSLIYNGHQLNRQLDQLHESTAN<br>GFTLQGLDSHLFDPANTPLLNRVTFSNETLQRVIQSMSLT<br>QPQKGRKRRGRVSYTQLGINQLGAVYEALLSYRGFFATD<br>DLYEVAAKGQNINPDELETGFFVTQSQLNEFDEDSEWVY<br>DIEDKKRKLRVHPKGKFIYRMAGRDREKSASYYTPEVLTK<br>SLVKYTLKERLTSDVTADDILNLTVCEPAMGSAAFLNEAV<br>NQLAEAYLTRKQQELGQRIPHEDYQHELQRVKMHIADHN<br>VFGVDLNPIAVELAEVSLWLNALSGGHNVPWFGYQLFTG<br>NSLIGARREVYPASTLKKQAKDGLWYNHAPRRLNPFSLLE<br>QAGEGGRKEGEIYRFLLPDPGMVGYNDSVAKQLRPDAFK<br>AIKDWKKAFCAPFEAQEIRTLQTLSDAVDRLWREHTQMLE<br>LHRRRTEDSYPLWGQQGMAEHHTSTKEKDQLRTTGIFNT<br>NARIASPYRRLKLAMDYWCALWFWPLDKAEQLPDRQKW<br>LFDLNTILNSAGTFEFVPTQEGLFSAEPAEGEDLFAKPIED<br>LFAVDEPQQTLRAETQAVRDVSTQQGELNLEKLFKNPFF<br>KTLAIANELGEHFCFFHWELAFSDIYAKRGGFDITLGNPP<br>WSALNFNMERVIGDFSPELILRTMSGEEKVALEARMLKVD<br>DVVSSVIKELEEISGARSFLGSIGNYPFTSSMKCDLYKGFA<br>ELQFNILSRSGVSGLLHQNSIFEEENGGSMREKVYQHLIY<br>HFSFVNQEKLFPIGNTRSFSLNVTSCAERNICAQMIFGLYH<br>PSTIDECFTSQGVSELGKKNLNGRWNRQGSLARIIDLNEG<br>YLKFVDYFFEQSMSAESWRSARLISFRSGLESAALLKLNK<br>FSKNNLKFSAGERNFGDLECVERTGFVDQKKFVLQSAHI<br>GLANAFSQTPKRVCDTHRAYTALDISGIGPDYLPRSNFVV<br>EGYNPDLSLSKIVARCRVDKEGERTLLSGWAPDLAQQEG<br>VSYFSLGDNAENLLWATLLSSLPYDYFHRVIGKPHFRLNA<br>LRKLPVLNFSVSKRNELQVRGVCLFSTLEVYKEIWGAVWS | 71535 to 66688 | 99 | 46 | Mtase Type II |

| | | | | | |
|---|---|---|---|---|---|
| | NKFRGAKWTSNNKCLLQDFFSNLTPEWQPHNGLRSDYS<br>RRQALVEVDVLVSQALGLTLEELLTIYRVQFPVMRQYEAD<br>TWYDQTGRIVFTPSKGLVGVGLPRKARKSDLGEGTHYSV<br>ESPDFNAHDIALGWEDIQYLSQGVVRKTYQDDTLPGGPW<br>ETTITYHAPFFKPDREEDYRVAWAIFEESAGCA | | | | |
| **Contig_98** | MIPGLLASEVSAALREFIITGFETETAPFRGEFRRLVEEQQ<br>DGEAFIKGPYVTVGLPFLSGQSGCDFFSGFKTEFPPHAH<br>QEQAWRRLAANGKVANTLVATGTGSGKTECFMYPVLDF<br>CQKAGKPGIKAIVIYPMNALATDQAKRFAKEIYNQPSLKGL<br>RVGLFVGGDGQGVKAMGPDQVITDKETLRQNPPDVLLTN<br>YKMLDYLLMRPQDQKLWVNNGPDTLRYLVVDELHTFDGA<br>QGTDLSLLIRRLRARFNLAPERLICVGTSATLGGEDSVAGL<br>LQYASDIFSAPFPREAVITEQRQSPDEFIDTLSFLNLNPDIG<br>PETLQIALREGLNEYLLVAYELYFSKQPEMNLQEIPGRIAL<br>GKELKQHGQLANLLRHLRQGDKTPTFRELANRLAAQIPKR<br>FQRQPEQALIALLSLAAHARAETKLPLMQLRLQLWARELR<br>RIVGTLREPVPADDLEIEETDSIKRMPPLLAFGDDPPARDK<br>QQIRLPLVQCRECHGTAWLTRMEVAQPVNQQIELDLANIY<br>SAFFSNHQETGLLMPWQPSGDKQMSGPRLTHFRVCREC<br>GFTAGLEHSGGCKACQAGNEALVRVSRPDLLKEERVGQ<br>VNRVVHQHNCPYCSAKASLVVFGARAASLSAVAIHQLFSS<br>RDNDDRKLLTFSDSVQDAAHRAGFFAARTWQNNVRMAL<br>TQLLENQTGPVPLLQIPELFERYWLEHEGQHGHLALPYYL<br>REFMPPDKRFDGDFEFFEQSGEVRNPARHLKIIRNRMLW<br>QVLEDLGWRAQVGRSLNRLGIAALEWPLDKVQQAAKRW<br>ATEVNNTLGYRIDSEAAQQYMQGLMLHLVHLGAFALEDL<br>ASYRRNAGKSYLLNLLDYAPASGPSAARPRYPAASNGDG<br>FEALSGSQGATWYQRWLACLNPGELVDRKQLESVLGAS<br>LKALCSTGLLKEELSERGVQLWAVVPDVLQVTTEVHNVE<br>CPGHRSLLLPASHAKAWLGMPLLNAAHPDLHYLQVIPTRD<br>SLYRNLFRHGVIHRVIAHEHTGLLATPERIRVENSFMRKG<br>GKPWEYNLLSATPTLEMGIDIGNLSSVLLCSVPPAQANYL<br>QRVGRGGRRDGNSFVLTVANGRPHDLVFYADPSRMLDT<br>PVEPPAVFLKARYVLRRQLLAYAMDCWTRESRGDNLVPS<br>NMQPVLDAVEKVQEDRFPYTLLNFLKHNMQEIWDGFSAY | 66533 to 60099 | 98 | 37 | Tpye III RE Helicase |

| | | | | | |
|---|---|---|---|---|---|
| | VARELSGDDLELLRQFLFGGPQHLDDHLQLYVLGRLKVV<br>ADERANMAARSADLKKQIDKLAKAPQDEHTRDELAELER<br>EREGFNRLRYVMNRRETLNFLTDEGLLPNYAFPEEGATL<br>HSVIFRSEKATSGDGAERELIKREYEYQRPAQAALTELAP<br>ESVFYAGNRKVKISRVETAKGRNIQDWRFCPRCHYSAAA<br>DNPTAGFNDKLCPRCHTSQWGDSSARTKMLKMTQVYAF<br>TNARDALLNDHSDDREPVFFSKQMLIDFKPADIHITWVLD<br>DKDRPFGFEFIRSATFLEVNFGRRDGDEMMFEVAGTQLQ<br>RSGFAVCRECGSVQSRKVMAGKGEPAHLKSCSYAKGVK<br>KLVSGGDDSGLDNCLYLYRQFTSEALRILLPRLATGGTEE<br>QINSFVAALQLGLKRRFGGKVDHLRVAHQSEPIGETDERR<br>HFIVLYDSVPGGTGYLHELLSRAENMQAVFRMAYDVMEA<br>CDCYEHTMDGCYRCLLEYRNAYGMENTSKAIALEMLKDI<br>VEGEHEWVQNSESLSSLSGNPWVESELEARFPEALARFS<br>GMDCVSGQKVRVSADIIHGKTGYRLTIGEQAYEMEPQVD<br>LGKAEGVQFASRPDFVLWPASADFKPVAIFLDGYQFHGQ<br>KSSEDLIKRQSLMHAGFVVWTLNWYDVNKVIGDKALDVP<br>LLTGMTSPAQHHQAIAGLAKLAGAANTAKHLSMPTFDLLM<br>HFLADQIPQALSEQALFFILQCLPAASLADAEVRAKLLQSL<br>HGLPVSFTDQQPEPIALAGAVEVKDAQGAALISLGLIAGAD<br>LVRQFDLNQALVSFCYDLKQGSEEAARYQWQRFWAAVN<br>FLQFLPLFYAWTPQSKHDGTAAGLLWPLVGAGAEKSVEA<br>HQLPDWFQLLDENVATALGEHQIAWPMSAVVGDAVMDE<br>QGEVVGEAELLLPEQKIALLLEHLEDQQAAMAYLQKAGWI<br>IVSSADDLAKAITQLNSGA | | | | |
| **Contig_98** | MPVVIVENDTSQWEDETGAVYHFPKRYQAWLAQGTEVIY<br>YKGRIKDKAFASVRLSTDPHYFGKARIGQVYADRRSDKG<br>DLFALIENFTPFEDAVPSKIDGDYLETIPASRMSNYWRDG<br>VRPISQSDYDAILSHATLLPSRADAFVPDTEDDPLTFESAS<br>EGSKTSYFGTRYERRKDLRVKAIAIHGLDCKACGFDFEEA<br>YGEHAKGFIHVHHVVPISDFGGEKAVNPETDLVTLCANCH<br>AVVHRKRDKTLSVDELKGMLRGRWVIESQ | 51421 to 50618 | 100(99) | 97(91) | HNH/ RE |

