

Development of a high throughput cell-free metagenomic screening platform

By

Walter Nevondo

A thesis submitted in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY



Department of Biotechnology

University of the Western Cape

UNIVERSITY *of the*
WESTERN CAPE
Bellville

Supervisor: Professor Marla Trindade

October 2016

Keywords

Metagenomics

Lignocellulase

Beta-xylosidase

Cell-free protein synthesis

FACS

In vitro compartmentalisation

Double emulsion

Next generation sequencing

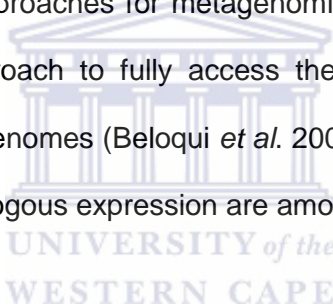
Enzyme screening

R. erythropolis



Abstract

The estimated 5×10^{30} prokaryotic cells inhabiting our planet sequester some 350–550 Petagrams (1 Pg = 10^{15} g) of carbon, 85–130 Pg of nitrogen, and 9–14 Pg of phosphorous, making them the largest reservoir of those nutrients on Earth (Whitman *et al.* 1998). However, reports suggest that only less than 1% of these microscopic organisms are cultivable (Torsvik *et al.* 1990; Sleator *et al.* 2008). Until recently with the development of metagenomic techniques, the knowledge of microbial diversity and their metabolic capabilities has been limited to this small fraction of cultivable organisms (Handelsman *et al.* 1998). While metagenomics has undoubtedly revolutionised the field of microbiology and biotechnology it has been generally acknowledged that the current approaches for metagenomic bio-prospecting / screening have limitations which hinder this approach to fully access the metabolic potentials and genetic variations contained in microbial genomes (Beloqui *et al.* 2008). In particular, the construction of metagenomic libraries and heterologous expression are amongst the major obstacles.



The aim of this study was to develop an ultra-high throughput approach for screening enzyme activities using uncloned metagenomic DNA, thereby eliminating cloning steps, and employing *in vitro* heterologous expression. To achieve this, three widely used techniques: cell-free transcription-translation, *in vitro* compartmentalisation (IVC) and Fluorescence Activated Cell Sorting (FACS) were combined to develop this robust technique called metagenomic *in vitro* compartmentalisation (mIVC-FACS). Moreover, the *E. coli* commercial cell-free system was used in parallel to a novel, in-house *Rhodococcus erythropolis* based cell-free system. The versatility of this technique was tested by identifying novel beta-xylosidase encoding genes derived from a thermophilic compost metagenome. In addition, the efficiency of mIVC-FACS was compared to the traditional metagenomic approaches; function-based (clone library screening) and sequence-based (shotgun sequencing and PCR screening).

The results obtained here show that the *R. erythropolis* cell-free system was over thirty-fold more effective than the *E. coli* based system based on the number of hits obtained per million double emulsions (dE) droplets screened. Six beta-xylosidase encoding genes were isolated and confirmed from twenty-eight positive dE droplets. Most of the droplets that were isolated from the same gate encoded the same enzyme, indicating that this technique is highly selective. A comparison of the hit rate of this screening approach with the traditional *E. coli* based fosmid library method shows that mIVC-FACS is at least 2.5 times more sensitive. Although only a few hits from the mIVC-FACS screening were selected for confirmation of beta-xylosidase activity, the proposed hit rate suggests that a significant number of positive hits are left un-accessed through the traditional clone library screening system. In addition, these results also suggest that *E. coli* expression system might be intrinsically sub-optimal for screening for hemicellulases from environmental genomes compared to *R. erythropolis* system. The workflow required for screening one million clones in a fosmid library was estimated to be about 320 hours compared to 144 hours required via the mIVC-FACS screening platform. Some of the gene products obtained in both screening platforms show multiple substrate activities, suggesting that the microbial consortia of composting material consist of microorganisms that produce enzymes with multiple lignocellulytic activities.

While this platform still requires optimisation, we have demonstrated that this technique can be used to isolate genes encoding enzymes from mixed microbial genomes. mIVC-FACS is a promising technology with the potential to take metagenomic studies to the second generation of novel natural products bio-prospecting. The astonishing sensitivity and ultra-high throughput capacity of this technology offer numerous advantages in metagenomic bio-prospecting.

Declaration

I declare that *Development of a high throughput metagenomic screening platform* is my own original work, that it has not been submitted for any degree or examination in any other university, and that all the sources I have used or quoted have been indicated and acknowledged by complete references.



Acknowledgements

I am especially grateful to PROF. MARLA TRINDADE for affording me the opportunity to undertake this study. Your thoughtful guidance has contributed immeasurably to the completion of this project and development of my career.

I owe much gratitude to Mr. LONNIE VAN ZYL and DR HEIDE GOODMAN. Your unwavering help and technical guidance contributed significantly to the success of this project.

Special gratitude to PROF ADRIENN EDKINS for kindly allowing me to use instruments in her laboratory.

Special thanks to MASTER ORGANICS for kind donation of the compost samples and THE NATIONAL RESEARCH FOUNDATION (NRF) for financial support.

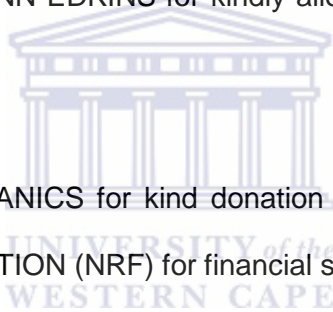


Table of Contents

Chapter One: Literature Review	5
1. Introduction to Metagenomics	5
1.1. Sequence-based bio-prospecting approaches	10
1.1.1. PCR-based methods.....	10
1.1.2. NGS-based methods	12
1.2. Function-based metagenomic screening.....	17
1.2.1. Cloning vectors	18
1.2.2. Heterologous hosts.....	19
1.2.3. Screening approaches	20
1.2.4. Challenges of function-based metagenomics.....	23
1.2.4.1. Promoter and Codon recognition	23
1.2.4.2. Instability of biological molecules	24
1.2.4.3. Insoluble aggregates and protein folding.....	25
1.2.5. Next generation metagenomic bio-prospecting	25
1.2.5.1. Cell-free protein synthesis	26
1.2.5.2. <i>In vitro</i> compartmentalisation (IVC).....	28
1.2.5.3. Fluorescence Activated Cell Sorting (FACS).....	30
1.3. Metagenomic as a tool for novel lignocellulases discovery.....	31
1.3.1. Hemicellulose	32
1.3.1.1. Hydrolysis of hemicellulose.....	33
1.3.1.2. Beta-xylosidases.....	34
1.3.1.2.1. Biochemical properties of beta xylosidases	35
1.3.1.2.2. Classification of beta-xylosidases	37
1.4. Aims and objectives of the study.....	38

CHAPTER TWO: Phylogenetic analysis of thermophilic horse manure compost metagenome (HMM), assessment of glycoside hydrolases (GH) and PCR-based bio-prospecting of beta-xylosidases.

2. Introduction	39
2.1. Materials and methods.....	41
2.1.1. Samples collection.....	41
2.1.2. Metagenomic DNA extraction	41

2.1.3.	Metagenomic DNA purification.....	42
2.1.4.	Sequencing of metagenomic DNA and reads assembly.....	43
2.1.5.	Analysis of microbial diversity and sample coverage.....	43
2.1.6.	Identification of glycoside hydrolases (GHs) and beta-xylosidase reads	44
2.1.7.	Calculation of abundance index (AI)	44
2.1.8.	Primers used in this chapter.....	44
2.1.9.	PCR amplifications.....	47
2.1.10.	Cloning of PCR products	47
2.1.11.	Preparation of electro-competent cells.....	48
2.1.12.	Electroporation.....	48
2.2.	Results and discussions.....	49
2.2.1.	Metagenomic DNA extraction	49
2.2.2.	Next generation sequencing of HMM	49
2.2.3.	Microbial diversity of horse manure metagenome (HMM)	52
2.2.4.	Abundance of Glycoside hydrolases (GHs).....	55
2.2.5.	Abundance of beta-xylosidases	57
2.2.5.1.	Beta-xylosidase ORFs	59
2.2.6.	PCR screening for beta-xylosidase encoding genes	61
2.2.6.1.	Degenerate primer design	61
2.2.6.2.	PCR screening based for beta-xylosidases from HMM	62
2.3.	Conclusion.....	65
CHAPTER THREE: Traditional functional screening for beta-xylosidases from a horse manure metagenomic fosmid library.		67
3.	Introduction.....	67
3.1.	Materials and methods.....	69
3.1.1.	Metagenomic library construction.....	69
3.1.2.	Evaluation of insert size of the metagenomic library.....	70
3.1.3.	High-throughput library picking and screening	70
3.1.4.	Secondary screening of positive clones	70
3.1.5.	Fosmids extraction, sequencing and analysis	71
3.2.	Results and discussions.....	71
3.2.1.	Verification of metagenomic library	72
3.2.2.	Screening for beta-xylosidase activity	72

3.2.3.	Sequence analysis of positive fosmid showing beta-xylosidase activity.....	75
3.2.3.1.	Clone P17C1	76
3.2.3.1.1.	Analysis of conserved residues of ORFs 18 and 19.....	79
3.2.3.2.	Clone P15A9	80
3.2.3.2.1.	Analysis of conserved residues of ORF P15A9-7	83
3.2.3.3.	Clone P118H6	84
3.2.3.3.1.	Analysis of conserved residues of ORF P118H6-1	86
3.2.3.4.	Clone P117H11	88
3.2.3.4.1.	Analysis of conserved residues of ORF P117H11-13	90
3.3.	Conclusion.....	92
CHAPTER FOUR: Development of a cell-free metagenomic screening platform using <i>E. coli</i>		
S30 CFPS system as transcription-translation machinery.....		
4.	Introduction.....	95
4.1.	Materials and methods.....	97
4.1.1.	Bacterial strains and plasmids	97
4.1.2.	Development of double emulsion (dE)	97
4.1.3.	Activity screening using a fluorescence activated cell sorter (FACS)	98
4.1.4.	Sequencing and Identification of beta-xylosidases.....	99
4.1.5.	Amplification of <i>mbgIX</i>	99
4.1.6.	Cloning of PCR products.....	99
4.1.7.	Preparation of electro-competent cells.....	100
4.1.8.	Electroporation.....	100
4.1.9.	Plasmid isolation.....	101
4.1.10.	Plasmid digestion and gel electrophoresis	101
4.1.11.	Sub-cloning	102
4.1.12.	<i>In vivo</i> expressions.....	102
4.1.13.	Enzyme essays	103
4.2.	Results and discussions.....	103
4.2.1.	Development of double emulsions	103
4.2.2.	IVC-FACS screening of beta-xylosidases activity.....	105
4.2.3.	DNA recovery from dE droplets and amplification	108
4.2.4.	Sequencing and reads assembly	108
4.2.5.	Identification of sorted genotype	113

4.2.6.	MBglX confers beta-xylosidase activity	114
4.3.7.	Catalytic amino acids residues and active sites of MBglX	116
4.3.	Conclusion	118
CHAPTER FIVE: mIVC-FACS screening of beta-xylosidase activity using a <i>Rhodococcus erythropolis</i> CFPS system.....		
		121
5.	Introduction	121
5.1.	Material and methods	122
5.1.1.	Bacterial strains and plasmids	122
5.1.2.	Actinobacteria cell-free protein synthesis (CFPS) system	123
5.1.3.	Validation of the <i>R. erythropolis</i> CFPS system.....	123
5.1.4.	Development of double emulsion and IVC-FACS screening	124
5.1.5.	PCR amplifications and cloning of genes isolated through IVC-FACS	124
5.1.6.	Development of pTip-RC4.....	125
5.1.7.	Preparation of <i>R. erythropolis</i> electro-competent cells	126
5.1.8.	<i>In vivo</i> expressions	126
5.1.9.	DNA recovery, sequencing and sequences analysis.....	127
5.2.	Results and discussions.....	127
5.2.1.	Development of actinobacteria based CFE	127
5.2.2.	IVC-FACS using the actinobacteria based CFPS system	128
5.2.3.	DNA recovery and sequencing	131
5.2.4.	Identification of possible beta-xylosidase genes.....	133
5.2.5.	Cloning of five ORF's	134
5.2.6.	<i>In vivo</i> expression in <i>E. coli</i> and <i>R. erythropolis</i>	136
5.2.7.	MBglX cannot be expressed by the <i>R. erythropolis</i> system.....	139
5.2.8.	Analysis of genes isolated through the <i>R. erythropolis</i> based CFPS system.....	141
5.2.8.1.	Analysis of XylA and XylB	141
5.2.8.2.	BgAX has the general GH3 catalytic motifs.....	144
5.2.8.3.	XylT is a transferase/beta-xylosidase.....	145
5.2.8.4.	Catalytic residues of BgCX	148
5.3.	Conclusion	150
CHAPTER SIX: GENERAL DISCUSSION AND CONCLUSION		153
REFERENCES		163
APPENDICES.....		191

List of Figures

FIGURE 1: General process in metagenomic gene discovery	8
FIGURE 2: General representation of NGS based metagenomic studies.....	13
FIGURE 3: Summary of DNA sequencing process.	17
FIGURE 4: Representation of function-based metagenomics.....	18
FIGURE 5: Schematic representation of p18GFP expression system.	21
FIGURE 6: Representation of METREX screening system.....	21
FIGURE 7: Schematic representation of the PIGEX.	22
FIGURE 8: General representation of cell-free protein synthesis.	27
FIGURE 9: General representation of <i>in vitro</i> compartmentalisation.	29
FIGURE 10: Overview of IVC system.	30
FIGURE 11: Screening of enzyme activity through fluorescence activated cell sorter.	31
FIGURE 12: Structure of hemicellulose.	33
FIGURE 13: Schematic representation of biodegradation of hemicellulose backbone	34
FIGURE 14: Agarose gel electrophoresis of extracted metagenomic DNA.	49
FIGURE 15: Breakdown of metagenomics sequence data.	50
FIGURE 16: Nonpareil curve of sequenced HMM sample.	52
FIGURE 17: Microbial diversity of horse manure metagenome.	52
FIGURE 18: Diversity of bacterial orders in horse manure metagenome.	54
FIGURE 19: Abundance of GHs in HMM.....	56
FIGURE 20: Beta-xylosidase reads obtained from HMM.	58
FIGURE 21: Agarose gel electrophoresis image showing low stringency PCR amplicons.....	62
FIGURE 22: Incidence rate of the positive according to the enzyme activity.	68
FIGURE 23: Restriction digest of 20 randomly selected fosmid clones.	72
FIGURE 24: Activity analysis of positive fosmid clones.....	74
FIGURE 25: Restriction analysis of positive hits.	75
FIGURE 26: Gene organisation of contig P17C1.	78
FIGURE 27: Phylogenetic and catalytic analysis of ORF P17C1-18 and P17C1-19.	80
FIGURE 28: Gene organisation of contig P15A9.1 (A) and contig P15A9.2 (B)	82
FIGURE 29: Phylogenetic and catalytic analysis of ORF P15A9-7.	83
FIGURE 30: Gene organisation of contig P118H6.1 (A) and contig P118H6.2 (B).....	85
FIGURE 31: Phylogenetic and catalytic analysis of ORF P118H6-1.	87
FIGURE 32: Gene organisation of contig P117F11.1 (A) and contig P117F11.2 (B).	89
FIGURE 33: Phylogenetic and catalytic analysis of ORF P117H11-13.	91
FIGURE 34: General overview of a novel mIVC-FACS.....	96
FIGURE 35: Number of events observed vs.storage time.....	104

FIGURE 36: FACS recording produced during beta-xylosidase activity screening	107
FIGURE 37: Image of agarose gel electrophoresis of DNA recovered and MDA-amplified	108
FIGURE 38: Agarose gel image of double digested pET21a harbouring <i>mbglX</i>	115
FIGURE 39: Activity obtained from crude extracts expressing MBglX,	116
FIGURE 40: Multiple sequence alignments of MBglX with four GH3 beta-xylosidases..	117
FIGURE 41: MUX hydrolysis using a novel <i>R. erythropolis</i> CFPS system.....	128
FIGURE 42: FACS recording produced during beta-xylosidase activity screening.....	130
FIGURE 43: Agarose gel electrophoresis images showing pTip-RC4.	135
FIGURE 44: Agarose gel image of double digested pTIP-RC4.....	136
FIGURE 45: Bar charts showing increase in fluorescence intensity hydrolysis..	137
FIGURE 46: Fluorescence intensity as a result of MUX hydrolysis.....	137
FIGURE 47: Agarose gel image of double digested pTIP-RC4 harbouring <i>mbglX</i>	140
FIGURE 48: <i>In vivo</i> expression of <i>bglX</i> in <i>E. coli</i> and <i>R. erythropolis</i>	141
FIGURE 49: Sequence alignments of XylA and XylB with other GH43 proteins.	143
FIGURE 50: Sequence alignment of BgAX with other GH3 proteins..	144
FIGURE 51: Domain architecture of XylT.....	145
FIGURE 52: Multiple sequence alignment of XylT.	147
FIGURE 53: Alignments of catalytic residues of GH5 enzymes with BgCX.	149
FIGURE 54: General overview of this study.....	154
FIGURE 55: Comparison of hit rate.	156
FIGURE 56: Flow chart illustrating the identification of highly active novel genes	158
FIGURE 57: Comparison of mIVC-FACS and clone library screening systems.....	160
FIGURE 58: Beta-xylosidase encoding which were identified in this study.....	161

List of Tables

Table 1: Culturability of bacteria	5
Table 2: Enzymes discovered from extreme environments	7
Table 3: Comparison of different generations of sequencing platforms	15
Table 4: Properties of some of the characterised fungal beta-xylosidases	36
Table 5: Properties of some of the characterised bacterial beta-xylosidases	37
Table 6: Degenerate primers used in this study	46
Table 7: Summary of MiSeq data	51
Table 8: Predicted beta-xylosidases encoding ORFs from HMM	60
Table 9: Motif positions in the aligned sequences	61
Table 10: NCBI blastx results of sequences obtained from PCR-based screening	63
Table 11: Summary of ORFs identified in this chapter	93
Table 12: Bacterial strains and plasmids used in this chapter	97
Table 13: Amount sequence data generated per dE droplets	110
Table 14: Bacterial strains and plasmids used in this chapter	122
Table 15: Primers used in this chapter	125
Table 16: Amount of DNA obtained per dE droplet	132
Table 17: GH hits identified from NCBI	133
Table 18: Blastx results of genes contained in P15A9	192
Table 19: Blastx results of genes contained in P118H6	193
Table 20: Blastx results of genes contained in P117F11	194

List of abbreviation

°C	Degrees Celcius
Amp	Ampicillin
Amp ^R	Ampicillin resistance gene
Bp	Base pairs
BLAST	Basic local alignment search tool
BSA	Bovine serum albumin
CAM	Chloramphenicol
cAMP	cyclic Adenosine monophosphate
CAZy	Carbohydrate-Active Enzymes database
CTAB	Cetyl trimethylammonium bromide
CFPS	Cell-free protein synthesis
C-terminus	Carboxy terminus
ddH ₂ O	Distilled deionised water
dE	Double emulsion
dH ₂ O	Deionised water
DNA	Deoxyribonucleic acid
dNTPs	Deoxyribonucleotide triphosphates
DTT	Dithiothreitol
EDTA	Ethylenediamine tetraacetic acid
EtOH	Ethanol
FACS	Fluorescence activated cell sorter
g	Gram
<i>g</i>	Gravitational force
gDNA	Genomic DNA

GH	Glycoside hydrolase
h	Hour
HEPES	2-[4-(2-hydroxyethyl)piperazin-1-yl]ethanesulfonic acid
HPLC	High-performance liquid chromatography
IPTG	Isopropyl- β -D-thiogalactopyranoside
IVC	<i>In vitro</i> compartmentalisation
K ₂ HPO ₄	Dipotassium hydrogen phosphate
kb	Kilo bases
kDa	Kilo daltons
<i>k</i> cat	Catalytic turnover
<i>K</i> M	Michaelis-Menten constant
LB	Luria broth
LB-amp	Luria broth with Ampicillin
L	Litre
LMP	Low melting point
MgCl ₂	Magnesium chloride
mIVC	Metagenomic <i>in vitro</i> compartmentalisation
μ g	Microgram
μ L	Microliter
μ M	Micromolar
mg	Milligram
mL	Milliliter
mm	Millimeter
mM	Millimolar
M	Molar
min	Minute

MW	Molecular weight
NaCl	Sodium chloride
NADH	Nicotinamide adenine dinucleotide
NCBI	National Centre of Biotechnological Information
ng	Nanogram
nm	Nanometer
N-terminus	Amino-terminus
OD ₆₀₀	Optical density at 600 nm
ORF	Open reading frame
PAGE	Polyacrylamide gel electrophoresis
PCR	Polymerase chain reaction
pfu	Plaque forming unit
PVPP	Polyvinyl polypyrrolidone
RNA	Ribonucleic acid
rpm	Revolutions per minute
rRNA	Ribosomal ribonucleic acid
s	Second
SDS	Sodium dodecyl sulfate
sE	Single emulsion
sp	Species
TAE	Tris acetate ethylenediaminetetraacetic acid
TEMED	N,N,N',N-tetramethylethylenediamine
Tris	2-amino-2-(hydroxymethyl)-1,3-propanediol
tRNA	Transfer ribonucleic acid
U	Enzyme units
UV	Ultraviolet

V	Volts
V_i	Initial velocity
V_{max}	Maximum velocity
vol	Volume
v/v	Volume per volume
w/v	Weight per volume



Chapter One: Literature Review

1. Introduction to Metagenomics

Microorganisms are essential for all life on Earth. Every life process in the biosphere depends on the capacity of microorganisms to transform the world around them. The conversions of key elements of life such as carbon, nitrogen, oxygen, and sulfur, into biologically accessible forms are largely directed by and dependent on microbes (Gadd 2010). Through fermentation and other natural processes, microorganisms create or add value to many foods that are staples of the human diet (Bourdichon *et al.* 2012). However, for decades, the knowledge of microorganisms has been limited to just less than 1% of those that can be cultured in standard laboratory conditions (Ekkers *et al.* 2012). Amann *et al.* (1995) summarised the culturability of bacteria from different environments (Table 1). In their summary, only 0.3% of soil bacteria have been cultured under standard laboratory conditions while less than 0.1% of seawater bacteria are cultured.

Table 1: Culturability of bacteria sampled from different environments

Habitat	Culturability (%)
Seawater	0.001-0.1
Freshwater	0.25
Mesotrophic lake	0.1-1
Unpolluted estuarine water	0.1-3
Activated sludge	1-15
Sediments	0.25
Soil	0.3

Modified from Amann *et al.* (1995)

Different approaches have been developed to improve culturability of these microorganisms, including simulation of natural environments in the laboratory (Kaeberlein *et al.* 2002) and co-culture of different microorganisms (Tanaka 2004). Recently, Tanaka *et al.* (2014) demonstrated that when phosphate is autoclaved together with agar, total colony counts lower remarkably compared with those grown on agar plates in which phosphate and agar were autoclaved separately and mixed immediately before solidification. This finding suggests that there is a need to re-evaluate what has been traditionally called standard microbial culture procedure. Indeed, the problem of microbial “culturability” is more complex than the toxicity of autoclaved phosphate/agar mixture.

Just over a decade ago, Handelsman *et al.* (1998) introduced metagenomics, a culture-independent approach for analyzing mixed microbial genomes directly from the environment. This approach has become a new benchmark for exploring the inaccessible metabolic potential of microorganisms and understanding their genetic diversity. Significant progress has been made through metagenomics (Felczykowska *et al.* 2012). The advancement of next-generation sequencing technologies results in the generation of a large amount of sequence data derived from various environments. These data have revealed enormous phylogenetic and metabolic diversity of microbial communities living in a variety of ecosystems (Morozova & Marra 2008). Some achievements of metagenomics include the discovery of novel antibiotics (Singh & Macdonald 2010; Craig *et al.* 2010), antitumor compounds (Piel 2002), and numerous industrial enzymes from extreme environments (Table 2).

Table 2: Enzymes discovered from extreme environments through metagenomics studies.

Enzyme	Source
Esterase	Seahore sediments Hydrothermal fields Intertidal zone Tidal flat sediment Red sea brine pool
Lipase	Deep sea
Alkaline phosphate	Tidal flat sediments
Glycoside hydrolase	Baltic sea
Phospholipase	Hot spring
Funarase	Marine water
B-glucosidase	Hydrothermal spring
Laccase	Marine water
Mercury reductase	Red sea brine pool

Adapted from Barone *et al.* (2014)

The process of metagenomics gene discovery is a complex multi-step procedure that involves sample collection, extraction of environmental DNA, and screening of desired gene or gene products through either sequence-based (Section 1.1.2-1.1.3) or function-based methods (Section 1.1.4) (Figure 1). Irrespective of the approach employed, there are critical aspects that need to be considered. These pertain to the choice of sampling environment and collection of samples that properly represent the environment. The choice of environment is mainly determined by the general objective of the study. For instance, to identify catabolic enzymes which digest dietary fibers, Tasse *et al.* 2010 sampled the human gut microbial community. Thermostable enzymes have been screened for from samples collected in high-temperature environments (Kang *et al.* 2011).

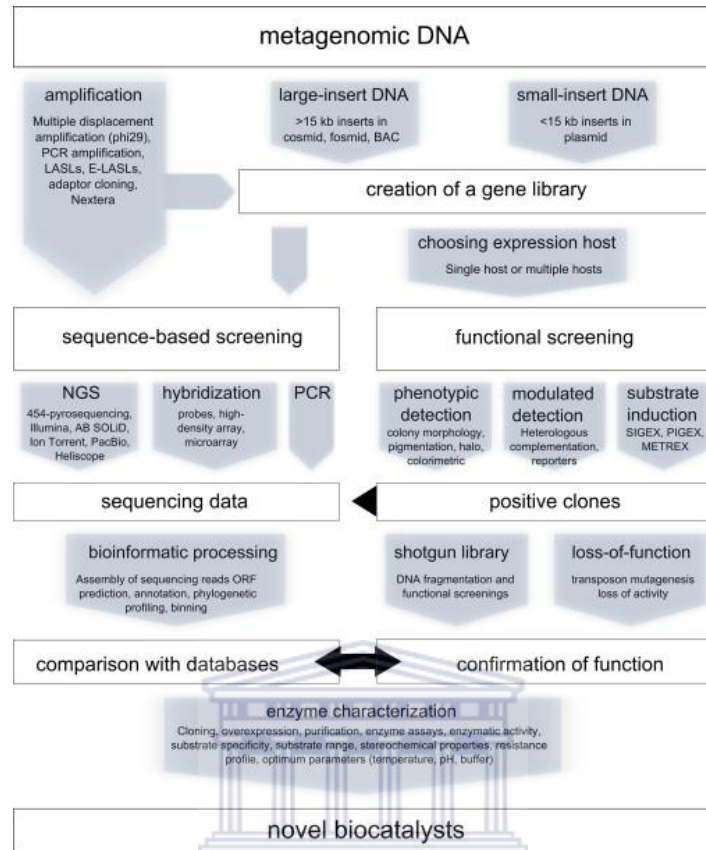


FIGURE 1: General process in metagenomic gene discovery and microbial communities' studies. Samples are collected from a variety of environments followed by isolation of mixed microbial genomes (metagenomic DNA). Bio-prospecting of isolated metagenomic DNA is done through sequence or function-based approach. In the function-based approach, large insert or small insert libraries are constructed followed by functional screening. Sequence based approaches can be PCR-based or direct NGS based. Adapted from Leis *et al.* (2013).

Metagenomic DNA is extracted either through direct or indirect methods (Courtois *et al.* 2001). The indirect method involves separation of microbial cells from soil and other particles followed by extraction of metagenomic DNA from separated bacterial cells. The major advantage of the indirect method is that it allows the extraction of high molecular weight metagenomic DNA (Courtois *et al.* 2001). In the direct method, metagenomic DNA is extracted from the environmental sample without separating bacterial cells from other particles. It is generally accepted that the direct extraction method accesses the entire microbial community because

loss of the particularly rare representatives is minimised and may also introduce less bias towards specific species during separation of microorganisms from other particles (Daniel 2005). However, this method is also known to yield low molecular weight DNA because of the harsh process involved in DNA extraction (Sharma *et al.* 2008).

Courtois *et al.* (2001) however, found no significant differences in bacterial diversity when comparing direct and indirect extraction methods from soil collected from an upper arable field. In another study, Delmont *et al.* (2011) reported that full pyrosequencing of both direct extracted and indirect extracted DNA from the same prairie soil showed statistically similar sequence distribution of the majority of the metabolic functions and species. Although some microbial functions differed at the 95% confidence interval in bootstrap analyses, the overall functional diversity was the same.

Many protocols for the extraction of environmental DNA have been published during the past decade, some of which have been commercialized. Although there are differences with respect to shearing, purity and quantity of the isolated DNA, the basic concept of enzymatic cell lysis and detergent treatment is mainly preferred (Rondon *et al.* 2000). Some protocols apply mechanical lysis through the use of bead beating (Voget *et al.* 2003), freeze drying (Kennedy *et al.* 2007) or sonication (Sharma *et al.* 2007). Cells lysis is generally followed by solvent extractions to obtain a crude preparation of nucleic acid. The desired molecular size of the DNA is important and needs to be considered during cell lysis. Mechanical lysis is not particularly suitable if high molecular weight DNA is required, specifically for the preparation of large insert metagenomic libraries (Nair *et al.* 2014).

The most critical aspects of extracting DNA from the environmental samples are getting metagenomic DNA that fairly represents all the environmental microorganisms, and extracting

DNA of high purity (Nair *et al.* 2014). In particular, soil and marine sponges contain high concentrations of enzyme inhibitory compounds such as polysaccharides, humic acids and clay minerals that are often co-extracted (Nair *et al.* 2014; Singh *et al.* 2013). These require additional purification procedures to improve purity of DNA and subsequently downstream application of the extracted DNA. While purification steps reduce DNA yield, significant improvements in downstream processing of DNA has been reported (Singh *et al.* 2013; Purohit & Singh 2009).

1.1. Sequence-based bio-prospecting approaches

Two methods are used for the sequence-based approach, PCR or next generation sequencing followed by *in silico* analysis to identify the gene of interest. In PCR-based homology screening, conserved regions of similar gene classes are used to design low stringency primers to amplify a target gene (Salah *et al.* 2012). This can be conducted from cloned or uncloned DNA. In the next generation sequencing method, metagenomic DNA from the environment is sequenced and the gene of interest is identified using *in silico* analysis. Alternatively, DNA from the environments can be fractionated and cloned into suitable vectors before sequencing. The sequence data can then be analysed for genes of interest, to predict metabolic pathways or comparison of metabolic activities of metagenomes from different environments. These two approaches are discussed in the following sections.

1.1.1. PCR-based methods

The PCR-based approach relies on the gene probes designed from known genes which share sequence similarity with the desired gene (Simon & Daniel 2011). As such, this approach can

only be used to screen for genes belonging to known families, which is one of the limitations of this method. PCR-based methods can be applied on metagenomic libraries or directly on isolated metagenomic DNA. When metagenomic DNA is directly used, time-consuming procedures of cloning, transformation and clone maintenance can be avoided. However, unlike when metagenomic DNA is used directly as template, PCR from clone libraries allows the recovery of the full gene when more intensive PCR techniques such as gene walking are applied.

Polz & Cavanaugh (1998) have shown that PCR-based strategies bear the risk of bias through unequal amplification of mixed template. In their study to explore potential causes and the extent of bias in PCR amplification of 16S rRNA, genomic DNA of two closely and one distantly related bacterial species were mixed and amplified with universal, degenerate primers. Quantification and comparison of template and product ratios showed that there was considerable and reproducible over amplification of specific templates. Using mutagenised templates containing AT- and GC-rich priming sites, they showed that GC-rich templates were amplified with higher efficiency, indicating that different primer binding energies may also influence template amplification. Interestingly, gene copy number was found to be an unlikely cause of the observed bias. Biases were reduced considerably by using high template concentrations, by performing fewer cycles, and by mixing replicate reaction preparations.

On the contrary, PCR amplification of specific sequence targets within a community has also been shown to depend on the abundance that those sequences represent relative to the total DNA template (Gonzalez *et al.* 2012). Using quantitative, real-time, multiplex PCR and specific Taqman probes, Gonzalez and colleagues showed that the amplification of 16S rRNA genes indicate that the relative amplification efficiency for each bacterial species is a nonlinear function of the abundance that each of those taxa represent within a multispecies DNA template. Low

proportion taxa in a community were under-represented suggesting that a large number of sequences need to be processed to detect some of the bacterial taxa.

The advantage of a PCR-based screening strategy is its independence on gene expression and production of foreign proteins in the library host. The PCR-based strategy has led to the successful identification of different enzyme encoding genes. PCR primer pairs targeting 10 environmental clades and sub-clades of the dimethyl sulfoniopropionate (DMSP) demethylase protein, DmdA, were designed and used to amplify *dmdA* genes from composite free-living coastal bacterioplankton DNA (Varaljay *et al.* 2010). Morimoto & Fujii (2009) combined PCR-denaturing gradient gel electrophoresis (DGGE) to retrieve full lengths of functional benzoate 1,2-dioxygenase alpha subunit genes (*benA*) from a soil metagenome. Nitrite reductase (*nirK*) encoding genes were isolated from ammonia-oxidizing archaea of soils and other environments using metagenomic PCR amplification (Bartossek *et al.* 2010). Other genes which have been isolated through this method include [Fe-Fe]-hydrogenases (Schmidt *et al.* 2010), [NiFe] hydrogenases (Maroti *et al.* 2009), hydrazine oxidoreductases (Li *et al.* 2010) and chitinases (Hjort *et al.* 2010).