| | | | | | |
|---|---|---|---|---|---|
| **Contig_101** | MNMRNPEDALLDSWQHNAQAWIDAVRSGSIESRRQVTD QAILLAILGRQPERVLDLGCGEGWLLRALGDRGVEAVGV DGDRALVDAARAAGSAEVHLASYAQLAAGQAYVGKDYDL ICANFALLQQDIIPLLAAMNALLAPGGALVIQTLHPWSVAD GDYQDGWREESFAGFAGDWQVMPWYFRTLASWLNALD MAGLRLVSLQEPQHPQSALPQSLLLVAERP | 63047 to 63718 | 100 | 99 | Mtase Type 12 |
| **Contig_102** | MHSNPVSISLSSYGADFVRQRGQEQFLDLLAAAGVTRVE LREELFTCAPDTAALAAAIAALRLECLYSTPLELWTAQGVP DPQLAQKLETARALGAVALKVSLGHYHAGCDVAALATLLP AHGPLLLVENDQTAHGGRIEPLQQFFQRADELGLTLGMT FDIGNWQWQGEPARHAARQLGRWVRYVHCKAVQCRAD GRLVAVPPEASDLQEWAELMAEFTPGVVRAVEYPLVSDD LLALTRAQVRDLAALGQGVSSEELSHA | 5839 to 6624 | 100 | 100 | AP Endonuclease |
| **Contig_103** | MGPDPTVMAFDDATNTGQADPCAFEVFHAVQALEHAEQ FAGVGHVEAHAVVANADLGFARVLHGADADARRGAPTC VFDGVGQQVVQGHVDERRVTDHLGQLGDVPDNFAVLVV GRQFAANGLDQGVEVDLGQAQGRAAHLREVEQVVDQAS GKMRGFLDVLQKAPAALAEALALDFAEQFGVAGNMAQG CAQVMGHAVGKRFELLVRATQFASQLRQFLGLAQDNPQ HCRAQLLHAFDDQRVPGFAVAAQFFLPAIETVPRVEVALR PDLFVRLPGPVHRAHLLGTEPQQVVRVDLRNGHCQHTAC MGRELGQFIDVAAQVVAVEHALFATLG | 8863 to 7862 | 35 | 29 | HNH Endonuclease |
| **Contig_108** | MNCRGCGSALHLPLIDLGTAPPSNAYLRAEQLAGAEQWV PLKVSVCEQCWLVQTEDYTRAEQLFDADYAYFSSYSSSW LGHAEAYVAGMAERFALSADSRVVEIAANDGYLLQYVAR RGIPCLGVEPTRSTADAARAKGLEIREVFFGRDVATQLVA EGWSADLMAANNVLAHVPDINDFLGGFATLLKPTGVATFE FPHLLSLIAEHQFDTLYHEHYSYLSLTAVQILCQRNGLEIFD VQELPTHGGSLRVFVQRTDGQRRTVEPAVTGLLALENDV GVRSAGFYSTLAPAAERIKLQLLRFLLDAKAAGKRVVGYG AAAKGNTLLNYAGVKADLLAWVADASPHKQGKFLPGSRI PIVAPERLAEEQPDYVLVLPWNLLHEISEQQAGIREWGGQ FVIAVPELTVL | 15221 to 13998 | 100 | 99 | SAM-dependant Mtase |

| | | | | | |
|---|---|---|---|---|---|
| **Contig_113** | MTENQQQYDVFADVYEILFDDSLYLSWFTYATETMQAFD KFLEGQDYWVDLGAGGGQFAILMAQAGYPIKGLDLSEKM VSAAKANAKEAKLDLEFWQDDMTTFELKDQAAVISCFCD TINYLADANAVKATFAHIYQQLQDGGIFMFDVHSIHQINDIY PETSFVVEWDDAVFTWTSDQFRGENTIDHTINVFVQNTE DNSYQRFEEMHYEQTLSIEAYQEILTQVGFKNIRVTADFS QDAPDETSKRIFFSAQK | 3857 to 3093 | 100 | 99 | Mtase |
| **Contig_114** | MSTPLDTNTLKARQQAAWASGDYAVIGTTLQLVGERLAE ACDLRWDEQVLDVAAGNGNATLAAARRGCCVTSTDYVP ELLKRGEERARAEHLNVVFQVADAEALPFADGTFDAVLST FGVMFAPDQAQAARELSRVCRPGGRIGLANWTPQGFVG QMFKTLGRHVPPPAGALPPSRWGDEEQLRVMFEGALGE LKVSRQHFNFRYRSAAHFIEVFRTWYGPVHKAFASLEPEA AGALERDLTQLLNESNVGGSSSLVVPSEYLEVVIIRG | 66255 to 65446 | 100 | 99 | SAM-dependant Mtase |
| **Contig_60** | MTTRPASMTWRSAIPARTLMMSNPTLSVSKFAGLGPLFG GLARNAVLARLGRLRHGHLRLLSHGQQWSFGDAHSPLQ AEVEILDDITWGLIAGNGSIGAGEAYIHGYWRSPDLALVTR LFVANLEVLDALEGGLARFGRPALRLQHKLNRNNRRGAR RNILAHYDLGNALFERLLDPTMMYSAAQFEHPGQTLEQA QLHKLERICQKLELSPADHLLEIGCGWGSLAIHAATRYGC KVTTTTLSEAQYSHTLQRVQALGLEQRVKVLREDYRDLQ GTFDKLVSIEMIEAVGHRYLPVYFRQCASLLKPDGLMLLQ AITIRDQRYEQARRSVDFIQRYIFPGGALPSLSVMLETASR HTALNLLHMEDFGQDYAHTLRHWRENLRQARAALTDLGY DDMFQRLWEFYLCYCQGGFEERTIGVAHLLWAAPQARR APLPGSA | 36371 to 37690 | 95 | 92 | SAM-dependant Mtase |
| **Contig_68** | MQAPVDQPRPTSRLRIGNEARRARGNLGDTQIDQHRLRR LPAGVPGLQRYRPVIPLPQAFEETSRMLGVMAKRWRQLY QQATQAVAQLPALSKKPVQRLLAPAEPLLMADGLGHLHR KAEMLWYGRSPTGVGFGAMGSVEGAVDFHRIEAAGVAL QVAAGLRKCLDVTARQAPASATQMNHSTTCWFATQRLAL TRDAPGIPNTPQQRPPHGFPCALVVDPFV | 39416 to 40084 | 61 | 35 | Endonuclease |

| | | | | | |
|---|---|---|---|---|---|
| **Contig_68** | MLPHFTAPRRSPGRATRPGPGAPGRAAWPSLHRGTGRR<br>RYRPHRAPAAHGHPAQDDARLGHVHQHGQRRPAPPTG<br>VDQRAGQPQGNDYNACPNLKYVFPMKSLLYAVSLLFSFT<br>LPALASPPATFTEAKVVAKQKVYMDQASSAMGDLYCGCK<br>WAWVGKSGGRIDAASCGYQTRKQQNRAERTEWEHIVPA<br>YTFGNQRQCWKNGGREHCVDDDPVFRAMEADLFNLYPA<br>VGEVNGDRSNFNYGMVAGNAGQYGQCTTKVDFVQRAA<br>EPRDEVKGLVARTTFYMYDRYKLSMSRQQQQLLMAWDK<br>QHPVSAWEKERDRRIAAIMGHANPFVTGERKWTANYKPV<br>GSGVVQAVPAKTAKPEAKPSLASAGSVGAVLGNRNSHVY<br>HLSVGCPGYTQVTAKNQVTFATEGEAQAAGYRKAGNCR | 56013 to 57272 | 76 | 81 | Endonuclease I domain |
| **Contig_62** | MMTSPIQTLEQHLLAALDPAPQETRRLFHGRGRCWAGLE<br>QVTVDWLQGVLSVALFREPAEGQLAELEAMLRTIAERPQ<br>WTGQAILLQHRYLPDSPGQWLLGEPCQQREVVEDGLTYL<br>LDLGVRQNNGLFLDMRYGRRWVREQAAGKRVLNLFAYT<br>CGFSVAAIAGGAEQVVNLDMAKSALSRGRDNHRLNGHD<br>ASRVAYLGHELFKSWGKVRKYGPYDLIIIDPPTFQRGSFVL<br>TQDYAKILRRLPELLSEGGTVLACVNDPGIGPEFLIEGMVE<br>QAPSLKFVERLENPPEFPDVDPAGGLKALVFRQA | 64392 to 65318 | 99 | 99 | Mtase |
| **Contig_5** | MTVRKVVTRRSNHYRGYFPSLKNKKPVPWESQLEGALFR<br>LLELSPAVIGYVPQPSEERVPSLQGYFKYYPDVQVFLADG<br>REWWFEVKPHDRLKIASVRQRLDAAERYFNATARNFSVIT<br>EKLIEAEPLATNLRRLMYHRRGPELSHQALEEVMATFNE<br>WPPMTVADLLYVVGEGKAWRLLGLGVVGIDLDRSIDTDS<br>PVFLQGGHRHANLFP | 62408 to 61773 | 100 | 100 | TnsA endonuclease |
| **Contig_5** | MLAQLPPALQSLHLPLRLKLWDGNQFDLGPSPQVTILVKE<br>PQLISQLSHPSMDQLGTAFVEGKLELEGDIGEAIRVCDELS<br>EALLTDEDDAPPQRRAHDKSTDAEAISYHYDVSNAFYQL<br>WLDQDMAYSCAYFREPDNTLDQAQQDKFDHLCRKLRLD<br>AGDYLLDVGCGWGGLARFAAREYGAKVFGITLSKEQLKL<br>GRERVKAEGLADKVDLQILDYRDLPQDGRFDKVVSVGMF<br>EHVGHANLALYSQKLFGAVREGGLVMNHGITAKHVDGRP<br>VGRGAGEFIDRYVFPHGELPHLSMISASICEAGLEVVDVE<br>SLRLHYAKTLHHWSENLENQLHKAAALVPEKTLRIWRLYL<br>AGCAYAFQKGWINLHQILAVKPYPDGHHDLPWTREDLYR | 104519 to 105700 | 99 | 100 | SAM-dependant Mtase |

| | | | | | |
|---|---|---|---|---|---|
| **Contig_6** | MISELVLNAAAGGVLGLLIGSFLNVVIHRLPLMMEQEWHA EGAQWAEEQKDKGARIELPPAKPAITLSQPRSRCPHCGH QIAWYENIPVLSYLFLRGRCAECKTPISLRYPVVELVCAAL FAFCLGRDGLTATGFAWCGFSAALLALALIDWDTTFLPDSI TLPLLWAGLIASALQWTSVPLQQSLWGAVAGYMSLWLIF WAFNLATGKEGMGYGDFKLFAALGAWFGWQALVPIILMA SVVGAVIGIALKINSKLREGGYVPFGPFLAGGGFVSLIWGP QAVLSFAGL | 13934 to 14800 | 100 | 99 | Mtase |
| **Contig_6** | MGKTDKKQSLQKPAKAGALDIAEQLGLKPGQSLELLKALH ILTREGKLNQDSRRKLKQVYHLYQFIEPLLAELSKDGHAVT LADHGAGKSYLGFILYDLYFKALAQGRIFGIETRAPLVEAS QKLAIELGFERMEFLNMSVAESTRADFMPSQFDVVTALHA CDTATDDAIAFGLEKQAKAMVLVPCCQAEVAACLRQTKA MSLARTPLAELWRHPIHTREMGSQITNVLRCLYLEACGYQ VTVTELVGWEHSMKNELIVARYTGQKKRSAAQRLRQLLA EFGLEGLAGVRYPHLQQTSAD | 102182 to 101280 | 100 | 100 | SAM-dependant Mtase |
| **Contig_6** | MSQLSEIIRDIERVSAEVVQKASSFQGAILADVAGRRGTM NARVAPVHQDMKLAGPAFTVEVRPGDNLMIHAAIALAKPG DILVIDGKGDQTAALMGTLMLSACKKTGLGGVIVDGAIRDK LELLELGFPVFSAGFNPAGPTKFVPGRINHPISCGGATVN PGDLVVGDADGVVVIERAKAPAMLALADRKVVDEAARIEA IARGDTASRWLPAALRAAGVLKEGETL | 301140 to 300457 | 100 | 88 | Mtase |
| **Contig_7** | MEKQDSLKPIVFERQKFDLDDIREMHTLYIENYPIVYILHQN KETKSRPKAYIGQTVHVHNRMRDHLKNKARKDLTDALFIG HQTFNQSATFNIETNLINYFIADNQFSLQNVSQTANLSMH NYYQKNLYDDDLFEDIWESLRHEGLAKETTDNLKNRDIYK LSPFKTLSEPQRALKENILNYCKRVMQDIRQEKSVKKKIYV IHGEAGTGKSVVLSSLFNTLQEEARQSHSVLARTKNFLLV NHSEMLKTYQQISKSLPYLKKKDFDKPTPFINNQKIEQADI VLIDEGHLLLSQPDAYNNFHGENQLDAIMDKAKVAVLIFDE KQYLKVKSKWQSDDLKAILKKYDADEYHLTDQFRMNASP AIMNWVNEFVGQKVTPLPPSDDHFEFKVFDDPTAFKDYIY QKNADKGLSRIVSTFDFAHKKNGDVYLVDEAGINLPWNTT NSKMTWSERPETINEVGSIYTVQGFDLNYVGIVLGPSIDY DFETERLTIDPSKYQDTGGYSGANRFASHKEAMAAKEQII LNSINVLMKRGIYGLAIYAVNDNLKRKLQDMRVEAMQDDI | 62231 to 60540 | 100 | 95(94) | Endonuclease (GIY-YIG) |