1.1.2. NGS-based methods

NGS-based metagenomic bio-prospecting involves direct sequencing of mixed microbial genomes or a clone library prepared from environmental genomes followed by *in silico* interrogation of sequence data for the presence of genes of interest or genetic markers (Figure 2). This method was greatly encouraged by the advancement in sequencing technology intertwined with the urgency to understand microbial metabolic dynamics from different environments.

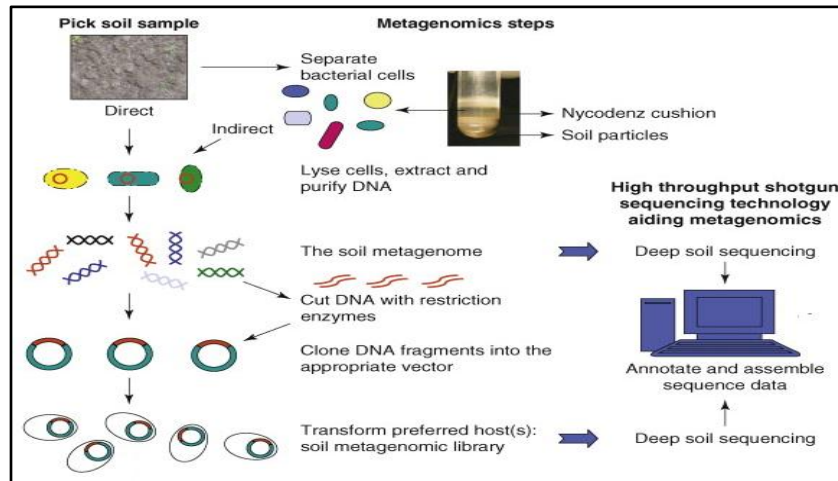


FIGURE 2: General representation of NGS based metagenomic studies. DNA is extracted from an environmental sample through a direct or indirect method. Mixed microbial genomic DNA may be fragmented through restriction enzymes and sequenced directly. Alternatively, a clone library may be constructed followed by sequencing. Adapted from Liu *et al.* (2012).

Indeed, sequencing technology has been nothing less than a paradigm shift for metagenomic studies. Sanger & Nicklen (1977) published a DNA sequencing method which was based on chain termination known as the Sanger sequencing method. One of the limitations of Sanger sequencing is the requirement of *in vivo* amplification of DNA fragments that are to be sequenced, which is usually achieved by creating a genomic library of the DNA to be sequenced. This introduces host related biases, lengthy cloning procedures and is quite labour intensive and expensive (Hall 2007).

The development of next generation sequencing (NGS) technology has supplanted the Sanger method. NGS platforms are cost-effective, high throughput and offer reasonable read lengths. The various technologies include the Ion Torrent, Roche 454 (Roche), Illumina HiSeq 2500, SOLiD 5500xl, PacBio RS II, and the Oxford NanoporeMinION (Table 3). The Ion Torrent is able to generate 8.2×10^7 reads of 200bp in up to 4 hours per run. While the error rate of this technology is comparable to that of 454 (Roche) GS FLX+, the Ion Torrent is much quicker than

the Roche 454 which requires up to 23 hours per run. In addition, the Ion Torrent is about 85 times more cost effective than the Roche 454. However, the Roche 454 GS FLX+ generates much longer reads (700bp) than the Ion Torrent. The Illumina HiSeq platforms on the other hand have much lower error rate compared to the Torrent and Roche 454. In addition, these platforms generate paired reads. However, they are much slower sequencing platforms than the Roche 454 and Ion Torrent. Although the Roche 454 has been discontinued, this platform has pioneered next generation sequencing. Like the Illumina HiSeq platforms, the SOLiD 5500xl requires up to 6 days per run. This platform has high error rate and generate short reads. The third generation platforms are generally high error rate platforms. SoLiD, like Roche 454 is no longer available on the market. The Oxford NanoporeMinION generate up to 38% errors per single pass while the PacBio RS II: P6-C4 has a 13% error rate.

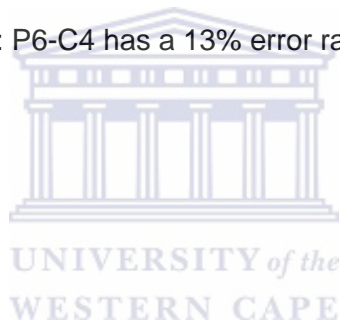
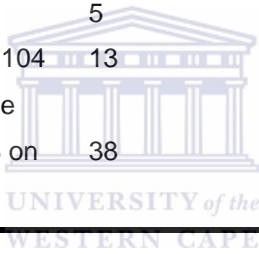


Table 3: Comparison of different generations of sequencing platforms

Platform	Generation	Read length (bp)	Single pass error rate (%)	No. of reads per run	Time per run	Cost per million bases (USD)
Sanger ABI 3730xl	1 st	600–1000	0.001	96	0.5–3 h	500
Ion Torrent	2 nd	200	1	8.2×10^7	2–4 h	0.1
454 (Roche) GS FLX+	2 nd	700	1	1×10^6	23 h	8.57
Illumina HiSeq 2500 (High Output)	2 nd	2 × 125	0.1	8×10^9 (paired)	7–60 h	0.03
Illumina HiSeq 2500 (Rapid Run)	2 nd	2 × 250	0.1	1.2×10^9 (paired)	1–6 days	0.04
SOLiD 5500xl	2 nd	2 × 60	5	8×10^8	6 days	0.11
PacBio RS II: P6-C4	3 rd	1.0–1.5 × 10 ⁴ on average	13	$3.5–7.5 \times 10^4$	0.5–4 h	0.40–0.80
Oxford NanoporeMinION	3 rd	2–5 × 10 ³ on average	38	$1.1–4.7 \times 10^4$	50 h	6.44–17.90

Adapted from Rhoads & Au (2015)



The two most widely used NGS technologies for metagenomic sequence data analysis are the 454 pyrosequencing and the Illumina technologies (Scholz *et al.* 2012). The 454 technology does not require cloning steps. This technology uses emulsion PCR which amplifies the DNA sample in an *in vitro* DNA system (Tawfik & Graffiths 1998). DNA molecules are sheered and attached to streptavidin beads. Individual DNA molecules are captured into separate emulsion droplets which act as individual amplification reactors that produce over 10^7 copies of the original template (Margulies *et al.* 2005). This is subsequently transferred into a well of a picotiter plate and the clonally related templates are analysed using a pyrosequencing reaction. The pyrosequencing approach measures the release of inorganic pyrophosphate (PPi) by chemiluminescence (Margulies *et al.* 2005). The solution of dNTPs is added to immobilised DNA template resulting in the release of PPi when ever the complementary nucleotide is attached to a template. In the Illumina technology, DNA amplification is achieved by a technique called single molecule array. A single stranded DNA molecule is attached to a solid surface using an adapter and amplified through solid-phase bridge amplification (Illumina, Inc.). DNA molecules are subsequently “bent over” and hybridises to complementary adapters, forming the template for the synthesis of their complementary strands. The templates are sequenced using DNA sequencing-by-synthesis. This approach employs reversible fluorescent labelled terminators with removable fluorescent moieties. Amplification is achieved by special DNA polymerases that can incorporate these terminators into growing oligonucleotide chains (Morozova & Marra 2008). The general overview of DNA sequencing technology is summarised in Figure 3.

The major challenge of NGS technology in metagenomic bio-prospecting is the assembly of sequence reads. In particular, assembly of large pathways from mixed microbial genomes can be challenging because of the highly conserved and repetitive organisation of genes in these pathways (Oulas *et al.* 2015).

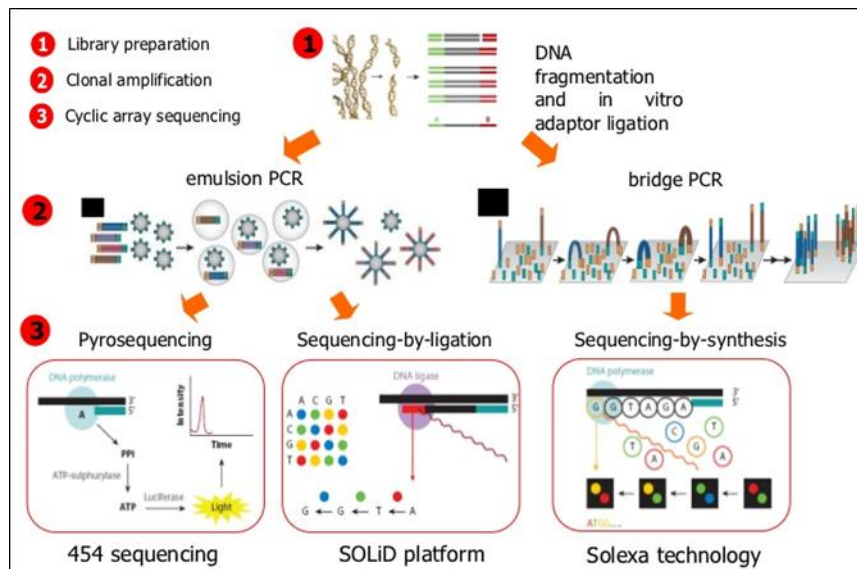


FIGURE 3: Summary of DNA sequencing process. A library of environmental DNA is first prepared through restriction digest. Depending of the sequencing technology, adaptors are attached to fragmented DNA molecules (Sequence by synthesis), or DNA fragments are amplified through emulsion PCR (Pyrosequencing and sequencing by ligation). Adapted from Liu *et al.* (2012).

However, with substantial amount of sequencing, assembly of sequences from metagenomic libraries can result in good draft or even complete genomes when the target species shows little intra-species variation. For example, fifteen draft genomes were constructed by direct assembly of a cow rumen metagenome (Hess *et al.* 2011).

1.2. Function-based metagenomic screening

The function-based screening involves identification of a heterologous host expressing the phenotype of interest (Uchiyama & Miyazaki 2009). This approach is not dependent on sequence homology, and is most useful in the discovery of novel genes and activities (Heath *et al.* 2009). The main setback of functional-based metagenomics is its dependence on expression vectors and an amenable heterologous host (Sharma *et al.* 2008). In most cases, tens of thousands to millions of clones need to be screened in order to fully represent every environmental gene, making this approach labour intensive and time consuming. A general representation of function-base metagenomics is shown in figure 4.

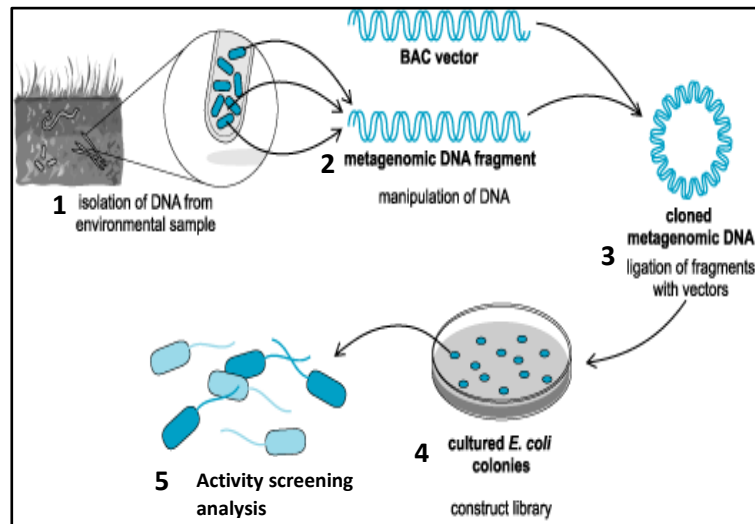
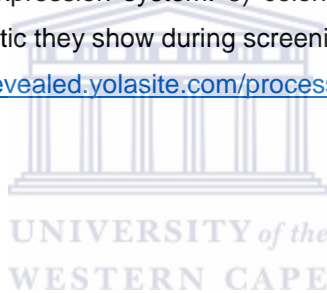


FIGURE 4: Representation of function-based metagenomics from sample collection to heterologous expression. 1) Microbial genomes are extracted from environmental samples. 2) The extracted DNA is size selected and end repaired. 3) The manipulated DNA is used to construct a clone library in plasmids, fosmids or BAC vectors. 4) The clones are then used to transform a surrogate host which is used as an expression system. 5) colonies are analysed for positive activity based on the phenotypic characteristic they show during screening.

Adapted from <http://metagenomicsrevealed.yolasite.com/process-1.php>



1.2.1. Cloning vectors

DNA from an environmental sample is first cloned into a suitable vector, creating a library of environmental DNA fragments (Daniel 2005). The efficiency of identifying a gene of interest from a library depends on a number of factors, including the size of the target gene and its abundance in the pool of environmental genes, the assay method, and the heterologous expression machinery used to host the library (Simon & Daniel 2011). Depending on the gene to be screened and the approach to be used for screening, short insert or large insert metagenomic libraries can be constructed (Simon & Daniel 2011).

Large insert metagenomic libraries, which are constructed in vectors such as a bacterial artificial chromosome (BAC), fosmids and cosmids, require high quality and molecular weight DNA. BAC vectors accommodate up to 300kb inserts. However, the most commonly

used large insert vectors are cosmids and fosmids. Fosmids are cosmids, except that they contain an F plasmid origin of replication and are low copy number vectors. This allows the clone to maintain stability and conserve energy (Kim *et al.* 1992). Cosmids on the other hand are plasmids, and have been associated with cell instability (Sudek *et al.* 2007), which result from cellular accumulation of the expressing vector and subsequently breakdown of membranes. These libraries are generally suitable for cloning of enzyme activities and pathways that are encoded by moderately large (20-50kb) gene clusters. However, high molecular weight DNA required for construction of this library may be difficult to obtain, making the large insert library technically challenging (Daniel 2005).

Small insert libraries are constructed in plasmid vectors which carry DNA fragment sizes of up to 20kb (Daniel 2005). Although it is generally accepted that large fragment libraries improves the probability of identifying the desired activity, Simon *et al.* (2009) used both plasmid and fosmid vectors to clone DNA polymerase I from the same metagenome and found no significant difference in hit rates between the two vectors. Plasmid vectors are generally high copy number, which allow detection of foreign genes that are weakly expressed. Since small DNA fragments are used in small fragment libraries, sheared DNA fragments can be cloned and expressed. However, a major drawback of small insert libraries is that a larger number of clones need to be screened in order to identify a positive clone. Moreover, these plasmid libraries are not suited to screening for activities that are encoded by large fragments of DNA (Schmeisser *et al.* 2007).

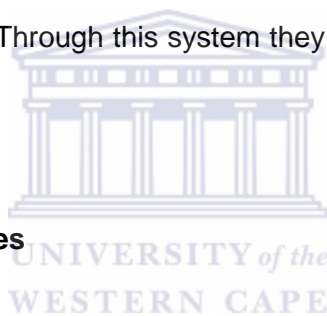
1.2.2. Heterologous hosts

Following insertion of environmental DNA into a suitable vector, the vector is transformed into a heterologous host. One of the most common systems used for metagenomic studies is *E. coli* and the pCC2FOS system (Jogler *et al.* 2009). This is mainly because *E. coli* is well studied, and its physiochemical and genetic machinery is well understood (Schlegel *et al.*

2013). However, not all environmental genes can be transcribed and translated into functional proteins in *E. coli*.

Over the past five years, a lot of effort has been invested toward the development of multiple host systems. Troeschel *et al.* 2010 developed a multiple host expression system using a shuttle vector that allows screening of metagenomic DNA in *E. coli*, *Pseudomonas putida* and *Bacillus subtilis*. This was done in an attempt to overcome stringency associated with inefficient transcription of target genes and improper folding of the corresponding enzyme in a single host system. Six different bacteria: *Agrobacterium tumefaciens*, *Bacillus graminis*, *Caulobacter vibrioides*, *E. coli*, *P. putida*, and *Ralstonia metallidurans* were used to host metagenomic DNA libraries to demonstrate the potential advantage of using multiple diverse host species (Craig *et al.* 2010). Through this system they were able to significantly improve the hit rate.

1.2.3. Screening approaches



Different versions of function-based screens have also been developed over the past decade (Uchiyama & Miyazaki 2009). Induced gene expression, also called substrate induced gene expression (SIGEX) was introduced by Uchiyama *et al.* (2005). This approach uses an operon-trap expression vector which carries a green fluorescent protein gene without a promoter (Figure 5). The green fluorescent gene is linked to a cloned gene that is under control of an inducible catabolic promoter. In the presence of IPTG, the recombinant gene is co-expressed with the green fluorescent protein gene, allowing for the positive clone to be identified by fluorescent activated cell sorting (FACS) (Uchiyama & Watanabe 2007). The advantage of this method is that it allows ultra-high throughput screening of the library through FACS. However, this method can only be used for discovery of inducible genes. In addition, it is also possible for effectors other than the specific substrates to induce transcription, resulting in false positives.

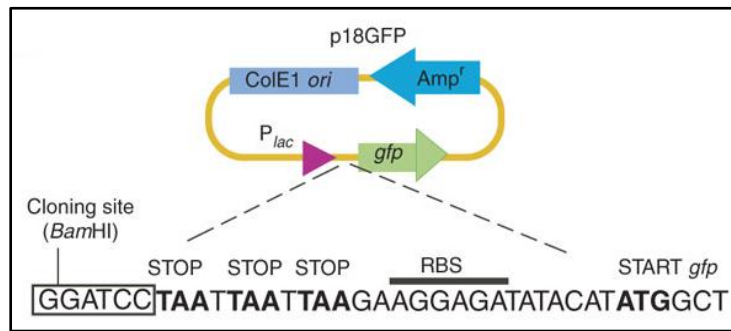


FIGURE 5: Schematic representation of p18GFP expression system. The *gfp* gene expression is under the control of the *lac* promoter. Source: Yun & Ryu (2005).

Another high throughput “intracellular” screening system called METREX (Metabolite-Regulated Expression), in which cloned metagenomic DNA fragments are in a host cell containing a biosensor for compounds that induce bacterial quorum sensing, was developed by Williamson *et al.* (2005). When the metagenomic clone produces a quorum-sensing inducer, the cell produces green fluorescent protein (GFP) and can be identified by fluorescence microscopy or captured by fluorescence-activated cell sorting (Figure 6). While this method is not based on the functional activity of the product, it can be applied for functional bioprospecting of metagenomic libraries.

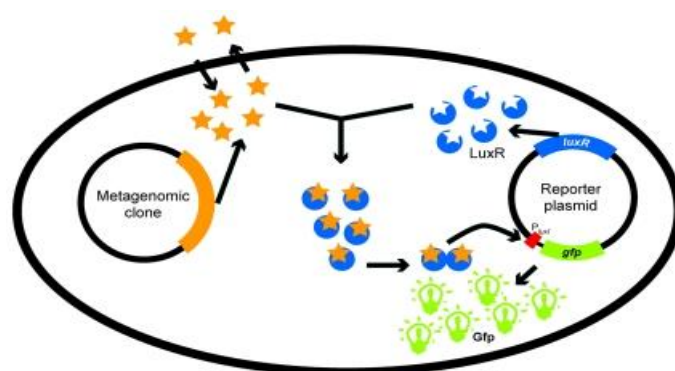


FIGURE 6: Representation of METREX screening system. In this intracellular screen, the biosensor that detects active clones is inside the same cell as the metagenomic DNA. The biosensor detects small diffusible signal molecules that induce quorum sensing. When the signal molecules reach a sufficient concentration, they bind the LuxR transcriptional activator which activates P_{luxR} and induces expression of target genes, in this case GFP. Adapted from Williamson *et al.* (2005).

Uchiyama & Miyazaki (2010) also developed a reporter-based screening method called product-induced gene expression (PIGEX) which was used to screen for amidases. A benzoate-responsive transcriptional activator, BenR, was placed upstream of the gene encoding green fluorescent protein and used as a sensor. *E. coli* sensor cells carrying the *benR-gfp* gene cassette fluoresced in response to low benzoate concentrations but were completely unresponsive to the substrate benzamide (Figure 7).

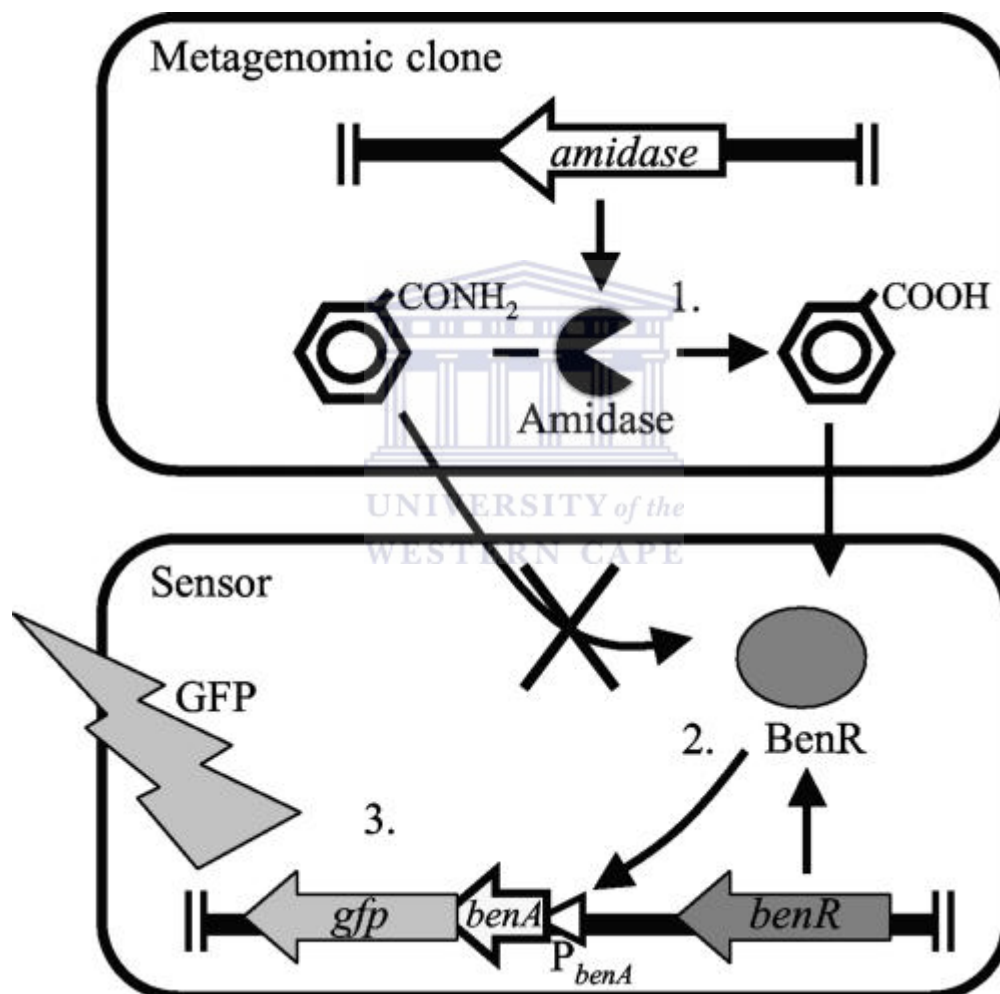


FIGURE 7: Schematic representation of the PIGEX. The amidase-positive clone catalyzes the conversion of benzamide to benzoate (step 1). Benzoate activates the transcriptional regulator BenR, which in turn activates the *benA* promoter (P_{benA}) (step 2) and induces the expression of a reporter gene (*gfp*; step 3). Enzymatic benzoate production is measured as GFP fluorescence. Adapted from Uchiyama & Miyazaki (2010).

1.2.4. Challenges of function-based metagenomics

Uchiyama & Miyazaki (2009) published a comprehensive review on the challenges of functional metagenomics in novel gene discovery. In their review, heterologous gene expression was identified as a major challenge. Prior to their review, Gabor *et al.* (2004) employed *in silico* analysis to estimate that only 40% of enzymatic activities can be identified by random cloning of environmental DNA in *E. coli*. Studies elsewhere have also highlighted heterologous expression as a challenge limiting the robustness of metagenomic techniques to fully access metabolic potential (Ferrer *et al.* 2009).

1.2.4.1. Promoter and Codon recognition

The major problems associated with heterologous expression are lack of promoter recognition and codon preference between different microorganisms. Generally, the promoter sequences vary between different microbial species, and as such some of the promoter sequences cannot be recognised by *E. coli* RNA polymerase. Consequently, the gene whose promoter cannot be recognised by the host organism polymerase cannot be identified (Huang & Lindblad 2013). On the other hand, strong constitutive promoters from environmental genomes might lead to the production of a high amount of the desired protein in the surrogate host, which might have a toxic effect on the host.

The degeneracy of the DNA code (20 amino acids are encoded by 61 different codons) results in different codons which can encode for insertion of the same amino acid into a protein. The frequency of codon usage in an organism directly reflects the amount of corresponding tRNAs produced by that organism (Stoletzki & Eyre-walker 2007; Chen & Inouye 1994). Therefore, foreign genes with rare codons are most likely to be expressed with defective proteins because their expression is limited to the available aminoacyl-tRNA. In addition, rare codons that are located close to the translation initiator have been shown to

result in ribosome arrest, preventing the entry of the next incoming ribosome (Wong & Chang 1986).

1.2.4.2. Instability of biological molecules

Biological molecules such as DNA (Pierce & Gutteridge 1985), messenger RNA (mRNA) (Baneyx 1999) and recombinant proteins (Dedhia *et al.* 2000) are highly unstable. DNA vectors, like mRNA and proteins, are the most important component of functional metagenomics studies. Vectors are foreign materials in the host system, making them highly susceptible to nucleases of the host (Pierce & Gutteridge 1985). This can reduce the copies of gene of interest in the metagenomic library and therefore reduce the probabilities of identifying the desired gene.

Messenger RNAs, which are produced during transcription of the desired gene, are also highly susceptible to exonuclease digestion. In particular, mRNA of foreign genes which have no significance to the host metabolic activities, are rapidly degraded by the host's exonucleases (Baneyx 1999). As a result, genes of interest that have fewer copies in the pool of metagenomic genes are most likely to be entirely destroyed at the level of mRNA.

Recombinant proteins have been shown to be susceptible to host protease degradation. Foreign proteins with high non-polar amino acids content at the C-terminus are known to be rapidly degraded, while proteins with polar or charged amino acids in the last five positions at the C-terminal are less susceptible to protease degradation (Dedhia *et al.* 2000). Since most *E.coli* proteins are synthesized in the cytoplasm, foreign proteins in the *E. coli* host are more susceptible to protease degradation. Although it is possible to construct metagenomic libraries in vectors that also encode for protein that direct a cloned gene product to the cytoplasm, the inner or outer membrane or the periplasmic space (Hoffman 1985), high

levels of a foreign protein in the membrane are also known to interfere with normal cellular functions and be lethal to the cell (Ghrayeb *et al.* 1984).

1.2.4.3. Insoluble aggregates and protein folding

Production of high concentrations of foreign proteins results in the misfolding of the protein of interest and its subsequent degradation by cellular proteases or its deposition into biologically inactive aggregates known as inclusion bodies (Blackwell & Horgan 1991). Inclusion bodies result from the failure of the host system to repair or remove defective proteins (Fahnert *et al.* 2004). In most instances, inclusion body formation is minimized by using weak promoters (Sharma *et al.* 2007), the use of low copy vectors (Devasahayam 2007) and incubation at low temperatures (Glick & Whitney 1987). However, it is difficult to find a good balance between conditions that allows bacterial growth and efficient expression of recombinant protein.



Protein folding is also a problem in metagenomic screening for active enzymes. Most proteins need to be folded to adopt a functionally active conformation during translation (Gershenson & Gierasch 2011). Proteins that are naturally excreted are usually incorrectly folded in heterologous hosts because the cytoplasmic environment does not favor the formation of di-sulphide bonds and glycosylation (Andersen & Krummen 2002). In the *E. coli* system, protein folding is facilitated by molecular chaperones and foldases. The co-expression of molecular chaperones and the gene of interest have been shown to facilitate target protein folding and enhance production of recombinant active proteins (Hartl *et al.* 2011). However, this process could further escalate the labour required for metagenomic bio-prospecting.

1.2.5. Next generation metagenomic bio-prospecting

In 2009, Ferrer *et al.* proposed *in vitro* metagenomics as an alternative solution to help overcome the limitations associated with heterologous expression discussed above. Cell-free expression reactions can be compartmentalised for screening through FACS in order to increase the throughput. Although no work has been published yet on *in vitro* metagenomics, Uchiyama *et al.* (2005) employed FACS to increase the throughput of screening a metagenomics gene library (SIGEX as discussed above).

1.2.5.1. Cell-free protein synthesis

Cell-free protein synthesis (CFPS) is a coupled transcription / translation system in which protein synthesis is separated from metabolic activities associated with cell growth and maintenance (Shrestha *et al.* 2012). This enables direct protein synthesis in a test tube wherein the transcription, translation, and protein folding components such as aminoacyl-tRNAsynthetases, ribosomes, translation initiation and elongation factors, chaperons and foldases are provided by cell extracts (Kigawa *et al.* 2004). Protein synthesis is activated by addition of essential components, which include amino acids, energy system, cofactors, salts and nucleotides followed by incubation at the desired temperature for roughly 3 hours (Figure 8) (Carlson *et al.* 2011). Numerous proteins which are difficult to express *in vivo* have been produced in a CFPS system, including toxic proteins (Avenaudo *et al.* 2000), membrane proteins (Isaksson *et al.* 2012), virus particles (Bundy & Swartz 2011), and proteins with unusual amino acids (Bundy & Swartz 2010).

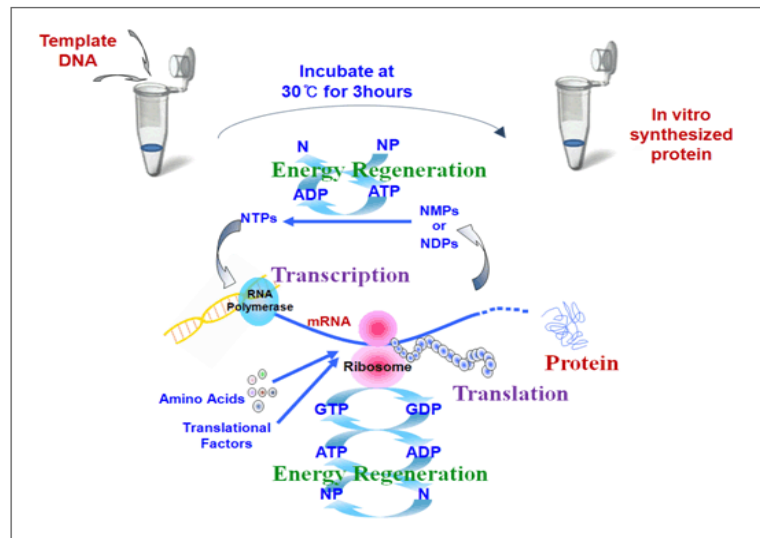


FIGURE 8: General representation of Cell-free protein synthesis. The main components of the system include energy regeneration components, RNA polymerase, transcriptional factors, ribosomal RNA and amino acids. Source: Carlson *et al.* (2011)

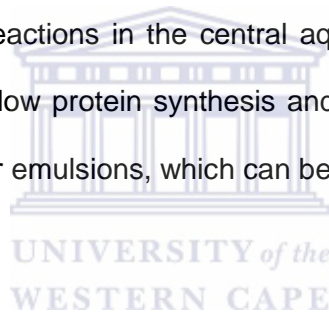
Until recently, IVC approaches have used bacterial cell extracts for transcription and translation. However, for efficient expression, complex multi-domain proteins require eukaryotic systems and the post-translational modifications these systems provide. The use of wheat germ (Yonezawa *et al.* 2003) and rabbit reticulocyte (RRL) (Ghadessy & Holliger 2004) extracts have extended the application of IVC to a wider spectrum of protein targets, although it is also possible to exchange these commercially available reagents for the individual reaction components. This is exemplified by the PURE synthesis method: a cell-free translation system using recombinant elements (Shimizu *et al.* 2005).

Although cell-free extracts from different organisms have been reported, *E. coli* based extracts are commercially available and commonly used due to its simplicity and well known machinery (Carlson *et al.* 2011). Since the early protocols for preparation of cell-free extracts from *E. coli* (Moore *et al.* 1966), several improvements have been made to reduce costs and improve the efficiency and yield of this system (Shrestha *et al.* 2012). These include the use

of high density fermentors to improve biomass yield (Liu *et al.* 2005), the use of sonication for cells lysis (Shrestha *et al.* 2012) and low speed centrifuge (Kim *et al.* 1996) to reduce costs and different energy sources to improve yields (Kim & Kim 2009).

1.2.5.2. *In vitro* compartmentalisation (IVC)

In order to facilitate screening through FACS, the CFPS reaction is commonly compartmentalized in cell-like oil droplets (Bernath *et al.* 2004). This process is called *in vitro* compartmentalisation. The general process of developing *in vitro* compartmentalisation involves encapsulating a CFPS reaction, which contains a fluorescent substrate and DNA molecule, with oil surfactant mixture in a test tube to develop numerous oil droplets. Each of these droplets contains CFPS reactions in the central aqueous phase (Miller *et al.* 2006). The droplets are incubated to allow protein synthesis and then re-coated with an aqueous layer to develop water in oil water emulsions, which can be sorted through FACS (Figure 9).



These droplets function as cell-like compartments, in each of which a single gene (or suite of genes) is transcribed and translated to give multiple copies of the protein it encodes (Miller *et al.* 2006). Each droplet measures about 2µm in diameter and contains a volume of about 5 femtolitres of CFPS reaction. A 1ml mixture might contain about 10E10 droplets, which represents a large library of proteins in a single micro-centrifuge tube (Bernath *et al.* 2004).