| | | | | | |
|---|---|---|---|---|---|
| **Contig_9** | MQLSLFGVLNAYEAGAQVNAKAYERLGQDLGITAEAWAV RQPVGTAGQPHSPLKRRVRWYQQTLKRLGLLEPVAGKR GVWRPTAAGRRTIEQRQQDLEPAAPGLVQLGFSTELGMA LWADCKDAFSRLDEPVHLVLTSPPYPLARQRDYGGPERA EYVDWLCACLEPVVARLASGGSLFLSVSNDIFETGSPARS LYRERLVLALHERLGLHKMDEWIWHNPSKAPGPVAWASK RRVQVNTAWEPIYWFSNDPQACFADNRRVLQPHSEKHA RLIASGGTKTAAVFADGANRRRAGAFGAQTAGRIPRNLIT VPHNCPSQTALRAWAKAEGIPIHGATMPLALAEHVVRFAS EPGQLVADPFGGWATTALASELNLRRWVITERMRAYLYA SQWRMALSFHQPAHSKG | 94217 to 92994 | 100 | 84 | Mtase |
| **Contig_9** | MNKAHTTGYVAPLGRYVPPVLFNPYTGEPRDARDIASDP KGVLIVPPGAQLAATNNPAAAPVVLPEPAYTLRVRGVFQD TTPAANAFGIPDGEHKLFTEPQVRALLAGVSAPAAEGSEI RWVSNGLIGARPTTHDLREAIAAAEDCQRCDCTDCTKAM TLPLKLVEAINTAVGSNEWQGDSVVTDLLEPACKVVANLT HRAQADQKARMKIAAALGHEGVNFAWSYLTGAIKELVKA DGEHLELQPQALAFQQRVQPWMMACFGPEISADRIERNH RFLEEALELVQSCGSTASEAHQLVDYVFARPVGDPMQES GGVMVTLAALCLASGLDMHACGEAELARIWTKVEAIRAK QAAKPKHSPLPMYAPQTQDALRLAFVLSEIRRDGMDALG AALHDADGFPLDHKASLEAIDRAANAAASGK | 91023 to 89752 | 37 | 57 | HNH |
| **Contig_9** | MPTRWPGRRAPPGVHHRQAGVAAQRRKWHRNPAHPGP DRRSLVHRLHVVHQGLSHRRDSGCQQAHALRDCRALHG LRAVHSRLSGRLHRTGQCQCRGHGLVSLERCPGRACPT PLWSASAAHGSQGQCAGAHHGTASRGRSRCSGRHDGS ASTRQKGCDCRHSGQGQGSARPELIPNKPLSLMQKAPA ALDSIASPGAAKLSQGLIRR | 81878 to 81255 | 39 | 43 | Endonuclease III |

| | | | | | |
|---|---|---|---|---|---|
| Contig_9 | MKRDDFTLPLDLGRELIIDNFAGGGGTSKGLEWAFGRPV DIAINHDPEALAMHAINHPYTKHLCESVWEVDPIAVTGNRP VGLVWLSPDCKHFSKAKGGTPVSKKIRGLAWVGLRWIAK TKPRIMMLENVEEFQDWGPLVVDANGNARPDPKKKGRT FKSFERQLRAHGYSVEWRELRACDQGAPTIRKRLFLVAR RDGIPIHWGGPSHAAPTDHRVIAGLLAAHRTAAQCIDFDL PAESIFGRKRDLVTNTLRRVAKGVFRHVLNTATPFIVNTRN GEREGQEPRIRDVNAPYWTVTSQGSQGALAAPVLAPFIN EHANGSNQRTMPADQPLRTVCAQVKGGHFSVVTPELQP LKPVDGAAQIIVPLRGTSEQHLGGHSVQSPLSTVSAGGRH HALAAAHITKFNTGAVGSSLNEPLPTVTAGGKPKRPSTGIT MGMVAAHLVDMGHGEGPAGGKRWSHGTRNIEMPLNTV TASGATSALAAVCLEQAYGGFYDGDGRTADEPLSTITTSG TQQRLITACLVKYYSEGGQDSSCSEPMHTVPTKARMALV QTAKVPASRLAPEHTERAKLCADLLREHLPEQFPEPADVV LMWHNGQWWALVDITLRMLKPRELARAQGFPDSYQIEEI PDPAILFKDGVQAVDDPRDIPRIKLTTTAQVRMIGNSVSPY MAAALARWNFQHEAQMYA | 103909 to 101843 | 100 | 96 | Mtase |
| Contig_9 | MLVWSKHFPSWARRRLLMTDSAEIVAQAALHRSRQLTEIT SLAALTAGMSGQDRRLFARELKELAARGELRRFRFDGRL HFAPLSRAGDAAFVAKFALESTACGCDEGSCMPWSGKF FKPRQGPVVNLDGKEYVLRRVIYEVRTGKKLAQSESVRP SCGEANCIHPQHLAKEPRNTPLIGRARLPSTREKLARAQQ AKAAYSMEEVQALRGRCIRGELTRVAAAKCLGVSVETME RMVNGLAWRDFSNPYQSLTA | 117463 to 118227 | 29 | 26 | HNH |
| Contig_9 | MMSTRFELYEGDCLQTLRGLAENCVDSIVTDPPYGLAFM GKKWDYDVPSTEIWVECLRVLKPGGHLLAFAGTRTQHRM AVRIEDAGFEIRDMIAWLYGSGFPKSRNIANDMQEPGNAA AWAGWGTALKPALEPITLARKPLSGTVAGNVIAHGTGALN IGDCRVPAEPMPPNTGSGGLPRRSEDEQRGPGVVSQPH EAGRWPANLIHDGSDEVVSLFPAQAGAAAPVLRRHGDKF RNSYGGFAGNDNEGGSSFHGDSGSAARFFYVPKASRSD RNEGCEGLERKPLLWSSGTQNPGSFQAEGTDKRSQNNH PTVKPTALMGYLCRLVTPSGGVVLDPFMGSGSTGKAAMR EGFSFIGCELNAEYLAIAKARIEHELARATSAPETTTRINQR DLFAEVSL | 119955 to 121154 | 98 | 67 | Mtase |

| | | | | | |
|---|---|---|---|---|---|
| **Contig_10** | MRGFGCGHCYSFRKMTSNTQPPTPAAIPAIAASHCAPAP SAWIARFAHLLRPQGSVLDLACGMGRHTRFLSALNHALT SVDKAPEATRSVADIAETITADIENDAWPLTGRSFDGVVVT NYLWRPLWTQILNSVRPGGVLLYETFAQGNEAYGKPSRP DFLLAPGELLQVCAGWSIVAYEHGLLEQPARVVQRIAAIR PDGAAATVAPAALLQA | 15806 to 16447 | 100 | 100 | SAM-dependant Mtase |
| **Contig_9** | MLAKRLFETEVDALPHPGQGLLLAAQGQAAVGGLQLPEA ADCSPKALHPLPQQGTDLQHPRLPQRAGAGVIGVAVHAQ QAQARCDLGLGTRGGMGVDVGLVDDHQVGQLHHALFDS LQVVARIGQLQHAEHVRHAVHGDLALAHAHGLDDDHVVT GGFADQHGLARLFGHTAQRAAAGAGTDIGLQPHGQLLHA GLVAEDGAARDGAGRIDRQHCHTQPLLYQVQTQGFDEG GFADTRHATDAKAKRLARMRQQSGEQAVCLFTVVGPGG FKQGNGLGHGPALHGGIAVQDAVLHLLRGHVAIAFAIDAC RPIRDRWPS | 183556 to 182600 | 26 | 27 | RE Type II |

| | | | | | |
|---|---|---|---|---|---|
| **Contig_9** | MSANDQFDPSRAVPVNSAASQKPEDDFNPGRAVPMNAD<br>NALVRGWKSMSNSVGITKDLATGDAAAVARRVKEFDDYN<br>RVNPGSQQGQELSKAWEQGDGITGGIAGVAGEIAKDWK<br>EAPNWVGGVRSVAGNAKAMGDGLVAQVPNMLAPMVGM<br>VGGGVAGGAAAGPVGLAAGAWAGASAGNTLVEGGGML<br>MDRLNKAGINPQDTAAVEKFTREQGDAALGDAAIKGAIIG<br>AVDAATAGAGGKILNAPARAAADRALTKMGVDMTDGAAV<br>KAAQKSDAFKDLVAKDATFQAASSGAGNVARNVGVAALD<br>PAGEFAGEFVGQGVATGDWDTKNAALEAISSVGQSALMY<br>GGQKAYQAATSPLRDKRANESDKADGEPEAGNPVPLML<br>TNQPPDTLFTHPDGSTGRRSEMESYLNSLNDGTPAGERR<br>MMKERARLLGYSSPEPYEFEIPDLVPWQEDYRNHSFGLP<br>YQPLQDHLGAVQAAVRGGTQFTDQQTVDAVRTTMEDAW<br>LEANVKQQPGSAVEAASIAGQQMEQELAAAAPGETGSTA<br>AVDAQVRTSRILSGLQNAMDSGAMGTAQSINYLNEALTRI<br>GEQPLAQDEAARARRLLDAHAAFTGKTTPAPLPGDLAAQ<br>NPLVDEFADNAALESLIRRKPSERLGIDPSAGPNSKAAAM<br>AVDAAQSEPAAAATDVAQPLSLQNPLLALPSSRENAAPAA<br>SPFNINGDLDGREADQAQQITAKQAQARPAQGSEVGGR<br>QDQSISGIGSAPGHLGQGVGYGTAPAINNGPQVPAQNAT<br>QGQASEAQTATETGSRETREPSTFAKYAGQSESSEANG<br>AINSEAKQAAQPASDAMTIGSTPKTGDAVSVRDGVVYLG<br>NYEALDYDSGEPITVPADATRGQVAQALRDGGVLTDRQK<br>VFGLKDREHERISGPVAPAVQAPAPAPAANAPAPSPARL<br>RQQKAIERINSEKGAYFFSKSKADAFLADNDLQDGYEVVQ<br>QEKAFAIKAKQAIAAAVEAEAPAAGAANPVANQNSGTGIA<br>GRDLGDGWVEFAPESGTKSVPRAEMPQIKAEHRGAMVN<br>FMNARGVAHEEATVPASSLKPTQREFSREKVEKAKTYAG<br>GNRAILTSNDGHVLDGHHQWVAALENGEDVRTIRLDAPID<br>GLVELAHEFPSSTVAEGAETGQSGTADKSLEAPVSATQS<br>NVDFLKALAEHRKDLPEGFEIVPMGDTISLKEGGKFAVNG<br>LSRDRAGVVEALRLAKERDQKNKAGSASTAPEHGQAGV<br>DGRDEVKDVTDVTPKSEQSVVDQMQSAKHESERAAREA<br>AAKKAAEQEARKASREVERDLFRQMTAANGGMAGKSLD<br>QVTDALKQQLGSRDASTALNPVQISRLARAIQKQLHANAA<br>QLAKEAARDPQVKADAQAEALGNAFGEDVSPQSKGDFLL | 137556 to 152900 | 19 | 70 | SAM-dependant Mtase |