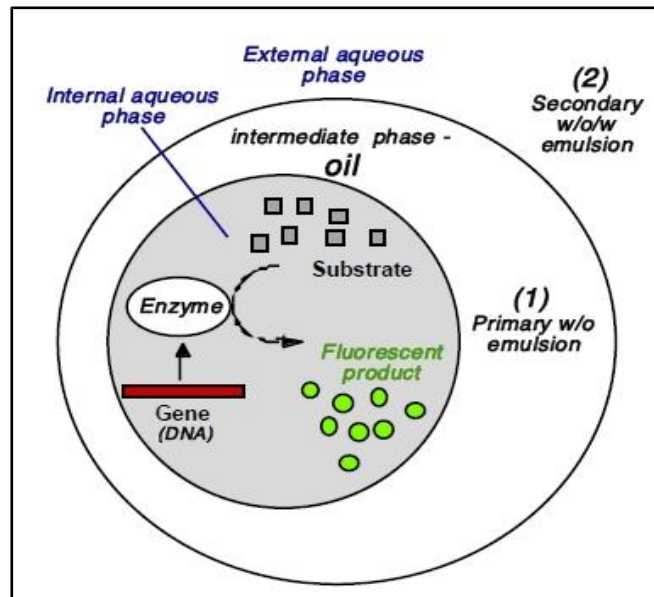


FIGURE 9: General representation of *in vitro* compartmentalisation. The internal aqueous phase contains cell-free protein synthesis components, template DNA and substrate. This phase is coated by intermediary oil that is surrounded by external aqueous phase to form W/O/W emulsion. Source: Bernath *et al.* (2009).

This technology allows the emulsion droplets to be manipulated in different ways, including the addition of components inside the emulsion (Leamon *et al.* 2006), merging and splitting of two droplets (Link *et al.* 2004), incubation and thermo-cycling of droplets (Schaerli *et al.* 2009), and sorting (Agresti *et al.* 2010). IVC has been used in the selection and evolution of numerous proteins and bioactive compounds, including evolution of beta-galactosidases (Mastrobattista *et al.* 2005), selection of restriction endonucleases (Doi *et al.* 2004), selection of ribozymes (Agresti *et al.* 2005), evolution of phosphotriesterase, (Griffiths & Tawfik 2003), evolution of hydrogenases (Stapleton & Swartz 2010), and selection of zinc finger DNA-binding Proteins (Sepp & Choo 2005). The general overview of this technology is represented in Figure 10.

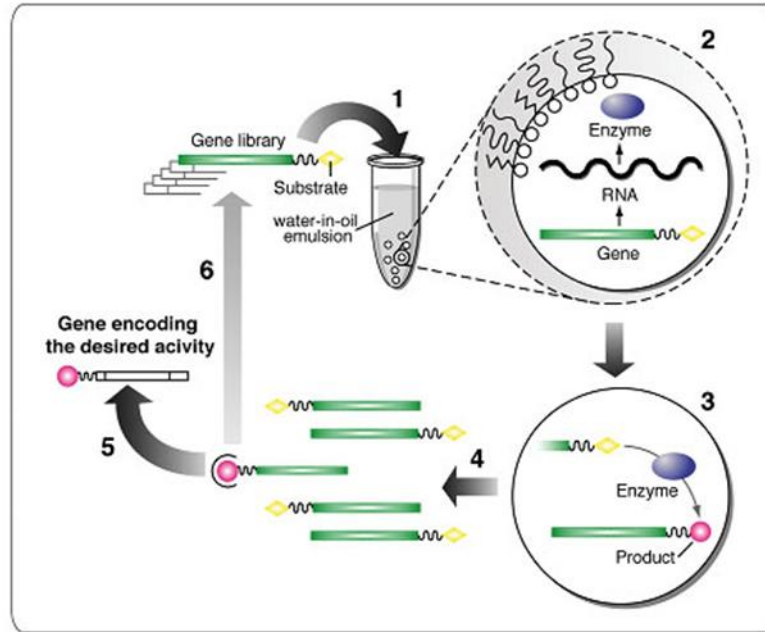
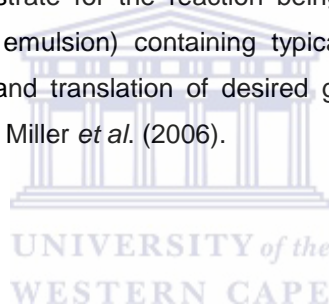


FIGURE 10: Overview of IVC system. An *in vitro* transcription/translation reaction mixture containing a library of genes linked to a substrate for the reaction being selected is dispersed to form aqueous compartments (in a water-in-oil emulsion) containing typically one DNA molecule. The reaction is incubated to allow transcription and translation of desired gene into proteins molecules that can be identified through activity. Source: Miller *et al.* (2006).



1.2.5.3. Fluorescence Activated Cell Sorting (FACS)

Fluorescence is one of the most sensitive and versatile ways of detecting biological activities and is extremely useful as a high-throughput screening (HTS) system (Mastrobattista *et al.* 2005). FACS can routinely sort over 10^7 clones per hour and has already proven to be a highly successful technique to select different proteins (Bernath *et al.* 2009). This technique allows analysis of individual cells based on the fluorescence and light scattering. Particles to be analyzed are introduced into a column of pressurized sheath fluid, and passed through a laser beam. The interactions of the particles with the light beam of specific wavelength result in the excitation of the particle, which releases fluorescent light (Herzenberg *et al.* 2002). The cytometer gathers information about the fluorescence characteristics released by the particle. The particles then pass through the stream for the break-off distance, where

particles containing fluorescent substances are charged. Charged drops then pass through two high-voltage deflection plates and are deflected into collection vessels or aspirated to waste (Figure 11). Depending on the nature of experiment and the general objective, collected droplets are broken to recover DNA molecule(s) (Bernath *et al.* 2004).

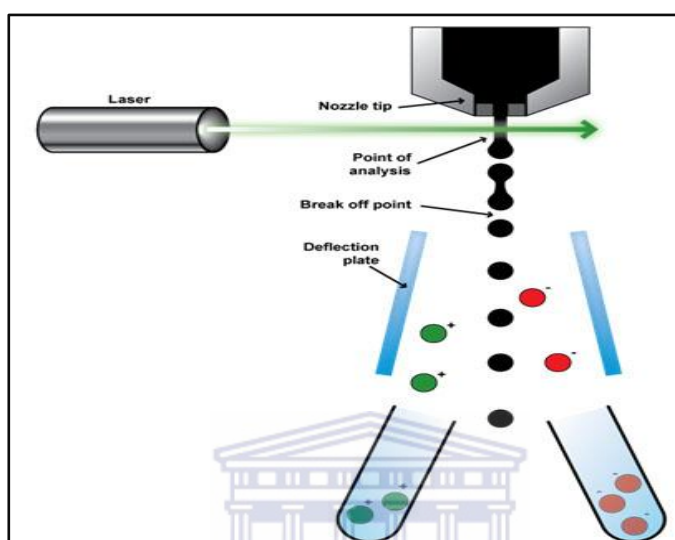


FIGURE 11: Screening of enzyme activity through fluorescence activated cell sorter. (<https://www.biw.kuleuven.be/m2s/cmpg/FACSCoreFacility/equipment/sorting>).

In this study, a high throughput screening system based on cell-free protein synthesis, IVC and FACS will be developed for bio-prospecting lignocellulase encoding genes from thermophilic horse composting manure. The next section will therefore review lignocellulase discovery through current metagenomics strategies.

1.3. Metagenomic as a tool for novel lignocellulases discovery

The depletion of fossil fuel has greatly influenced the development of lignocellulosic-based generation of biofuels, which in turn has resulted in a surge of enzyme screening research from different environments. The use of metagenomics for lignocellulase gene discovery has been well documented and very successful (Bastien *et al.* 2013). Hemicellulose in particular

is one of the three major components of plant cell wall that provide strength and stability, making it one of the most abundant biomass on earth after cellulose. Chemical and enzymatic hydrolysis of these materials results in different hemicellulose sugars being released. Fermentation of these, depending on the microorganism used, results in several different biofuels being produced including methanol, ethanol, butanol and methyl or ethyl esters.

1.3.1. Hemicellulose

Hemicellulose is mainly found between cellulose fibers, which is the inner component of lignocellulose and lignin, which forms an external protective layer (Perez *et al.* 2002) (Figure 12). The most common hemicellulose found in hardwoods is xylan (Lemmel *et al.* 1986). Soft wood mainly contains β -mannan as the principal component, although xylan is also common in many plants (Postma 2012). Xylans are composed of a D-xylopyranosyl backbone linked together through β -1,4-glycosidic bonds (Subramaniam & Prema 2002). This backbone is modified with various side chains depending on their source (Shallom & Yuval 2003). The side chains of hardwood xylans includes methyl-D-glucuronic acid linked to xylose through 1,2-glycosidic bonds, and an acetic acid linked to the xylose through an ester bond (Postma 2012). In softwood xylans, the side chains include L-arabinofuranose residues attached to the xylan backbone by glycosidic bonds. The hemicellulose found in grass species contain phenolic and ferulic acid side chains further attached to the arabinofuranose side chains through ester bonds (Shallom & Yuval 2003). Other hemicelluloses include beta-mannan, galactoglucomannans, arabinogalactans and arabinan (Howard *et al.* 2003).

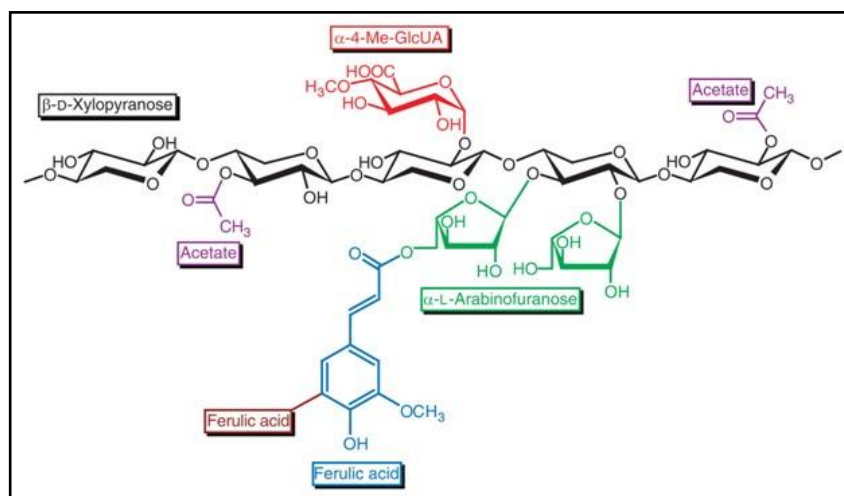


FIGURE 12: Structure of hemicellulose. The main chain is made of beta-D-xylopyranose. Side chains include acetyl groups, ferulic acid, alpha-L-arabinofuranose and alpha-4-Me-GlcUA. Source: Polizeli *et al.* (2005).

1.3.1.1. Hydrolysis of hemicellulose

The complete hydrolysis of the xylan backbone involves the activities of two enzymes, endo-1,4-beta-xylanase (EC 3.2.1.8) and beta-xylosidases (EC 3.2.1.37). Beta-xylanase randomly attacks the main xylan backbone releasing short xylooligosaccharides, which are then acted upon by beta-xylosidases to release xylose, the monomer of xylan (Perez *et al.* 2002). The release of xylan side chains involves numerous enzymes, including alpha-L-arabinofuranosidases (EC 3.2.1.55), which hydrolyse arabinose side chains; alpha-D-glucuronidase, which releases gluconic acid side chains; galactosidase, which releases galactose side chains and xylan esterase which hydrolysis the ester bond (Shallom & Yuval 2003). Hydrolysis of beta-mannan hemicellulose of softwood occurs through beta-mannanases (EC 3.2.1.78), which liberate short beta-1,4-mannooligomers (Chauhan *et al.* 2012). The short beta-1,4-mannooligomers are further hydrolyzed to mannose by beta-mannosidases (EC 3.2.1.78) (Shallom & Yuval 2003). Hydrolysis of arabinofuranosyl hemicelluloses involves alpha-l-arabinofuranosidases (EC 3.2.1.55) (Saha 2000). The side chains of arabinofuranosyl hemicelluloses are detached by alpha-D-glucuronidases that

cleave the alpha-1,2-glycosidic bond of the 4-O-methyl-D-glucuronic acid. Figure 13 shows a general representation of lignocellulose biodegradation.

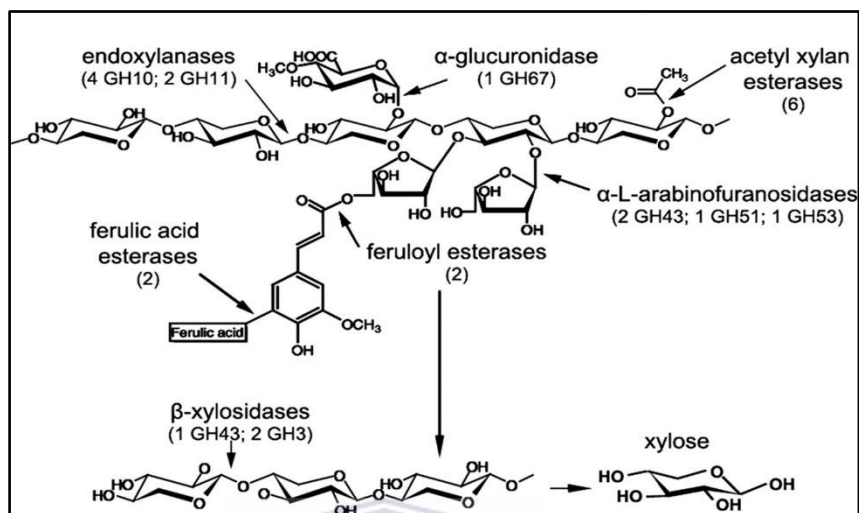
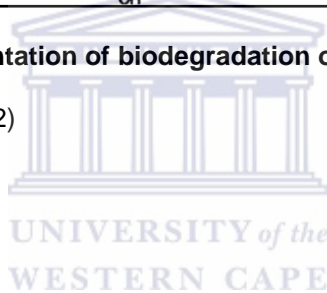


FIGURE 13: Schematic representation of biodegradation of hemicellulose backbone and side chains. Source: Perez *et al.* (2002)



1.3.1.2. Beta-xylosidases

Xylan 1,4-beta-D-xylosidases or beta-xylosidases are enzymes that catalyse the hydrolysis of single xylosyl residues from the non-reducing end of 1,4-beta-D-xylooligosaccharides (Jordan & Wagschal 2010). They are either extracellular or intracellular enzymes which contribute to the hydrolysis of xylan (Henrissat *et al.* 1991). Currently, these enzymes are used in combination with xylanase in different large-scale industrial applications. This includes the removal of ink in recycled paper (Sharma *et al.* 2006), processing wood pulp in the paper industry (Postma 2012), improving bread dough baking and nutritional quality (Jordan & Wagschal 2010), hydrolysis of xylosylated compounds from grape juice during extraction and liberation of aroma derived from xylosylated compounds of grapes during wine making (Rensburg & Pretorius 2000), and hydrolysis of xylan to D-xylose residues for subsequent reduction to xylitol (Granström *et al.* 2007). The most attractive potential

application of beta-xylosidases is in hydrolysing lignocellulosic biomass for the production of biofuels such as ethanol and butanol.

1.3.1.2.1. Biochemical properties of beta xylosidases

The most important biochemical properties of any industrial lignocellulytic enzyme includes high k_{cat} and k_{cat}/K_M , low affinity for monosaccharide inhibitors, good stabilities at extreme pH values and temperature, low levels of adsorption/inactivation by biomass feedstocks, and low cost of enzyme production (Hu *et al.* 2011). Beta-xylosidases are produced by different organisms, including bacteria and fungi. A few thermostable xylosidases have been reported. This includes beta-xylosidase with an optimum temperature of 70°C and pH 6.0 isolated from *Bacillus thermantarcticus* (Lama *et al.* 2004); a thermostable 150kDa beta-xylosidase that shows stability at 70°C from *Bacillus stearothermophilus* (Nanmori *et al.* 1990). Terrasan *et al.* (2013) reported another beta-xylosidase with optimal temperature and pH value of 75°C and 4.0 respectively.

In *Thermoanaerobacterium saccharolyticum* JW/SL-YS485, a beta-xylosidase with maximum activity at the temperature of 65°C and pH 6.0 has been reported by Shao *et al.* 2011. This enzyme was also shown to have K_M of 28mM for *p*-nitrophenyl- beta-d-xyloside, and a V_{max} of 276 U/mg. In addition, the enzyme retained 70% activity in the presence of 200mM xylose, making it one of the most suitable for various industrial applications. Another high xylose and thermotolerant beta-xylosidase was recently isolated from the extremely thermophilic bacterium *Thermotoga thermarum* DSM. The optimal activity of this enzyme was obtained at pH 6.0 and temperature of 95°C and was active at 500mM xylose (Shi *et al.* 2013). A beta-xylosidase with optimum temperature of 65°C and pH of 5.0 has been reported in *Colletotrichum graminicola*. This enzyme is interesting because when combined with xylanase, crude extracts from the same organism, and cellulase from *Trichoderma*

reesei, the resulting cocktail is able to hydrolyse raw sugarcane waste with a glucose yield of 33.1% in just 48 h, demonstrating excellent potential for hydrolysis of lignocellulosic material (Zimbardi *et al.* 2013). Table 4 and 5 summarises some of the characterised beta-xylosidases.

Table 4: Properties of some of the characterised fungal beta-xylosidases.

Organism	Optimum temperature (°C)	Optimum pH	K_M (mM)	V_{max} (umol/min/mg)	Specific activity (U/mg)
<i>Aspergillusawamori</i> K4	70	4			19.58
<i>Aspergillusfumigatus</i>	65	4.5			
<i>Aspergillus japonicas</i>	70	4	8.7	114	112
<i>Aspergillusnidulans</i>	50	5	0.31	25.6	
<i>Aspergillusochraceus</i>	70	3.3-5	1.1	39	
<i>Aspergillusphoenicis</i>	75	4.4-5	0.66		
<i>Fusariumproliferatum</i>	60	4.5	2.36	75	53
<i>Paecilomyces thermophile</i>	55	2.27	0.77		43.4
	60	2-4	4.3	0.48	8.96
	50	4	0.75		353.6
<i>Penicillium herquei</i> S1	30	6.5			224.6
<i>Penicillium herquei</i> S2	75	4			171.8
<i>Penicilliumjanczewskii</i>	50	7		114	
<i>Sporotricum thermophile</i>	70	4.4-5	1.1		3.4
<i>Trichodermaharzianum</i>	60	4	0.05		
	55	4.5	0.42		10.8
<i>Trichoderma viride</i>	60	2.5	5.8	1.7	149.4
<i>Talaromyces emersonii</i>	50	5.5	0.13		3.996
<i>Xylaria regalis</i>					

Modified from Kousar *et al.* (2013)

Table 5: Properties of some of the characterised bacterial beta-xylosidases.

Organism	Optimum temperature (°C)	Optimum pH	Km (mM)	Vmax (umol/min/mg)	Specific activity (U/mg)
<i>Aeromonascaviaerec</i>	50	6	0.34	33	172.9
<i>Bacillus halodurans</i>	45	7	1.9	0.65	34.2
<i>Bacillus stearothermophilus</i>	70	6	1.2		
<i>Bacillus thermentarciticus</i>	70	6	0.5		160
<i>Caldocellumsaccharolyticumrec</i>					
<i>Clostridium acetobutylicum</i>	65	5.4	10	64	49
<i>Clostridium cellulolyticum</i>	45	6.0-6.5	3.5		
<i>Streptomyces Sp.</i>	35	7.5	0.40	19.6	17.5
<i>Thermoanaerobacter ethanolicus</i>	45	7.5	13.5		15.11
<i>Thermomonospora fusca</i>	93	5.9		0.89	66
<i>Thermoanaerobacterium saccharolyticum</i>	40-60	5-9	28	276	8.0
	65	6.0			45.8

Modified from Kousar *et al.* (2013)

1.3.1.2.2. Classification of beta-xylosidases

There are 132 characterised beta-xylosidases listed in the carbohydrate active enzymes (CaZy) database. These proteins are classified into eleven glycoside hydrolase (GH) families based on their amino acid sequence (1, 3, 5, 30, 39, 43, 51, 52, 54, 116 and 120) and tree carbohydrate binding module (CBM) families (6, 22 and 42) (<http://www.cazy.org/Glycoside-Hydrolases.html>, accessed July 2015). Most of the GH39 and GH43 beta-xylosidases have been reported from bacteria such as *Bacillus* sp. and *Clostridium* sp. (Wagschal *et al.* 2008). The majority of fungal beta-xylosidases that have been described in the CaZy database belong to families 3 and 43. One of the more interesting characteristics of GH3 and GH43 beta-xylosidases is that they involve bi-functional activities of beta-xylosidases/alpha-l-arabinofuranosidases, while others have xylanase activity (Barker *et al.* 2010). Further discussion of the classification and characterisation of beta-xylosidases will be presented in the coming chapters.

1.4. Aims and objectives of the study

The aim of this study is to develop an ultra-high throughput cell-free approach for screening hemicellulose encoding genes from uncloned metagenomics DNA. This approach is designed to overcome most of the limitations associated with conventional function-based screening approaches as indicated above. Three widely used techniques: Cell-Free Transcription-Translation (CFTT), *In vitro* compartmentalisation (IVC) and Fluorescence Activated Cell Sorting (FACS) will be combined to develop a robust technique of tapping the metabolic potentials contained in metagenomic DNA. Although commonly used in the scientific domain, these techniques have never been used in combination to screening of metagenomic DNA library. The versatility of this technique will be tested by screening beta-xylosidase encoding genes derived from thermophilic horse manure.

The main objectives of the study are:

- I. Develop metagenomic IVC-FACS (mIVC-FACS) screening system for enzymes bioprospecting
- II. Analyse microbial diversity and glycoside hydrolase content of the thermophilic composting horse manure metagenome obtained by next generation shotgun sequencing
- III. Apply PCR base screening method for beta-xylosidase encoding genes from composting horse manure
- IV. Apply classical function-based method for screening of beta-xylosidase encoding genes from composting horse manure
- V. Compare Actinobacteria based mIVC-FACS screening with a commercial *E. coli* extract for screening of beta-xylosidase encoding genes from composting horse manure
- VI. Compare IVC-FACS with classical function-based metagenomic screening system

CHAPTER TWO: Phylogenetic analysis of thermophilic horse manure compost metagenome (HMM), assessment of glycoside hydrolases (GH) and PCR-based bio-prospecting of beta-xylosidases.

2. Introduction

The paucity of enzymes that efficiently deconstruct plant polysaccharides represents a major bottleneck for industrial-scale conversion of lignocellulose biomass into biofuels. The ability of microorganisms to produce these enzymes have been reported in different environments, including insects (Engel & Moran 2013) and animal guts (Bäckhed *et al.* 2004), composting materials (Martins *et al.* 2013) and forest soil (Hong *et al.* 2007). In particular, composting materials and animal manure are resident to diverse lignocellulose degrading microbial consortia (Yu *et al.* 2007). This is mainly because composting does not only offer the necessary energy and carbon source for growth, but the fluctuating temperatures and acidity insure the cycle of different microbial species (Wu *et al.* 2009).

A series of microorganisms produce lignocellulytic enzymes that convert lignocellulosic material to carbon and energy for growth (Cayuela *et al.* 2008). This process involves variations in temperature that influences the composition of microbial communities within the environment at any given time. Mesophilic microorganisms dominate the microbial community during the early stage of composting. At this stage, organic acids producing mesophiles such as *Lactobacillus* spp. and *Acetobacter* spp are predominant (Golueke *et al.* 1954). The increase in temperature resulting from microbial metabolic processes within the compost results in a shift in the microbial population towards thermophilic organisms. This temperature can increase from about 30°C to as high as 70°C (Neher *et al.* 2013). Yañez *et al.* (2009) reported that Actinobacteria and *Bacillus* spp. are the dominating microorganisms

during this stage. The thermophilic phase has also been shown to be essential for rapid degradation of lignocellulose. In particular, organic compounds like lignin are mainly degraded by thermophilic microfungi and actinomycetes. The optimum temperature for thermophilic fungi is 40–50°C that is also the optimum temperature for lignin degradation during composting (Tuomela *et al.* 2000). The maturation process is characterised by gradual decrease in temperature and the microbial population shifts back towards mesophilic organisms (Guo *et al.* 2007). The tolerance of microbial communities to temperature changes and dynamic redox conditions demonstrate their potential to produce lignocellulolytic enzymes which can withstand industrial conditions (Allgaier *et al.* 2010).

Shotgun sequencing-based analysis of microbial communities is a valuable means of interrogating mixed microbial genomes of uncultured microorganisms (Culligan *et al.* 2014). Analysis of metagenomics sequence data for microbial diversity and their metabolic potential not only provides a platform for the identification of novel functionalities, but also provides understanding of interdependencies underlying microbial life (Mardis 2008). In some functional studies, inspection of metagenomics data for biocatalyst encoding genes enrichment has been applied as the first step towards exploring functional novelty. For instance Hess *et al.* (2011) sequenced and analyzed 268Gb of metagenomic DNA from microbes adherent to plant fiber incubated in cow rumen. From these data, they identified 27,755 putative carbohydrate-active genes and expressed 90 candidate proteins, of which 57% were enzymatically active against cellulosic substrates.

This chapter investigates the microbial phylogenetic diversity of the thermophilic stage of composting horse manure metagenome (HMM) through shotgun sequencing and also assesses its suitability for bio-prospecting of glycoside hydrolases, specifically beta-xylosidase. In addition to sequence analysis of HMM, PCR-based bio-prospecting of beta-xylosidases based on low stringency primers was done to compare the efficiency of this method with cell-free method that this study is developing.

2.1. Materials and methods

2.1.1. Samples collection

Horse compost manure was collected from a commercial compost farm located in the Western Cape Province of South Africa (34°S 2' 53.35', 18°E 31' 45.71' E). The compost source material consisted of an unspecified mix of horse manure, wood chips and sawdust, with lesser amounts of plant debris. The temperature of the sample was measured at 70°C. The collected samples were stored at room temperature until extraction of metagenomic DNA.

2.1.2. Metagenomic DNA extraction

Metagenomic DNA was extracted using the chemical lysis method as described by *Zhou et al.* (1996). Each extraction was performed using 1.6g compost re-suspended in 5ml extraction buffer (100mM Tris-HCl, pH 8.0; 100mM EDTA, pH 8.0; 100mM sodium phosphate, pH 8.0; 1.5M NaCl; 1% CTAB). A volume of 20µl proteinase K (10U) was added and the mixture incubated at 37°C for 30 min. Cell lysis was performed through the addition of SDS (2% w/v) and PVPP (0.5% w/v) with further incubation at 65°C for 2 hours. Debris was pelleted by centrifugation at 6000g for 10 min at room temperature. The supernatant was carefully removed and added to 1 volume phenol:chloroform:isoamyl alcohol (24:24:1) followed by gentle inversion of the tube and centrifugation at 16000g for 10 min. The aqueous phase was removed, added to an equal volume of chloroform and transferred to micro-centrifuge tubes. Following centrifugation at 16000g for 10 min, the aqueous phase was again recovered and precipitated with 1 volume isopropanol at room temperature overnight. Crude nucleic acids were pelleted by centrifugation at 16000g for 20 min at room

temperature. The supernatant was discarded and the pellet washed with cold 70% v/v ethanol, air-dried and re-suspended in an appropriate volume of 1 x TE buffer.

2.1.3. Metagenomic DNA purification

The extracted DNA was further purified by a formamide purification method as reported by Liles *et al.* (2008). Briefly, 500µl of 20ng/ml DNA was mixed with 500µl of 2% agarose gel prepared in TAE buffer. The mixture was allowed to solidify in a 1ml plastic syringe. The solidified agarose-plug containing compressed metagenomic DNA was removed from the syringe and placed within a 15ml centrifuge tube containing 80% formamide and 0.8M NaCl in 20mM Tris-HCl buffer (pH 8.0), providing a 60 to 70% final formamide concentration. Formamide is known to denature genomic DNA and associated proteins which enables inhibitors to be released from the DNA helix, while sodium chloride enhances genomic DNA stability during denaturation (Rondon *et al.* 2000). The plug was gently suspended in a formamide solution, washed by gently inverting the tube several times. Formamide solution was replaced, and the plug was incubated overnight at room temperature with gentle stirring. The agarose plug was placed in a new 1% low melting point (LMP) agarose gel and metagenomic DNA was electrophoresed 3h at 70V. High-molecular weight (<23kb) DNA was excised from the LMP agarose. The agarose plug was transferred to a sterile 2ml micro-tube and incubated at 70°C for 10 minutes to melt the agarose. This was followed by additional incubation at 42°C to equilibrate the mixture. DNA-agarose mixture was treated with agarase (Fermentas) at 1U of enzyme per 10mg of agarose followed by gentle mixing and incubation at 42°C for 2h. The reaction was heat inactivated at 70°C for 10min followed by centrifugation at 9000g for 10min. Supernatant was removed and DNA was precipitated by adding 2.5 volume of absolute ethanol and 0.1 volume of sodium acetate (pH 7.0), followed by incubation at -20°C overnight. DNA was pelleted by centrifugation at top speed for 30 minutes.

2.1.4. Sequencing of metagenomic DNA and reads assembly

Next generation sequencing of metagenomic DNA was performed in an IlluminaMiSeq sequencer using the NexteraXT library preparation kit (Illumina) and a 10% phiX v3 spike as per the manufacturer's instructions (Preparation Guide, Part #15031942 Rev A May 2012) as well as the MiSeq Reagent kit V2 (500 cycle). Sequencing was performed at the Institute for Microbial Biotechnology and Metagenomics (IMBM), University of the Western Cape, Cape Town, South Africa. The raw reads were trimmed (bases with a Q-score less than 36 were trimmed from the 3'end) and de-multiplexed at the sequencing facility generating 2 x 250 bp reads, resulting in a set of paired (read pairs, forward and reverse of the same 250bp sequence).

Following removal of human contaminants and phiX spike, raw Illumina reads were assembled using the CLC Genomics Workbench (version 4.0.3; CLC Bio, Cambridge, MA, USA). Paired-end reads were assembled using the following parameters: mismatch cost 2, insertions cost 3, deletion cost 3, length fraction 0.5 and similarity of 0.8. Total amount of DNA was calculated by adding assembled DNA (bp) to the total amount of un-assembled, underreplicated DNA (bp).

2.1.5. Analysis of microbial diversity and sample coverage

Microbial diversity of HMM was analysed on the MG-RAST server (Meyer *et al.* 2008). Raw reads were uploaded to the servers and compared for homology in M5NR 16S rRNA databases using default parameters. The amount of sequencing required to cover the total diversity present in the HMM sample was done using Nonpareil pipeline (Rodriguez-R & Konstantinidis 2014).

2.1.6. Identification of glycoside hydrolases (GHs) and beta-xylosidase reads

A local library was created through the NCBI protein database search using the input “beta-xylosidase” (accessed May 2015). Three hundred sequences annotated as beta-xylosidase were downloaded from the database and uploaded into CLC workbench. Raw reads or assembled reads (contigs) were compared to the local library on CLC workbench using the following parameters: Maximum E-value of 1e-5; percentage identity of 20% and alignment length of 15. Sequences that met these criteria were further compared to proteins in the NCBI protein database. GHs sequences were identified by comparing raw reads to proteins in MG RAST M5NR proteins databases using default parameters (Maximum E-value cut-off: 1e-5; percentage identity cut-off: 60%; alignment length cut-off:15). The M5NR is an integration of different sequence databases (e.g. NCBI, KEGG or the EGGnogs, PATRIC, etc) into one single, searchable database.

2.1.7. Calculation of abundance index (AI)

The number of sequence reads with homology to beta-xylosidases or GHs in the local library was used to determine abundance index (AI) as described by (Aziz *et al.* 2015). AI describes the fraction of a metagenome that matches a given sequence. The total number of beta-xylosidase or GH sequence reads (n) in a pool of metagenome (g), represent the fraction (f) in the metagenome (g) and therefore the fraction (f) can be represented as $\sum n/g$. This value reflects the abundance of all sequences with similarity to a sequence of interest in a metagenomic library.

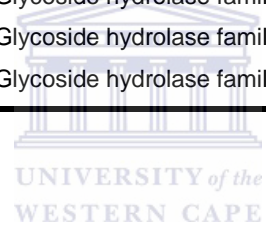
2.1.8. Primers used in this chapter

Degenerate primers were designed using two online services, COBALT (Constraint Based Alignments Tool) (Papadopoulos & Agarwala 2007) and MEME motif finder application (Bailey *et al.* 2009). A total of 23, 100, 100 and 8 sequences annotated as beta-xylosidases from glycoside hydrolase family 3, 39, 43, and 52 respectively were randomly selected and downloaded from the NCBI database. Sequences belonging to the same family were aligned using COBALT. After alignment, most of the sequences from family 39 and 43 were found to be highly redundant and were reduced to 24, and 19 respectively. The resulting sequences were analysed for conserved motifs using MEME motif finder application. The server was restricted to identify three most conserved motifs per family. Identified primers were analysed for secondary structure using DNAMAN 2.1 (LynnonBiosoft, San Ramon, CA, USA). Degeneracies were calculated manually. Table 6 summarises primers that were designed from those motifs.