GFDHALAGKTKSTLSGAGLADMVKGYEAASEWMGTEEG
AAWFEGKRRKKLENTGVDLRRHWEAMKAQMKAGESDM
QKAWKQIEAATNRAELFAPYLPEGSTPGWVAYVSQLRGI
TKTFKQFMFDGRASWYGSVAVSRKGKEAANLEFVLDGK
RYPGSMSIEQRQQWQTDPVFRMAQLRDAAELYIEKVQEL
TAFLDGATSLQEAAERFVDTMVADKHQEKARNRDTYYSR
YERINADRLDGVLYGVHGEGVFDRGVWNGDSYDFTKFR
AASPWASGLIANEATHALPTRATPLTPPKLDRVDRAMPKD
HRGGKDVTPAQFKAQFGFADVGFGNWVGAKNDQDHLN
YAFDAFMDLAEHFGFAPKNIGLGGVLHFTVGALGHGKFA
AHFSPNHPGPNGRVQVMNLTNTKGDGTVYHEWAHALDH
NLGGEWDRVKVMILNAFKFKAYEASDWERVARNFLVGG
SYWQGNKNQDKVDAAIQGLRYYANTGARRGLTAYKENA
DKLGKDYWGNDAELLARAVEAWSADSLGGINSYLVNQD
WVGDGKVTQASGYRGTPYPTGGERVVFGQYLTALAKSV
KFTDGKPTVTVTDFERNLPADMGAGEVRRRELLSREGM
QAYFEQVQEERAMAAEEKARLEVEKQAEDKAKVDAMAE
QALAELEAMSKPAVVDAPAPSESRGPLSNDDLSAIFDQAA
AELREQTQEQPNVSATNEAAAISETVAPAVQAAAGQADK
TAAKLIAEAAKLGVTGANEALSGLAKLFGGGKGGRLNSFP
AGFDEETYKAAKPHFKAALSSFQAAGKSLKDLFKLLIQNF
GDGVKDYAIQFAKDEGLSAQLGKAPTAGAARSPSGVLAD
WVKSQLERGAADGSFDWRALFEQADSAFGGTQAEGKYT
PKDAYDAMEAGVNQFILSRPGEFNPNASQDAAHIIVERLH
RITQMLPTQTKRTAEQDEFQQFSTVPALAYAANWAANMD
SSDTMLEPSAGIGGLAAFAKNAGAKLILNELSSRRAAVLR
EVFPAAKVFTENAEQIDNILPATEIPSVVVMNPPFSATAGRI
QGKRDTHVGAQHVEQGLKRLADGGRLVAIVGEGMNLDR
PAFADWWKGIRAKYDVRAVIPMDGSGYAKYGTTFDNAIL
VIDKVKPSNRPIVTTPAKTYSDLIGSLAEIRNDRPESIFPSN
DRDGLELDAAEHALAESSQAGRGSAQPEQSGSDQRSDV
GRAESGRGQGLGAGGGGRGSAGGSGVAGNDGAKSRP
GRGAGSNDARGSRDAAATGGGGSDAAQSALSIQATEQA
GDPGAGLSDSIFESYQPQRLQVPGAKPHPGPLVQSSAMA
AVLPPVPTYTPNLPKETIEKGLLSIAQIEAVVYAGQAHQELL
EPITVDGKDVAYRRGFFIGDGTGVGKGREISGIILDNMRQ

GREKAVWISEKPGLLPDAQRDYSGIGGDPKQIFNVSKTKA
EEQINADKGVAFLSYATLRSGAKSQENVATPTTKAQFAKQ
FPKGLDVVTTNGRGTFPLDHIDPKDDGGKVWVKVGGELK
FIRYTSVESIGGQSDWLSGRTAQPAEGSKKEGQSRLDQL
VNWLGKDFDGVIAFDEAHNAGNAVAQKGTRGQSQPSAQ
ALAVVDLQKRLPNARVVYVSATGATQVSNLSFATRLGLW
GPATPFASVQNFIAEMTAGGLAAMELVARDLKQMGAYMA
RSLSFEGVTYSRVEHQLSPLQQDIYNRLAEAWQVTLQNID
AALKTTGAVGENGKGSSSAKSAAMSAYWGAQQRFFSQV
ITSMQMPSVLEQMERDVADGKALVLQLVNTNEAQQNRSI
AKRREEDESADLEELDLTPRDVLMQMVEKSFPVTQFEEQ
EDDSGKKIRVPAKDSQGNPVINREAVAMREALLKDLKDIR
VPDGPLEIVLNHFGVDKVAEVTGRTQRVVRKLDKDGELK
AQLESRGPASARADANAFMADQKPILVFSDAGGTGYSFH
ADNTQKNKRKRSHYLIQPGWRADKAVQGFGRTHRTNQA
SAPHYYLASTNVPSQKRFLSAIARRLDQLGALTKGQRDTA
NQGMFSEKDNLESIYATQAVQQFFKDGQHQQLDGISFAE
FLRQTGLESIIDEETNRIAEDRMPDTRTFLNRMLSLKLDMQ
EKVFDAFMLRMEEKVEMAVERGEFDAGLQTIRALESRVV
ADDLAYTDPRSGAETRLVELELTQPTTIYPFPHSLQKAEY
VVNVKSGKVYAKTLVGKSTTKEGAVVDRFRFYGTGSVQS
KTAPEIAKAFRGSTKAEAMKLWAAENEARPKTYTERKHMI
VGAMLPIWDRLKTDGSIQVARTMTVDGSRLLGRVIDKKSL
PDVRKRLNVSSAASKMSPAQVMAQILKGDKAELANGWSL
ERARVSDDLRIELKNPSGGYISPAVRADLVGIGLVSERISW
AERLFVPTGAAGVPVLERLTKNRPVVDLQGEQDTSEAHF
GAGQGAKVGSRTDRAVMDMVREGKSAADILGLIASTSKT
PFNRKLSALLVKAGAAPRISMGGNMGADGGFNFLAKYSR
KLEELTLSEGAASRAEQIFLHEMTHAATLKSLDRKGIASLQ
MRNLYEHVKKQGGAAGAYGMKNVGEFVAEAFTNPEFQR
ALRGMKAPAGSSLQNAWDAFVRILKAVLGLPAKSEDALS
RALELGVLVMREDRALRQASQKDRATQSSVRQFLDGNA
VASMRGDEVPRLGGQALVVNWAAKHFESATKDGAVVHP
ELGQIKLDRRSAKDSLSHGYGKDKVQALYLLPEVLPKARV
LHTESRKEGQTGYVLGAPVEIGGKAYVAAMVVTQHEGRT
GLYVHEVVLREKLQGAINTGAPAQSQADIAQSSRGNPGAI

| | | | | | |
|---|---|---|---|---|---|
| | RSVLERIYAVNPDGNSDDVAHFGVSDVAEAAKNIGQGLK<br>AITATDVKKAGSHKLTDWLKLGLQFMGRRQLVDIYGDTIP<br>MAEYDRLVAQMEADKNDVGAEADNLATRWGKLKDEVQL<br>SELMHDATLAEIDADSSKEYVEGDDKAQSAALKRRFSALS<br>PEAKAVYREARDHYREHHKQVRNAIRERIMRAELSSKKR<br>QELLEKMDADFFGYVKGVYFPLARFGQYVVATKDQDGKV<br>ISVSRAETMTEAERMRDEMRKAFPSAKGFNVGRVTLSKD<br>FVATHDMVGRGFMSEVFAALDKHEIPADKRAELEDTLGQ<br>LYLSSLPDLSWAKHGIHRKGTPGFSQDARRAFAQNSFHG<br>ARYLAKLRYSDLMADELDNMQKHVDTMGAFKDDFDQPK<br>AQRVVDEMQKRHDQLMNPKTNSLSTALTSFGFIFHLGLS<br>PAAAMVNLSQTALVAYPVMGAKWGFQKSAAALLRASNES<br>VKGKNDIRTQLKDKDEIAAYDEAVRSGVIDVTMAHDLAGIA<br>QGEDAGVMWKLRPVMRAASFLFHHAERFNRQATFIASYR<br>LAREAGSDHYKAYEDAVKATYDGHFDYGSANRPRLMQG<br>NVARVVLLFKQFAQNMIYTLGRNAYLAAKGDKQALKTFAG<br>VVTMHAAGAGVLGLPLVGPLLALASALGGDDDDPWDAEI<br>ALRNMLADVFGQQVSEVIAKGFSRLTPWDISGRVGLDNLI<br>FPDIREGLEGKMWAQEMATGLLGPVVGIGINGARGAQLL<br>AEGDFMRGLESMMPVVPRNAIKSARFLQEGAKDSTGITIK<br>DDVSALGIAGQLVGFSPSEVRLAFEGRGAVMNADRRLNV<br>RRAELLGAFSHAVMKKDSEAQAAAREEIKAFNDKNPGRRI<br>TSPQMWQSVRARQRRIDQAQDGVYLPRNRRDAMDAGG<br>FAF | | | | |
| **Contig_10** | MPADTAFRHISEGASLLWRGDFQNARQLLLALGRRLDKK<br>SRRKSPAKSPAAAGFPHAFHLYRQSQAQRARMLASILIEL<br>DLQWHCALRRAPDWSQACSEAWGTVSAEVQAQSVLVPL<br>RDLLGVVGAHEWRKKGVEIPALDGARIHAHYGVFSPVRG<br>EYLDLVARASIPAAGIQQAWDIGVGTGVLSALLLKRGVKS<br>VVATDTSERALACACENLQRLGHASRVELQHADLFAQGQ<br>AGLIVCNPPWLPGKAASVLDQAIYDEDSRMLRGFLQGLAA<br>HLLPGGEGWLIISDLAEHLKLRTRDELLGWIEAAGLKVLGR<br>EDVRPHHGKVQDREDPLHIARSAEVTSLWRLAKAG | 138370 to 137318 | 100 | 100 | Mtase |