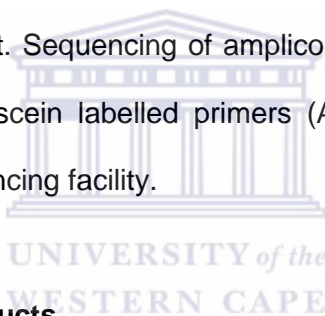
Table 6: Degenerate primers used in this study

Primer name	Oligo sequence	Annealing used (°C)	Target	Primer degeneracy	Estimated amplicon size (bp)
GH3DF	GARACNCCNGGNGARGAY	52	Glycoside hydrolase family 3	512	995
GH3DR	YTGNCNCGGRTACCACAT	52	Glycoside hydrolase family 3	64	995
GH39DR	RTTRTAYTCNGTDATRTG	52	Glycoside hydrolase family 39	32	1000
GH39DR2	YTCYTCRAANACRTC	52	Glycoside hydrolase family 39	64	150
GH39DF	GGNMGNCTNNGNCT	52	Glycoside hydrolase family 39	512	1000
GH43DF	TGGGARCARATHGGNCAY	52	Glycoside hydrolase family 43	96	294
GH43DR	CATRTGNCCRTARTCNGT	52	Glycoside hydrolase family 43	128	294
GH52DF	GGNGTNCAGWSNCTN	52	Glycoside hydrolase family 52	512	1100
GH52DR	CATRTCRTGNGTRAA	52	Glycoside hydrolase family 52	32	1100



2.1.9. PCR amplifications

PCR amplifications using degenerate primers were done in a 50µl reaction mixture consisting of 1x Phusion buffer, 200µM dNTPs, 0.5µM reverse primer, 0.5µM forward primer, 50ng template DNA, and 1U high fidelity Phusion polymerase. The thermocycle conditions were: 98°C for 30sec, 35 cycles (98°C for 10sec, 52°C for 30sec, 72°C for 1min), 72°C for 2 min. PCR amplicons were mixed with a loading dye (60% v/v glycerol and 0.25% w/v Orange G) and loaded into the wells of a cast gel followed by electrophoresis for 45 minutes at 100V. Agarose gel solution contained 0.7% w/v of agarose gel prepared in 0.5 X TAE buffer (0.2 % v/v Tris base; 0.5 % v/v glacial acetic acid and 1% v/v 5M EDTA, pH 8.0) and 10µl/L of 0.5µg/ml ethidium bromide. Amplicons were excised from the gel and purified using the Nucleospin gel purification kit. Sequencing of amplicons was done using an automated DNA sequencer 373 and fluorescein labelled primers (Applied Biosystems, USA) at the University of Stellenbosch sequencing facility.



2.1.10. Cloning of PCR products

The PCR products were purified using a Nucleospin PCR clean-up kit (MACHEREY-NAGEL, Germany) according to manufacturer's recommendations. Cleaned PCR products were cloned in pJET 1.2/ blunt (Fermentas) according to the manufactures recommendations. The reaction mixture contained 10µl of 2 X reaction buffer, 2µl PCR product (55ng/µl), 1µl pJETvector, and 5U of T4 ligase per µg of DNA and 6µl ddH₂O. The mixture was incubated at room temperature for 20 minutes and then used to transform electro-competent *E. coli* JM109 cell.

2.1.11. Preparation of electro-competent cells

E. coli JM109 or BL21 (DE3) from glycerol stock were plated on LB plates and incubated overnight at 37°C. A single colony from an overnight culture was transferred to 5ml LB broth in 50ml flask and incubated overnight at 37°C with shaking at 150rpm. This culture was used to inoculate pre-warmed 50ml LB broth in a 250ml flask. The flask was incubated at 37°C with shaking at 150rpm until optical absorbance (OD) at 600nm reached about 0.6. The cells were transferred to a 500ml centrifuge tube, incubated on ice for 15 minutes and harvested by centrifugation at 3500g (Beckman J-26 XP, USA) for 10 minutes. The pellet was gently re-suspended in 10ml of ice cold 0.1M CaCl₂, incubated on ice for 30 minutes and harvested by centrifugation as above. The pellet was again gently re-suspended in 10ml of ice cold solution of 0.1M CaCl₂ containing 15 % glycerol. Aliquots of 100µl were transferred into 1.5 ml tubes and stored at -80°C.



2.1.12. Electroporation

Electro-competent *E. coli* JM109 cells were removed from -80°C storage and thawed on ice. Fifty microliters cell suspension was mixed with 3µl plasmid DNA/ligation mixture (~10 ng) in a pre-chilled Eppendorf tube and incubated on ice for 15 min. The mixture was transferred to a pre-chilled 0.1cm electroporation cuvette (BioRad). Electroporation was performed using a BioRad Gene PulserR at 1.8KV, 25µF and 200Ω. Following electroporation, 500ml LB broth (10g/l tryptone, 5g/l yeast extracts and 10g/l NaCl) was added to the cuvette and the cells were gently re-suspended. The cell suspension was transferred to 2ml eppendorf tube and incubated at 37°C horizontal shaking for 1 hour. Transformed cells were plated onto LB plates (LB broth containing 15g/l agar and 100µg/ml ampicillin) followed by incubation at 37°C overnight.

2.2. Results and discussions

2.2.1. Metagenomic DNA extraction

The chemical lysis method yielded high molecular weight DNA, above 23kb (Figure 14). However, the high content of humic acid present in the extracted DNA prevented enzymatic activities in the subsequent processes, and therefore an extra purification step based on formamide treatment was incorporated. Formamide has been shown to lower melting temperatures (T_m) of DNA linearly by 2.4-2.9 degrees C/mole of formamide depending on the GC composition, helix conformation and state of hydration of DNA (Blake & Delcourt 1996). At high formamide concentration in the presence of salt, DNA maintains its stability allowing impurities such as humic acid to move out of the double helix without destabilising its structure (Liles *et al.* 2008).

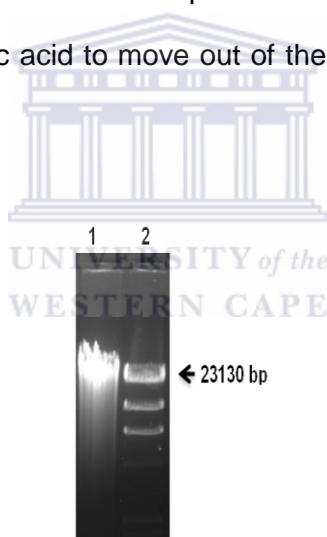


FIGURE 14: Agarose gel electrophoresis of extracted metagenomic DNA. Lane 1 shows metagenomic DNA extracted. Lane 2 is lambda DNA marker digested with *HindIII*.

2.2.2. Next generation sequencing of HMM

Metagenomic DNA was extracted and sequenced by the MiSeq next generation sequencing (NGS) platform. A total of 3,683,940 sequence reads were obtained from the metagenome totalling 889,980,646bp with an average length of 241bp. Sequences were quality assessed

through duplicate read inferred sequencing error estimation (DRISEE) (Keegan *et al.* 2012), kmer abundance spectra (Chor *et al.* 2009), and nucleotide positioning histogram (Meyer *et al.* 2008) in the RG RAST server. Eightysix percent of the sequences passed the QC pipeline. About 5% of these contained ribosomal RNA genes, 46% contained predicted proteins with known functions, 16% contained predicted proteins with unknown function and 19% of sequences were unknown (Figure 15).

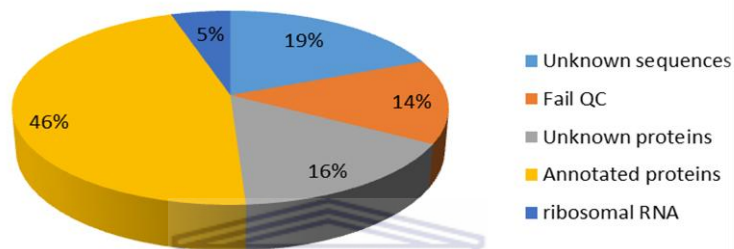


FIGURE 15: Breakdown of metagenomics sequence data. Raw sequence reads were analysed through MG RAST server using default parameters.

Raw reads were assembled into full length contigs using CLC workbench. A total of 16941 contigs were obtained with an average length of 1950bp (Table 7). Only 4% of the total metagenome was successfully assembled into full contigs. Although successful assembly of up to 30% of total metagenomic sequence data are known (Kanokratana *et al.* 2011), some studies resorted to analysing unassembled sequence reads due to difficulties of assembling highly complex metagenomics data (Warnecke *et al.* 2007).

While next-generation sequencing technologies allow the generation of reads from a complex microbial community for analyses, the assembling of a set of mixed reads from different species is a major challenge for metagenomic studies. The performances of available assemblers on metagenomic data are far from satisfactory, mainly because of the existence of common regions in the genomes of subspecies and species, which make the assembly problem much more complicated (Scholz *et al.* 2012).

Table 7: Summary of MiSeq data

Total number of reads	3683940
Average read length (bp)	241
Number of duplicate reads	243140
Number of reads after de-replicate removal	3440800
Number of assembled reads	137074
Number of contigs	16941
Average contig length (bp)	1950
Number of un-assembled reads	3303726
Total DNA after de-replicates removal (bp)	829232800

In addition, compared to traditional Sanger sequencing methods, the read length of NGS is shorter and the error rate is generally higher (0.5%–2% error per nucleotide) (Morozova & Marra 2008), making assembly even more challenging.

To determine the sample coverage of sequenced data, nonpareil pipeline (Rodriguez-R & Konstantinidis 2014) was used to estimate the sample coverage of sequenced data and the sequencing efforts required to obtain nearly complete coverage of the sample. Nonpareil relies on the observation that datasets with higher coverage are more redundant because the sequencing reads are nearly random, although some systematic biases have been noted for specific sequencing protocols (Dohm *et al.* 2008). Redundancy is defined here as the portion of reads in a dataset that match with at least one other read. Using this redundancy, Nonpareil examines the degree of overlap among individual sequence reads of a whole shotgun metagenome to compute the fraction of reads with no match, which is then used to estimate the abundance-weighted average coverage.

Figure 16 demonstrates that this sequenced data covered 59.55% of the metagenome, representing 834Mb of data. In order to achieve a near complete to complete coverage, an estimated 8.115Gb of metagenome still needs to be sequenced. This explained the low

number of reads which were assembled since most of the fragments which should be assembled into full contigs were not sequenced.

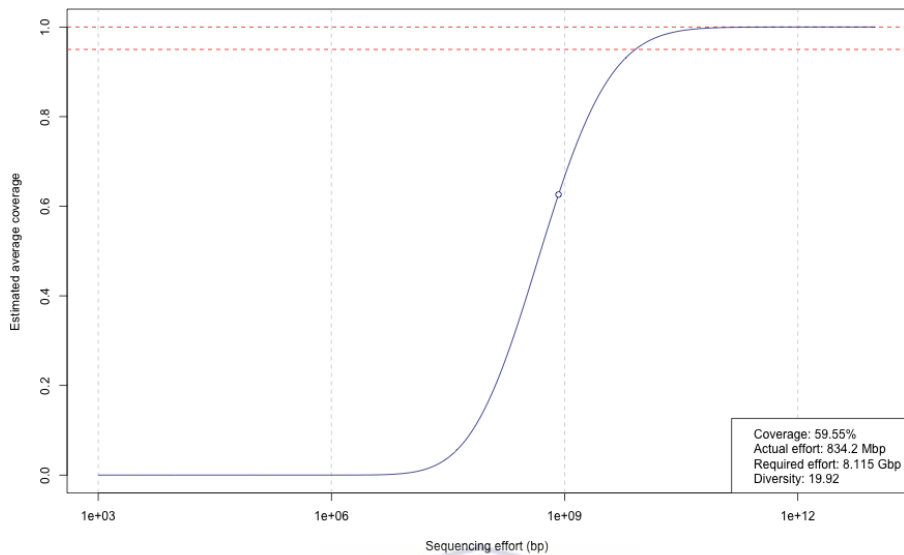


FIGURE 16: Nonpareil curve of sequenced HMM sample. The analysis was executed with default parameters (50% reads overlap, 96% identity).

2.2.3. Microbial diversity of horse manure metagenome (HMM)



Sequences that passed QC pipeline (without artificial de-replicates) were analysed for 16S rRNA sequences on the MG-RAST server using default parameters. Figure 17A shows that this metagenome is dominated by Bacteria (86.7%) and 11.9% Eukaryota domains.

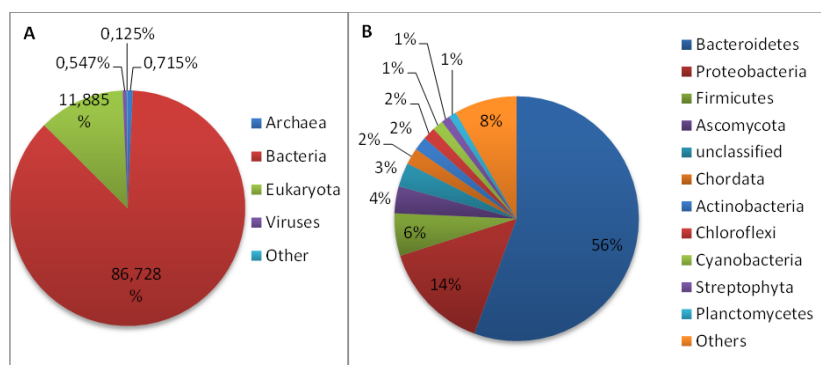


FIGURE 17: Microbial diversity of horse manure metagenome. (A) Microbial distribution within four domains of life and (B) Phylum diversity.

The *Bacteroidetes* phylum is the most dominant (Figure 17B). *Bacteroidetes* are mostly anaerobic and are widely distributed in soil, sediment, aquatic habitats, and animal guts (Mhuantong *et al.* 2015). Contrary to this finding, Uroz *et al.* (2013) showed that *Proteobacteria* were the most dominant phylum in organic materials such as compost. Danon *et al.* (2008) reported that *Gammaproteobacteria* were ubiquitous in cured and uncured compost, and that this class was the most abundant *Proteobacteria* in matured compost. However animal manure, unlike organic compost, is mainly dominated by gut microbial communities which has been reported to be dominated by *Bacteroidetes* (Green *et al.* 2004; Jami & Mizrahi 2012).

Singh *et al.* (2014) showed that *Bacteroidetes* was the most predominant phylum (30–60%), followed by *Firmicutes* (20–40%), *Proteobacteria* (8–10%), and *Actinobacteria* (3–5%) in a Buffalo rumen metagenome. This was consistent with a study by Pope *et al.* (2012) who also got similar findings from a cow rumen metagenome. In lactating cows, Jami *et al.* (2014) also reported the predominance of *Bacteroidetes*, *Firmicutes*, and *Proteobacteria* (core microbiota) in cow fed with 30% roughage. Indeed, bacterial species colonising the animal gastrointestinal track belong to the phyla *Firmicutes* and *Bacteroidetes*, while species of the phyla *Actinobacteria*, *Proteobacteria* and *Verrucomicrobia* exist in lower numbers (Eckburg *et al.* 2005).

The most abundant *Bacteroidetes* orders were *Chitinophagales*, *Cytophagales*, and *Pedobacterales* (Figure 18). *Chitinophagales* are known to encode numerous glycoside hydrolases (Glavina Del Rio *et al.* 2010). Like most bacteroidetes, *Cytophagales* are mainly chemo-organotrophs which metabolise organic macromolecules such as starch, agar and glycan. These molecules are mainly degraded by glycoside hydrolase family 3 enzymes, although some of these enzymes are also found in other families (CaZy).