| | | | | | |
|---|---|---|---|---|---|
| **Contig_10** | MMRANIFTPQHLESGGFNPAQQLVTRAQLQMLGQVGDD EPAFTTWQQVCRQPLQKAAQHATVFVINGLIEHRCCLAR QPGRVADNQTGLPLRKQIGMLQLHPGSMTQALQILAGAG QRTLAGIGGDHRLDAALEQQCREHARADTNIPGLLNTGC GNARTRHQIEILAAHGRENTVVRMNAGAIQGRNLHPLLAP LMGADHPQQIAQRHQHGLRLHFGRNCAPGLGAGLTPVR SPAQGAMPLQIQLDQNAGQHACTLGLGLAVQMKSMGKA CCSRALCRALASRLLVQPAAQCQQQLTRILKITAPQQRRT FADMAKRRISGHGVIGNFDAARLGQAAFAAPQSRMGLGL LLPLQQVESRCGSAD | 137402 to 138496 | 41 | 30 | Mtase |
| **Contig_10** | MNILTWNVQWCCGMDGLVSVERIVRHALQMGEQSGGLD VLCMQEIAVNYPDLQGRPGDQLAELKALLPGWQIFFGAS VDEFTPRGHQRFGNLIATRLPVLLVQHYLLPMPAEADMR CMQRMCSVVTVDDAALGPVRIMTTHLEYFSARQRMAQA GALRALQMQACALADAPPQPASDGSPYQTKPHTRHAVL CGDFNFEPHEPEYAVLSAPWVAGEEGCLQAGQWRNSW DVLYPGQPQPPTFRLVDRTWGAEPGACDLIWVSDSLCQ RVHTWSVDSATQASDHQPVMLTLG | 256717 to 255845 | 100 | 100 | Endonuclease |
| **Contig_10** | MHLRLETPARISPIHALQMTGGRSDSHTCTALDSRPLPSL FFHAQDLTLLEPAPYRGSTRSGLAADARELGHARARCQA AGLDFSAQRPQLARSQHGRHAGAMDRLSQPRRQRRQA ACPLDAIQRCTRSPAAVSAWSALECDGLLATHTAPAGHG LFGAGRGLPRLWQEQPGPALAGLGRRGCPCRLGLAGPT GRRQAALYLRPLAGRRRCY | 81657 to 82292 | 43 | 34 | SAM-dependant Mtase |
| **Contig_11** | MLTPQFVLPVHTELVVDLFAGGGGASTGIEQAIGRHVDIA VNHDPEAVSLHTANHPQTRHFCSDVFEVDPLAVTDGQPV GLLWASPDCKHFSKAKGGKPVSKKIRGLAWVVIKWAKLT RPRVICLENVEEFQTWGPLGVDSRPCPERKGQTFQRWV SQLRNLGYKVEWKELRACDFGAPTIRKRLFLVARRDGLPI SWPQPTHAQPDESGKVAKGFKAWRTAAECIDWSIAAPSI FERERPLADATCRRIAKGIDRYVVKTAKPYIVSLTHQGSDR TESLGEPFKTITGANRGEKALAVPTLVQTGYGERAGQAP RVPGLDKPLGTVVGSPKHALVQAFLAKHYTGVVGSDLQD SIGTVTSVDHHSLVTAHMTKFRAGSVGSAADEPLHTVTAG GTPARPSTGNTMGLVTANLVHLGHGEGKDGTKRFSHGIR DVAAPLNTVTAQGATAGLVTSHMVKLRNNQFGQSHEEPF | 4385 to 6175 | 100 | 93 | Type II M |

| | | | | | |
|---|---|---|---|---|---|
| | PTLTAGGGHAGEVRAFLVKYYSEGGQDASCGDPMHTIPT KDRMGLVMVHGEPYAIVDIGLRMLTPRELYRAQGFPESYII DRGAAGEAITKTAQVRMCGNSVCPPLSRAIVAANYSEAG QLRKVA | | | | |
| **Contig_10** | MHPKALLDACSELVKRALTFEHPADAVVSRFFRENRYLG PRERATLAETVYTVLRKKLLFEALAHSGSGARERRLAILGF AAVLREQAKKEGKVKSKDGQDSETFIKAALTPQELKWLAA CDGVKPEELMEAHRHNLPEWLVEPLKAQLGDGFWALAA SMEQAAPLDLRVNTLNDKRSDLRKELEKAGIKAEPTPFSPI GLRVDGKPALAKVDAFNRGAIEVQDEGSQLLALMLDAKR GEMVVDFCAGAGGKTLAIGAAMRNTGRLYAFDVSGHRLD ALKPRLARSGLSNVHPAAIAHERDERVKRLAGKIDRVLVD APCSGLGTLRRNPDLKWRQSVKAVQELTQKQAAILESSA RLVKAGGRLIYATCSILPEENEAIAEAFSAAHPEFVPLDAG EVLEQLKIADGDKLCSGGDEGRRYLRLWPHQHETDGFFA AVWVKKA | 330185 to 328857 | 100 | 100 | SAM-dependant Mtase |
| **Contig_3** | MHGHAGGALHQRLQNQGSSLLVMRLQPGLQTLCCAACH IGSGFARPGLACIGAGHGGRQTQQGAIGIAKQRNIGDRQR AYRLAVVAAGQAHEAVLGRMPLVAPVMGAHLQCDLGGR SAVTAVERMTEAGQAGQALGQLDHGGVGETGQHHMIEL AHLLGQRRPDMRMRVTEQIDPPGTDAVQITAALGVHQPG PGGVVNGNRRRGFVPLHLGAGMPDMAHAAAGKTVFHQK LLRRITRLVKTQAQRGLARTRGRAMQFAMRKPRRDGRG RFRDQLAADAPSPLRPSLRIKVVSASTSGLPVTSSLSP | 128438 to 129358 | 31 | 31 | HNH Endonuclease |
| **Contig_3** | MTKKWLASTWPLPALLVWLLAWVIFAGLARLLPWWLALL VGGGFSTAASLYGPNWWRRLIIAAGFPLSFLVLSASQLPA WGWLLPLALLLLIYPLNAWRDAPVFPTPLNALKGLAEVVQ LPGQALVLDAGCGMGDGLRALRSALPQARLNGLEWSWP LAIASALRCPWARVRRGDIWLADWSGYQLVYLFQRPESM GRAAVKAATEMQPGSWLVSLDFQLPGVEATACLQGNSR HKVWIYAIPLDGAERS | 282369 to 283118 | 97 | 72 | Mtase Type 12 |

| | | | | | |
|---|---|---|---|---|---|
| **Contig_3** | MFIGVEQHHALAAGNFHGNDLLLEAAGLDGGSGALLADD GQLVLHHAADLVALGHVFSSDAHMHFLPGVVQDAQHVV DALGIAHACAPARGHVEVGAAAHGFGAGANGHLAVAQR NRLSCRDDGLQARAAQAVDVEGRGFDGAAGVHRRHAG QVGVARVGGDDVAHHHMAHGVGSDAGTRNGGLDHGGG QLGVGNVLEAAAKSADGRARGADNEDVSGAHGLSPASR RQCLCLSDLKR | 303068 to 302355 | 42 | 28 | SAM-dependant Mtase |
| **Contig_3** | MRRAFMAGGNHHGGEGEVFRRNGVAVGGHARAARADV THLAAAVFGIKVGLELQGIEVVLARGKARNACAHPVFKAG AFGAVPSTGAVVLCDLAHMSLLKPLVNTKSDLLFVLLIKPG SFGETIIRRLARYQQAHFLVAAVETGIGVAFCPRPHRHGP ENRHKKRSAKRSCELALKDKMKNSQASRSSLRRILPTGV FGRSVRKTMWRGRL | 547655 to 547023 | 23 | 46 | Endo/exo/phosphotase |

**Table A3: Functional assignment of predicted ORFs. Functional annotation was performed by RAST**

| Category | Subcategory | Subsystem | Role | Features |
|---|---|---|---|---|
| DNA Metabolism | DNA Metabolism - no subcategory | Type I Restriction-Modification | Type I restriction-modification system, specificity subunit S (EC 3.1.21.3) | fig\|6666666.224586.peg.10536, fig\|6666666.224586.peg.13715 |
| DNA Metabolism | DNA Metabolism - no subcategory | Type I Restriction-Modification | Type I restriction-modification system, restriction subunit R (EC 3.1.21.3) | fig\|6666666.224586.peg.4435, fig\|6666666.224586.peg.7008, fig\|6666666.224586.peg.10537, fig\|6666666.224586.peg.13716, fig\|6666666.224586.peg.13797 |
| DNA Metabolism | DNA Metabolism - no subcategory | Type I Restriction-Modification | Type I restriction-modification system, DNA-methyltransferase subunit M (EC 2.1.1.72) | fig\|6666666.224586.peg.10535, fig\|6666666.224586.peg.13714 |
| DNA Metabolism | DNA Metabolism - no subcategory | Restriction-Modification System | Type I restriction-modification system, specificity subunit S (EC 3.1.21.3) | fig\|6666666.224586.peg.10536, fig\|6666666.224586.peg.13715 |
| DNA Metabolism | DNA Metabolism - no subcategory | Restriction-Modification System | Type III restriction-modification system methylation subunit (EC 2.1.1.72) | fig\|6666666.224586.peg.6353 |
| DNA Metabolism | DNA Metabolism - no subcategory | Restriction-Modification System | Type I restriction-modification system, restriction subunit R (EC 3.1.21.3) | fig\|6666666.224586.peg.4435, fig\|6666666.224586.peg.7008, fig\|6666666.224586.peg.10537, fig\|6666666.224586.peg.13716, fig\|6666666.224586.peg.13797 |
| DNA Metabolism | DNA Metabolism - no subcategory | Restriction-Modification System | Type I restriction-modification system, DNA-methyltransferase subunit M (EC 2.1.1.72) | fig\|6666666.224586.peg.10535, fig\|6666666.224586.peg.13714 |
| DNA Metabolism | DNA Metabolism - no subcategory | Restriction-Modification System | Putative predicted metal-dependent hydrolase | fig\|6666666.224586.peg.2713, fig\|6666666.224586.peg.2770, fig\|6666666.224586.peg.5710 |
| DNA Metabolism | DNA Metabolism - no subcategory | DNA ligases | ATP-dependent DNA ligase (EC 6.5.1.1) LigC | fig\|6666666.224586.peg.3342 |
| DNA Metabolism | DNA Metabolism - no subcategory | DNA ligases | Ku domain protein | fig\|6666666.224586.peg.3393, fig\|6666666.224586.peg.14435 |

| | | | | |
|---|---|---|---|---|
| DNA Metabolism | DNA Metabolism - no subcategory | DNA ligases | DNA ligase (ATP) (EC 6.5.1.1) | fig\|6666666.224586.peg.9746, fig\|6666666.224586.peg.13390 |
| DNA Metabolism | DNA Metabolism - no subcategory | DNA ligases | ATP-dependent DNA ligase (EC 6.5.1.1) clustered with Ku protein, LigD | fig\|6666666.224586.peg.3392, fig\|6666666.224586.peg.14430 |
| DNA Metabolism | DNA Metabolism - no subcategory | DNA structural proteins, bacterial | Chromosome partition protein smc | fig\|6666666.224586.peg.4304, fig\|6666666.224586.peg.4550, fig\|6666666.224586.peg.11928, fig\|6666666.224586.peg.13020 |
| DNA Metabolism | DNA Metabolism - no subcategory | DNA structural proteins, bacterial | DNA-binding protein HU-alpha | fig\|6666666.224586.peg.10184 |
| DNA Metabolism | DNA Metabolism - no subcategory | DNA structural proteins, bacterial | Integration host factor beta subunit | fig\|6666666.224586.peg.13808, fig\|6666666.224586.peg.14107 |
| DNA Metabolism | DNA Metabolism - no subcategory | DNA structural proteins, bacterial | Integration host factor alpha subunit | fig\|6666666.224586.peg.1063, fig\|6666666.224586.peg.15907 |
| DNA Metabolism | DNA Metabolism - no subcategory | DNA structural proteins, bacterial | DNA-binding protein HU-beta | fig\|6666666.224586.peg.1875, fig\|6666666.224586.peg.10735, fig\|6666666.224586.peg.13558 |
| DNA Metabolism | DNA Metabolism - no subcategory | DNA structural proteins, bacterial | DNA-binding protein HBsu | fig\|6666666.224586.peg.12348 |
| DNA Metabolism | DNA Metabolism - no subcategory | DNA structural proteins, bacterial | DNA-binding protein Fis | fig\|6666666.224586.peg.127, fig\|6666666.224586.peg.5460 |

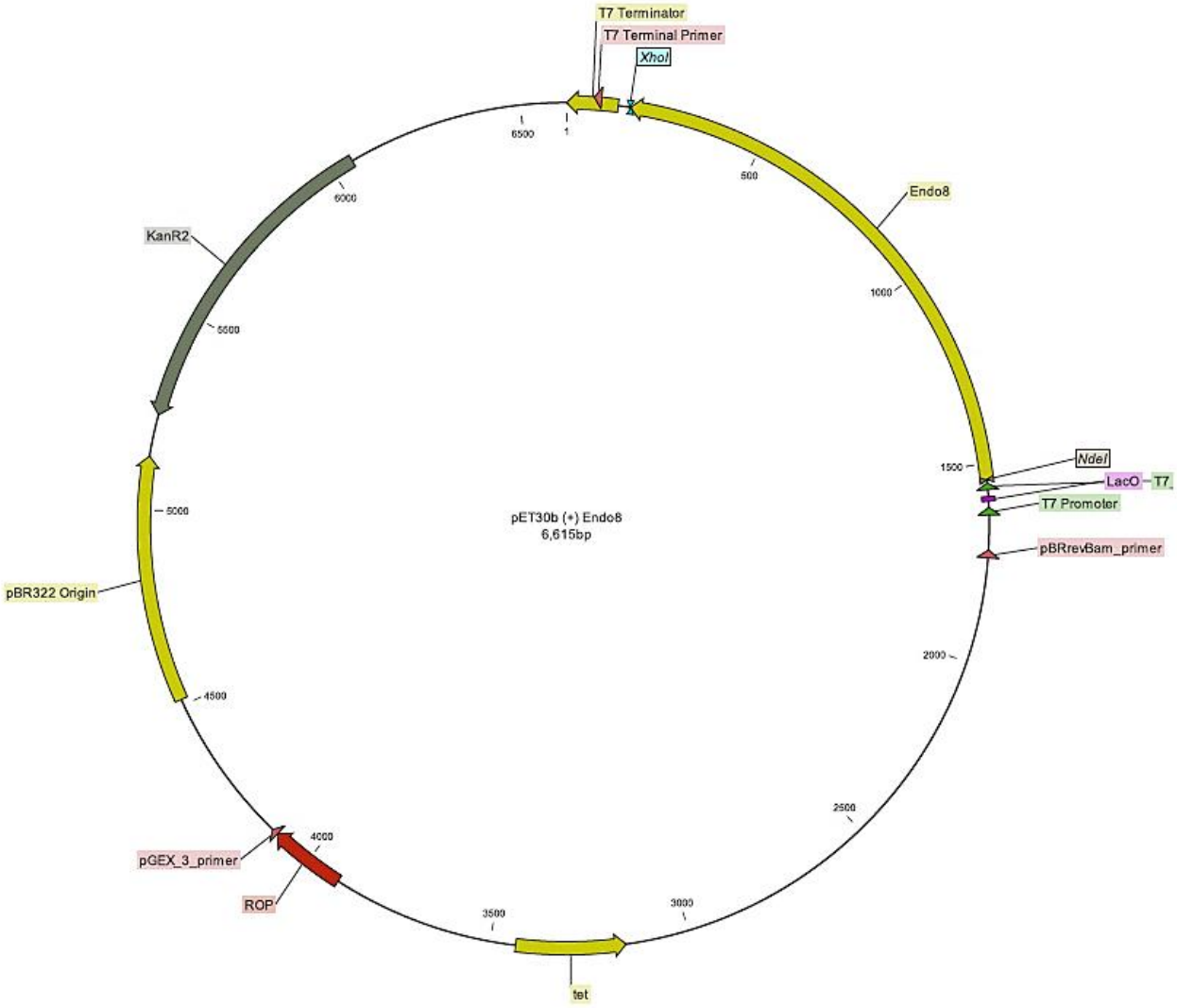



Figure A1: Schematic representation of the pET30b(+) vector map with endo8 ORF. Parts of the vectors backbone have also been included. The restriction sites used to insert the genes are also represented in the vector map. Image was constructed using CLCBio WorkBench Version 11, full suite.

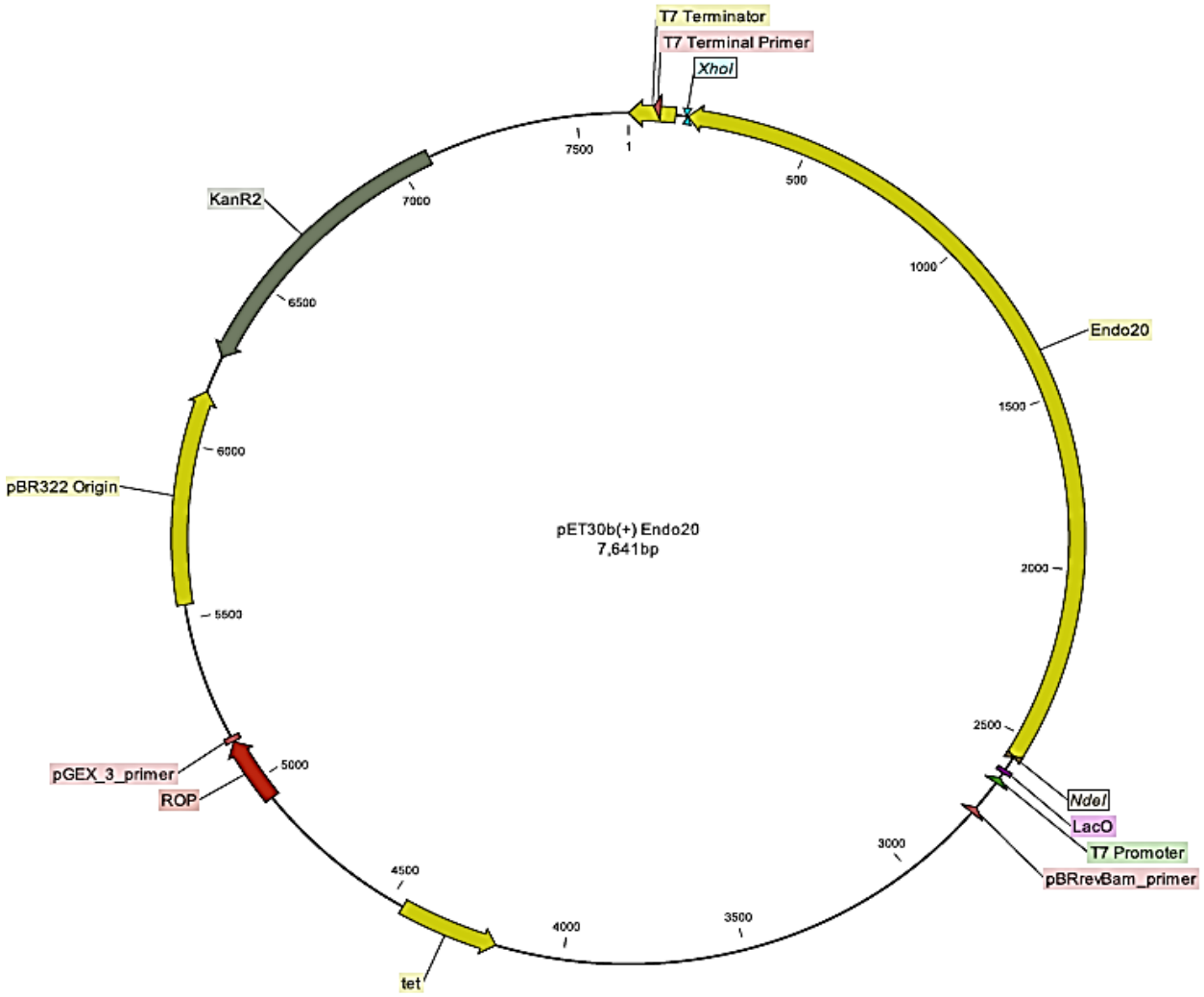



Figure A2: Schematic representation of the pET30b(+) vector map with endo20 ORF. Parts of the vectors backbone have also been included. The restriction sites used to insert the genes are also represented in the vector map. Image was constructed using CLCBio WorkBench Version 11, full suite.

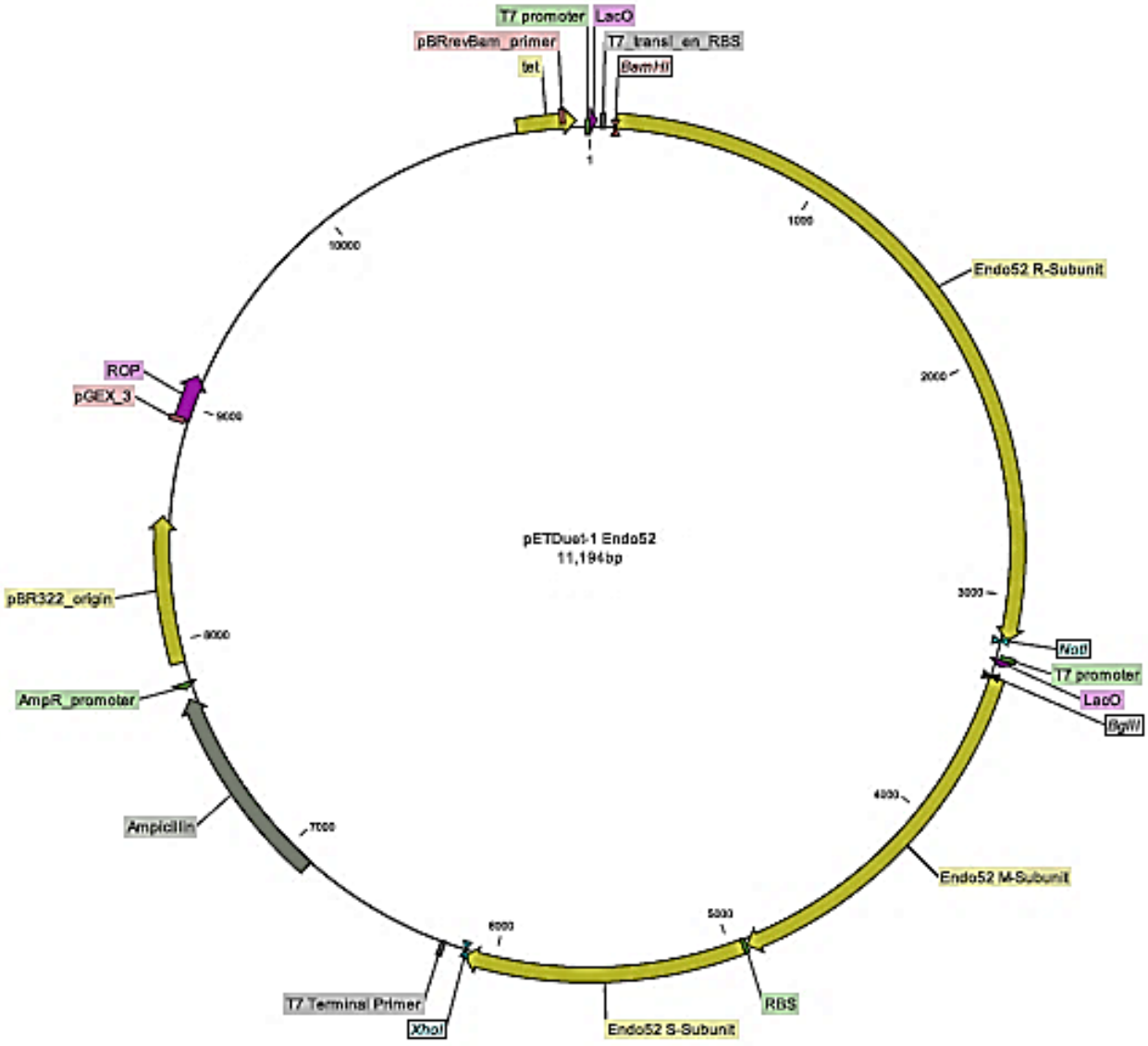



Figure A3: Schematic representation of the pETDuet-1 vector map with endo52 ORF showing the three sub-units. Parts of the vectors backbone have also been included. The restriction sites used to insert the genes are also represented in the vector map. Image was constructed using CLCBio WorkBench Version 11, full suite.

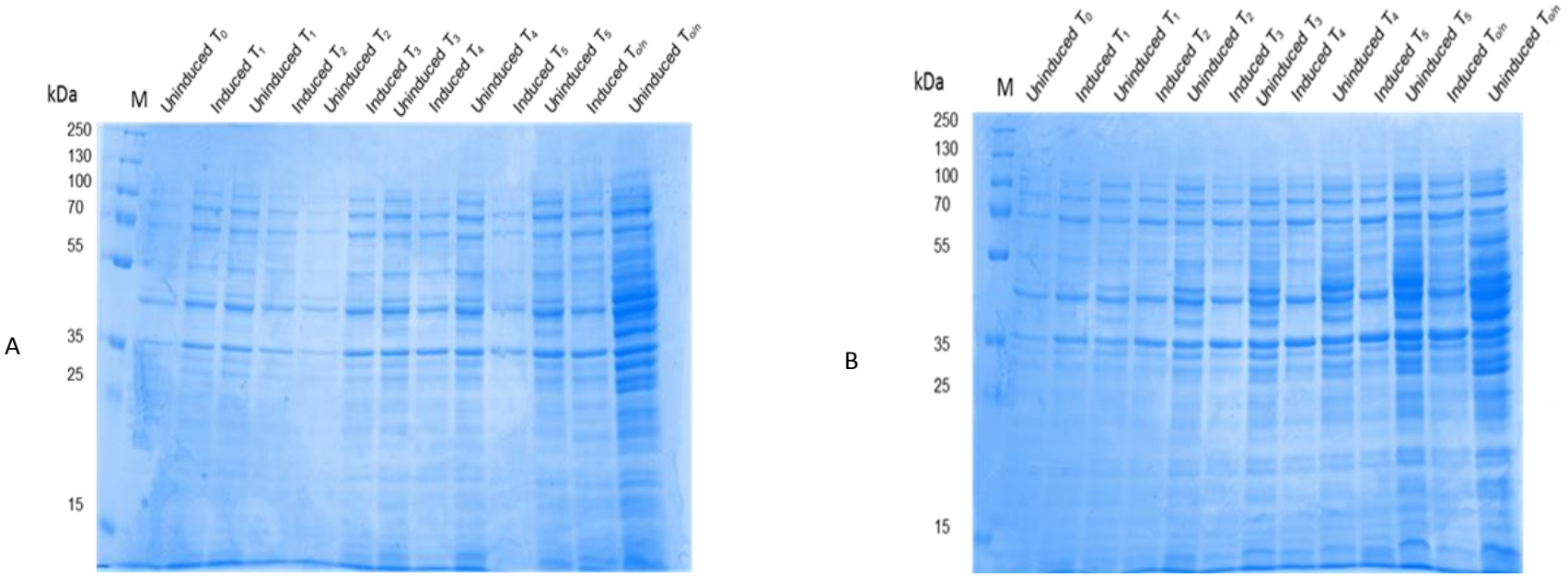



Figure A4: Optimisation of Endo20 using BL-21 as the expression host at three different temperatures (A) 17 and (B) 25 °C. M- Prestained Protein Ladder. $T_0$ - fraction before induction; $T_1$- 1 hour post induction and similarly other T labels refer to time post induction at an hour interval; $T_{o/n}$ - fraction taken the next day post induction.

C

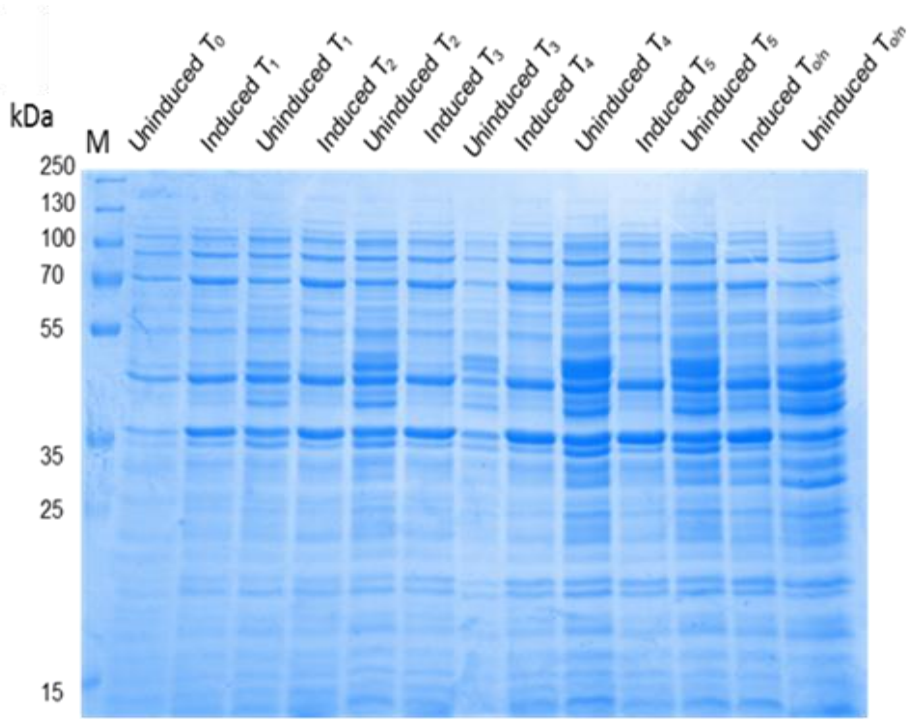



Figure A5: Optimisation of Endo_20 using BL-21 as the expression host at three different temperatures (C) 30 °C. M-Prestained Protein Ladder. $T_0$ - fraction before induction; $T_1$- 1 hour post induction and similarly other T labels refer to time post induction at an hour interval; $T_{o/n}$ - fraction taken the next day post induction.

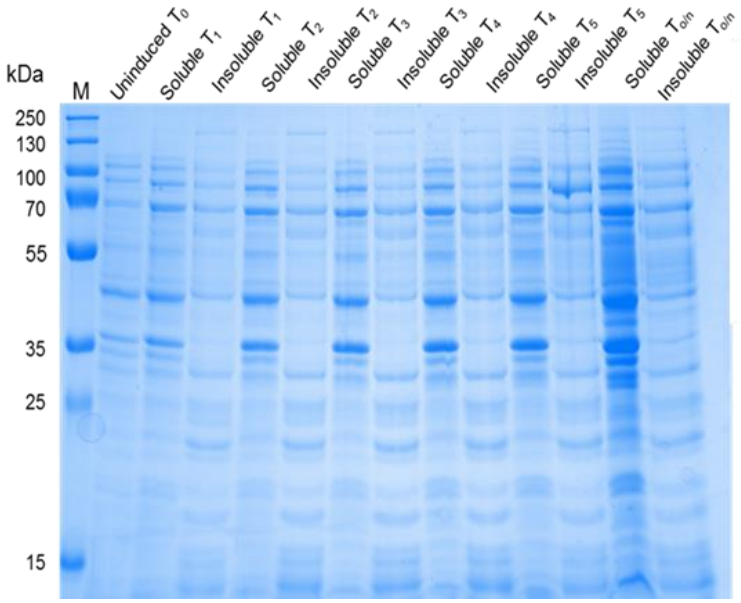



Figure A6: SDS-PAGE analysis of the soluble and insoluble fraction of endo20 using BL21 (DE3) at 25 °C sampled every hour for 5 hr post induction and overnight. M-Prestained Protein Ladder. $T_0$ - fraction before induction; $T_1$- 1 hour post induction and similarly other T labels refer to time post induction at an hour interval; $T_{o/n}$ - fraction taken the next day post induction.

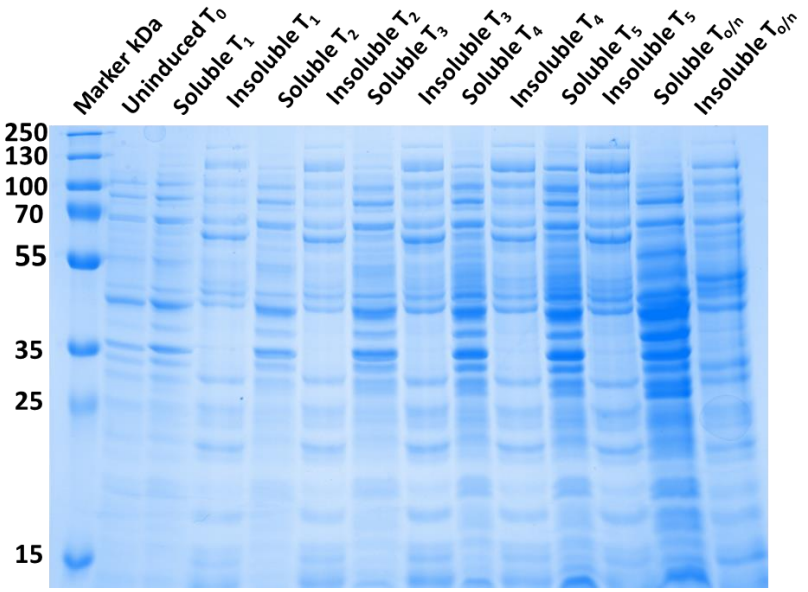



Figure A7: SDS-PAGE analysis of the soluble and insoluble fraction of endo52 using BL21 (DE3) at 25 °C sampled every hour for 5 hr post induction and overnight. M-Prestained Protein Ladder. $T_0$ - fraction before induction; $T_1$- 1 hour post induction and similarly other T labels refer to time post induction at an hour interval; $T_{o/n}$ - fraction taken the next day post induction.

**Table A4: Size determination of protein of interest.**

Size of the proteins of interest was determined using Image Lab 4.1 Software from BioRad.

Endo8

| Marker | | | Lane 4 | Lane 6 | Lane 8 | Lane 10 | Lane 12 | Lane 14 |
|---|---|---|---|---|---|---|---|---|
| Band No. | Band Label | Mol. Wt. (KDa) | Mol. Wt. (KDa) | Mol. Wt. (KDa) | Mol. Wt. (KDa) | Mol. Wt. (KDa) | Mol. Wt. (KDa) | Mol. Wt. (KDa) |
| 1 | 250 | 250 | 53.11 | 52.58 | 52.58 | 53.66 | 54.76 | 55.32 |
| 2 | 130 | 130 | | | | | | |
| 3 | 100 | 100 | | | | | | |
| 4 | 70 | 70 | | | | | | |
| 5 | 55 | 55 | | | | | | |
| 6 | 35 | 35 | | | | | | |
| 7 | 25 | 25 | | | | | | |
| 8 | 15 | 15 | | | | | | |
| 9 | 10 | 10 | | | | | | |

Endo20

| Lane 1 | | | Lane 8 | Lane 10 |
|---|---|---|---|---|
| Band No. | Band Label | Mol. Wt. (KDa) | Mol. Wt. (KDa) | Mol. Wt. (KDa) |
| 1 | 250 | 250 | 81.66 | 86.11 |
| 2 | 130 | 130 | | |
| 3 | 100 | 100 | | |
| 4 | 70 | 70 | | |
| 5 | 55 | 55 | | |
| 6 | 35 | 35 | | |
| 7 | 25 | 25 | | |
| 8 | 15 | 15 | | |

| | | | | |
|---|---|---|---|---|
| 9 | 10 | 10 | | |

| Endo52 | | | | | | | |
|---|---|---|---|---|---|---|---|
| Lane 1 | | | Lane 6 | Lane 8 | Lane 10 | Lane 12 | Lane 14 |
| Band No. | Band Label | Mol. Wt. (KDa) | Mol. Wt. (KDa) | Mol. Wt. (KDa) | Mol. Wt. (KDa) | Mol. Wt. (KDa) | Mol. Wt. (KDa) |
| 1 | 250 | 250 | 122.78 | 123.76 | 123.76 | 124.75 | 125.75 |
| 2 | 130 | 130 | | | | | |
| 3 | 100 | 100 | | | | | |
| 4 | 70 | 70 | | | | | |
| 5 | 55 | 55 | | | | | |
| 6 | 35 | 35 | | | | | |
| 7 | 25 | 25 | | | | | |
| 8 | 15 | 15 | | | | | |

**Table A5: Endo8's purity table using the quantified fraction**.

Purified Endo8 was used to determine its purity. Image Lab 4.1 software was used to determine the purity. Fractions from the purified were used and a 1:10 was used before loading the samples on the gel. Lane 1-3, 2.5, 5 and 10 µl were loaded of endo8 with sample buffer. Lane4-9 is the BSA and 10µl were added in each lane.

| Endo8 (1:10 dilution) | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Lane 1 | | Lane 2 | | Lane 3 | | Lane 4 | | Lane 5 | | Lane 6 | | Lane 7 | | Lane 8 | | Lane 9 | |
| Band No. | Band % | Band No. | Band % | Band No. | Band % | Band No. | Band % | Band No. | Band % | Band No. | Band % | Band No. | Band % | Band No. | Band % | Band No. | Band % |
| 1 | 92.79 | 1 | 95.10 | 1 | 0.21 | 1 | 100 | 1 | 100 | 1 | 100 | 1 | 100 | 1 | 100 | 1 | 100 |
| 2 | 1.83 | 2 | 1.57 | 2 | 1.09 | | | | | | | | | | | | |
| 3 | 5.38 | 3 | 3.33 | 3 | 88.20 | | | | | | | | | | | | |
| | | | | 4 | 4.00 | | | | | | | | | | | | |
| | | | | 5 | 6.50 | | | | | | | | | | | | |

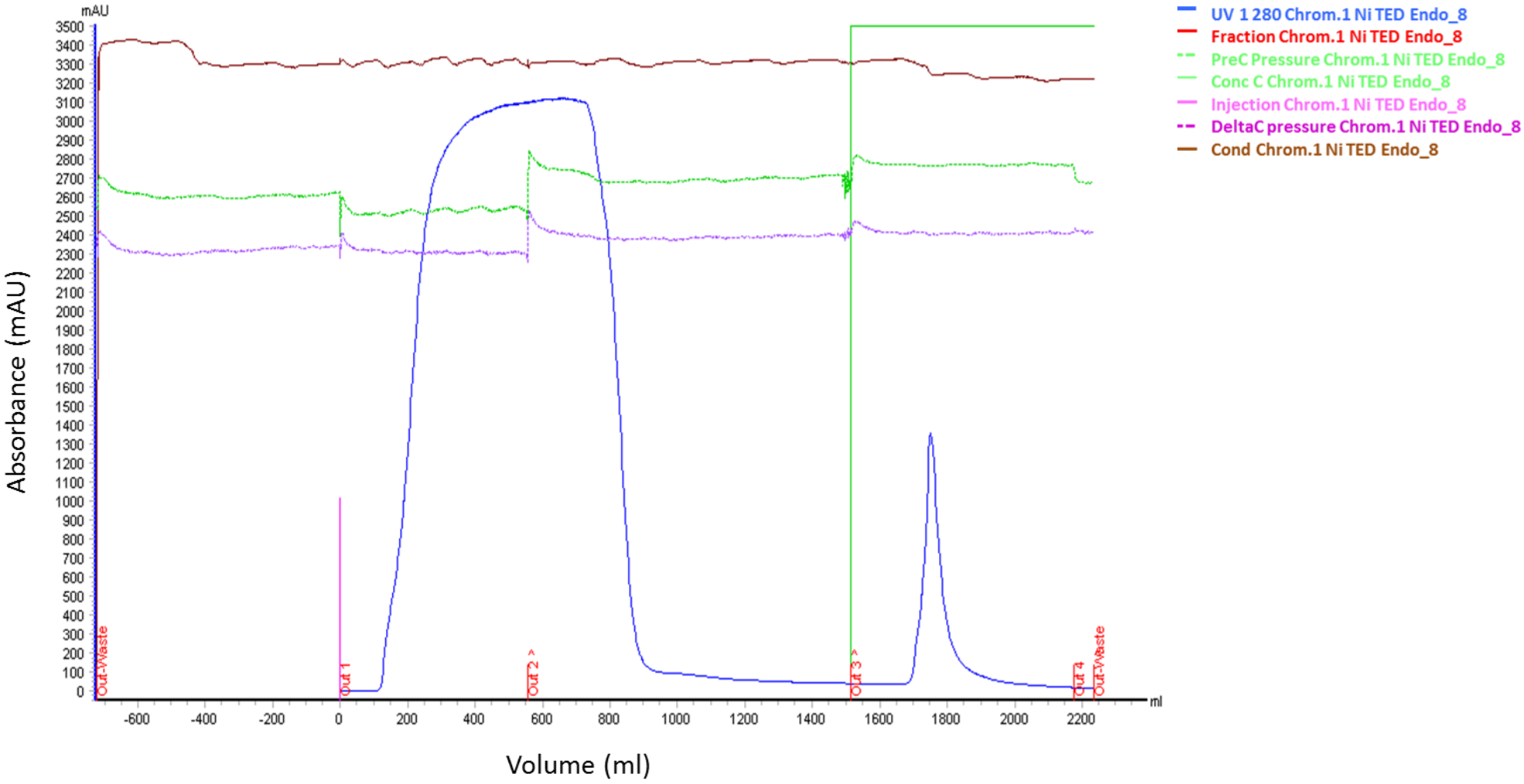



Figure A8: AKTA purification of endo_8, 8 hr post induction using HiScale50 Ni-TED ion-exchange chromatogram using absorbance monitoring and peak collection through outlet 3. Ni-TED buffer exchange chromatogram using conductivity monitoring and peak collection through outlet 2. Protein elution was monitored by absorbance (in milliabsorbance units) at 280 nm (blue), with the conductivity (brown) or pressure in the chambers within the system (purple and green, dotted) overlaid on the chromatograms.Soild green line appearing from out3 is the elution buffer, eluted from 100%,