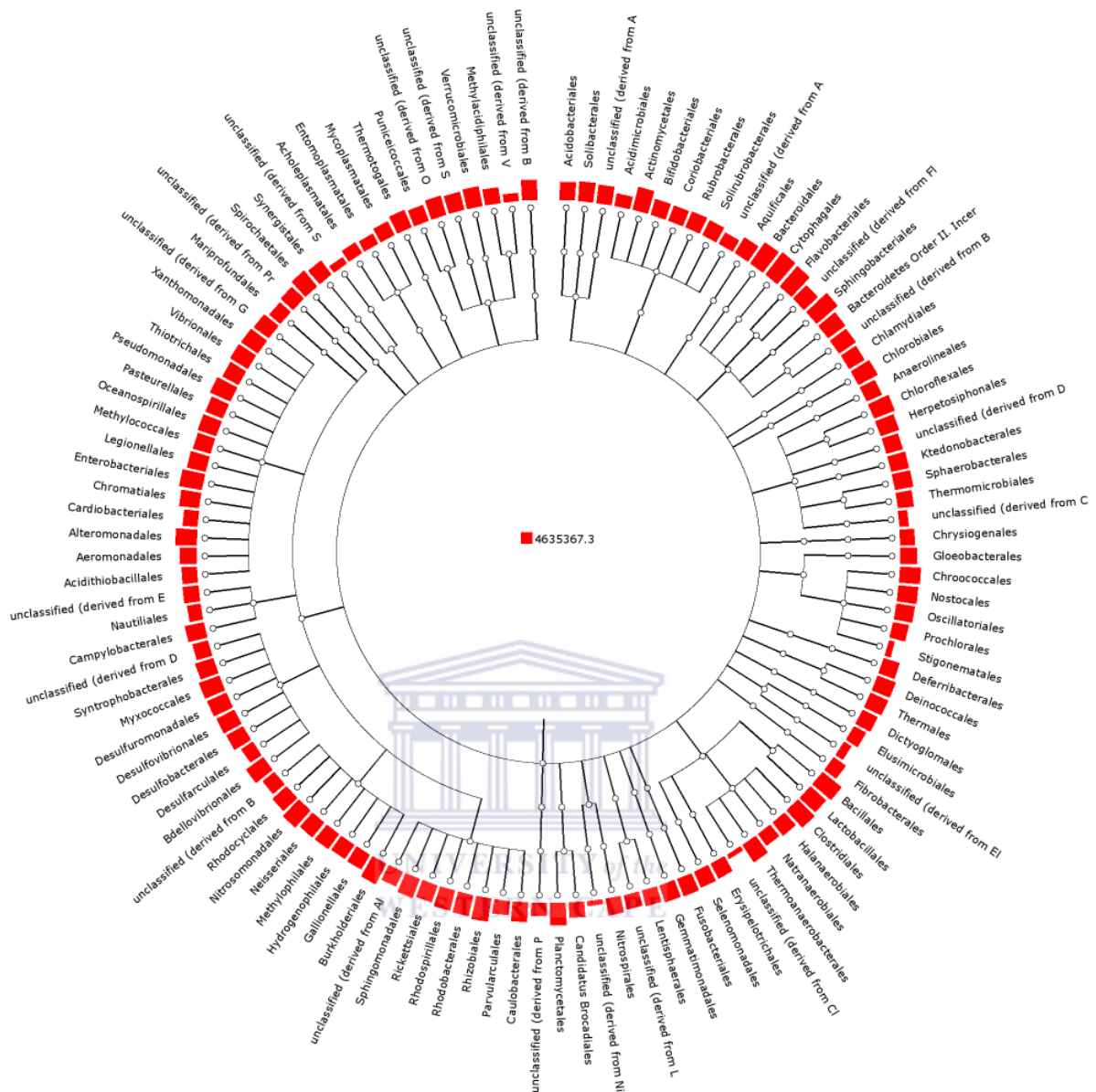


FIGURE 18: Diversity of bacterial orders in horse manure metagenome.The data was compared to the M5NR database using a maximum e-value of $1e-5$, a minimum identity of 60.

Bacillales are the most abundant *Firmicutes*. Some glycosyl hydrolases have been reported from this phylum. This includes *Bacillus coagulans* MXL-9 which was found capable of growing on pre-pulping hemicellulose extracts, utilizing all of the principle mono-sugars found in woody biomass (Walton *et al.* 2010); *Bacillus stearothermophilus* strains which were reported to produce xylanase and beta-xylosidase showing optimum activity at 60°C and 70°C respectively (Nanmori *et al.* 1990).

2.2.4. Abundance of Glycoside hydrolases (GHs)

GHs are a prominent group of enzymes that hydrolyse the glycosidic bonds among the carbohydrate molecules. Functional screening of metagenomics libraries is generally used to identify active ORFs using a functional activity based assay. Careful interrogation of the environment before functional analysis not only allows for discovery of appropriate biocatalysts but also increases the chances of identifying novel biocatalysts with unique structural and kinetic properties. To that end, HMM was analysed at the sequence level to identify sequence reads with homology to GHs. A total of 9161 reads were predicted from raw reads, representing an abundance index of 10 reads per megabase of metagenome. These reads represent 66 of 135 known families of glycoside hydrolases. Thirty-five families were represented by more than five reads (Figure 19) while the rest were represented by fewer than five reads (data not shown).

The highest published number of GHs hits per megabase of metagenome was from wood-feeding termite gut, where 20.44 hits were identified per megabase of metagenome (Warnecke *et al.* 2007). This is not surprising since termites are an extremely successful group of wood-degrading organisms (Thongaram *et al.* 2012). In the switch-grass adapted compost metagenome, Allgaier *et al.* (2010) identified 3.56 GHs per megabase. GH3 and GH43 reads were the most abundant. Polysaccharide utilization loci from *Bacteroidetes* were also found highly abundant in the microbiota of elephant faeces samples (Ilmberger *et al.* 2014). GH5 and GH9 family enzymes were the most abundant detected for *Bacteroidetes*, suggesting that bacteria of this phylum are mainly responsible for the degradation of cellulose. Enzymes from these families and those from GH10, GH39, GH11 and GH52 are involved in the hydrolysis of hemicellulose (Ilmberger *et al.* 2014).

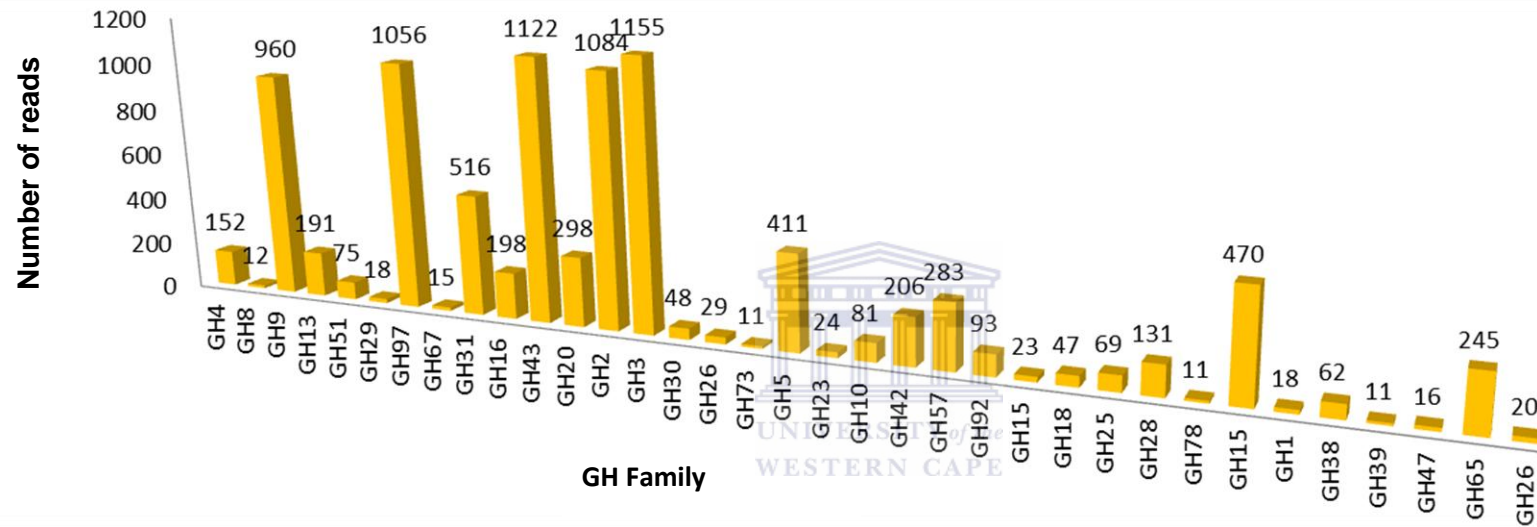


FIGURE 19: Abundance of GHs in HMM. The abundance is represented by the number of reads with homology to GHs from different families. The data was analysed through MG-RAST proteins databases by comparing reads using a maximum e-value of $1e-5$, a minimum identity of 60 %, and a minimum alignment length of 15 measured in amino.

Hemicellulose degradation requires a broader range of endo-enzymes such as endo-1,4-beta-xylanase (GH10) for hydrolysis of xylan, endo-1,4-beta-mannosidase (GH26) for mannan; and endo-1,4-beta-galactosidase (GH16) for galactan. Reads with homologies to all these genes were found in HMM. The high abundance of GH3, GH43, GH2 and GH97 might reflect the requirement to digest the higher abundance of xylan, as compared to mannan (GH26) and galactan (GH16).

2.2.5. Abundance of beta-xylosidases

Since the major focus of this thesis was on beta-xylosidases, a more focused analysis of this gene was conducted. A total of 817 reads with similarities to beta-xylosidase sequences were predicted from MHH (Figure 20). This represents an abundance index of one read per megabase of metagenome. Beta-xylosidases are members of ten GHs families which are classified based on sequence homology. Although the AI obtained in this study was established from reads with homology to beta-xylosidases, this hit could also include GHs which have no beta-xylosidase activity but share sequence similarities with beta-xylosidases. Conversely, hits identified based on homology may also exclude novel sequences with beta-xylosidase activity. In addition, since reads are only small fragments of the gene, homology search using this strategy can also exclude genuine beta-xylosidase sequences which have homology below the minimum search parameters used. However, this method is useful to estimate the probability of isolating the gene of interest when PCR-based or functional based screening is to be used.

The abundance of beta-xylosidase reads also confirms that the HMM material is rich in genes encoding hemicellulose degrading enzymes. These enzymes are produced by all major bacterial phyla: *Actinobacteria*, *Bacteroidetes*, *Firmicutes*, and *Proteobacteria* (Perez *et al.* 2002). Over 20% of the reads show homology to proteins derived from thermophilic bacteria (red bars in figure 20), which is in line with the thermophilic nature of this sample.



FIGURE 20: Beta-xylosidase reads obtained from HMM. The data was analysed through CLC workbench by comparing sequence reads to the library of beta-xylosidases using a maximum e-value of 1e-5, a minimum identity of 60 %. Only 35 reads are shown.

These include Xyl3A from *Fervidobacterium gondwanense*, Xyl3A from *Caldanaerobius polysaccharolyticus* and beta-xylosidase from the hyperthermophile *Thermotoga maritima* MSB8.

2.2.5.1. Beta-xylosidase ORFs

Complete beta-xylosidase ORFs were also analysed from assembled contigs. Twenty-seven full ORFs with similarity to beta-xylosidase were identified through homology prediction (Table 8). These represent one ORF per 33Mb of HMM. In a switchgrass-adapted compost community metagenome, Dougherty *et al.* (2012) identified twenty-two putative carbohydrate degrading ORFs, four of which were beta-xylosidase. While the authors did not indicate the size of the metagenome, the number of beta-xylosidase ORFs was significantly lower than what was obtained in this study.

Almost all ORFs identified were from the GH43 and GH3 families. This is in agreement with the analyses of GHs abundances (Figure 19) where these two families were identified as the most dominant sequences. As expected, the majority of these ORFs have closest identity to proteins from the *Bacteroidetes* phylum. These include glycoside hydrolase from *Spirosoma spitsbergense*, a glycoside hydrolase family 3 protein from *Flammeovirgaceae bacterium* 311, alpha-N-arabinofuranosidase from *Flectobacillus major*, glycosyl hydrolase family 3 from *Sunxiuqiniadok donensis*, glycoside hydrolase family 3-domain protein from *Gillisia limnaea*, etc. Other ORF hits include proteins with similarity to proteins from cyanobacterium *Hassallia byssoidea*, and thermophilic bacteria such as *Caldanaerobius polysaccharolyticus*, and *Thermopagus xiamenensis*. Only two ORFs have similarity to a characterised GH43 protein (Matsuzawa *et al.* 2015) from uncultured bacteria (highlighted in green) while the rest of the sequences are from genome sequencing projects. The majority of genes do not show sequence novelty, which is typical of sequence based genes discovery.

Table 8: Predicted beta-xylosidases encoding ORFs from HMM

Contig	Highest identity at NCBI database	Accession number	Identity %	Family
363	glycoside hydrolase <i>Spirosomaspitsbergense</i>	WP_020607374.1	74	GH3
963	hypothetical protein <i>Segetibacterkoreensis</i>	WP_018611809.1	63	GH3
1152	beta-glucosidase <i>Flexithrixdorothaeae</i>	WP_020528605.1	100	GH3
2113	alpha-N-arabinofuranosidase <i>Sediminibacter</i> sp. Hel_I_10	WP_051605581.1	92	GH43
2139	glycoside hydrolase family 3 protein <i>Flammeovirgaceae bacterium</i> 311	AHM63589.1	99	GH3
2202	glycoside hydrolase family 43 uncultured bacterium	BAS02080.1	90	GH43
2653	beta-glucosidase-like glycosyl hydrolase <i>Flammeovirgaceae bacterium</i> 311	AHM61425.1	64	GH3
3178	arabinan endo-1,5-alpha-L-arabinosidase <i>Draconibacterium orientale</i>	AHW60264.1	60	GH43
4595	alpha-N-arabinofuranosidase <i>Flectobacillus major</i>	WP_026998094.1	100	GH43
6904	glycoside hydrolase family 43 uncultured bacterium	BAS02080.1	90	GH43
7257	glycosyl hydrolase <i>Hassalliabysssoidea</i>	WP_039749676.1	82	GH43
8059	hypothetical protein <i>Nafulsellaturpanensis</i>	WP_017733633.1	100	GH3
8335	glycoside hydrolase <i>Thermophagusxiamenensis</i>	WP_010528840.1	90	GH43
8540	alpha-L-arabinofuranosidase <i>Flammeovirgaceae bacterium</i> 311	AHM62617.1	92	GH43
8711	xylosidase/arabinosidase <i>Flammeovirgaceae bacterium</i> 311	AHM62251.1	99	GH43
9012	xylan 1,4-beta-xylosidase <i>Flammeovirgaceae bacterium</i> 311	AHM62615.1	59	GH3
1069	xylan 1,4-beta-xylosidase <i>Flammeovirgaceae bacterium</i> 311	AHM62615.1	66	GH43
1070	1,4-beta-xylanase <i>Pedobacteroryzae</i>	WP_051189348.1	100	GH3
1087	glycosyl hydrolase <i>Hassalliabysssoidea</i>	WP_039749676.1	70	GH3
1115	glycosyl hydrolase family 3 <i>Sunxiuqiniadokdonensis</i>	KOH44958.1	82	GH3
1327	1,4-beta-xylanase <i>Pedobacteroryzae</i>	WP_051189348.1	88	GH43
1683	glycoside hydrolase family 3 domain protein <i>Gillisialimnaea</i>	WP_006989015.1	63	GH3
2281	1,4-beta-xylanase <i>Pedobacteroryzae</i>	WP_051189348.1	100	GH3
2874	glycoside hydrolase family 3 <i>Paenibacillus</i> sp. FSL R7-0273	WP_039873384.1	66	GH3
3397	glucan 1,4-alpha-glucosidase <i>Teredinibacter</i> sp. 1162T.S.0a.05	WP_045820119.1	92	GH43
3869	alpha-N-arabinofuranosidase <i>Rufibacter</i> sp. DG31D	AKQ44444.1	90	GH43
2670	Xyl3A <i>Caldanaerobiuspolysaccharolyticus</i>	AFM44649.1	92	GH3

2.2.6. PCR screening for beta-xylosidase encoding genes

PCR-based screening of glycoside hydrolases from metagenomic DNA has mainly been applied to cellulases and lipases/esterases (Bell *et al.* 2002). As far as we know, there are no reports of beta-xylosidase isolation from mixed microbial genomes through this approach. In this section, low stringency primers were designed from conserved domains of four glycoside hydrolases families where beta-xylosidases have been reported. These primers were used to isolate beta-xylosidase encoding genes from HMM.

2.2.6.1. Degenerate primer design

Low stringency primers were designed targeting glycoside hydrolases family 3, 39, 43 and 52 as described in 2.2.8. These families were selected because they contain most of characterised beta-xylosidases whose sequences are available in the databases. Table 9 shows three conserved motifs which were identified by the MEME server from each GH family. Low stringency primers were designed from highlighted regions because they show low degeneracy and annealing temperatures.

Table 9: Motif positions in the aligned sequences

Family	Motif	Sequence	Motif stat	Motif end	Motif size
GH3	A1	YGLTYWSPNINIFRHPRWGRGQ ETPGED	160	195	35
	A2	NHSRHHFNMVITQQDLEETYQPPFKVCVRDGVH SVMCSYNQVNGVPCCA	235	285	50
	A3	INAILWM WYPG QAGGQAIADIIFGKHNPGRPLPMT	557	586	27
GH39	B1	HITEYN TSYSPINPVHDTALNAAIARILSEGGDYVDSFSYWTFSD VFEE	275	325	50
	B2	NWKFCVGT GRLGL ALQKEYLDHLKLVQEKIGFRYIRGHLLCDDVGIYRE	17	67	50
	B3	YNFTYIDRIVDSYLALNIRPFIEFGMPKALASGDQTVFYW	77	118	41
GH43	C1	NPVIPGFHPDPSICRVGDDYLVNSSFQYFPGVPIFHSKDLVN WEQIGH C	26	76	50
	C2	EGPHIYKKGWYYLMIAEGG TEYGHM VTIARSRNIYGPYESCP	190	233	43
	C3	GIYAPTIRYHDGMFYMITTNV	91	112	21
GH52	D1	SYYGNTQLLEQEGKPIWVVNEGEYRMMNTFDLTVDQLFFELKMNPWTVKN	337	387	50
	D2	FTHD MGVANTFSRPHYSAYELYGIDGCFSHMTHEQLVNWVLCAAVYIEQT	416	466	50
	D3	MPKNMFFNAHHSVPVAFASFTLGFPGKSGGLDLELGRPPRQNVYI GVQSL	24	74	50

2.2.6.2. PCR screening based for beta-xylosidases from HMM

Purified metagenomic DNA was used as template in low stringency PCR amplification with the degenerate primers (Figure 21).

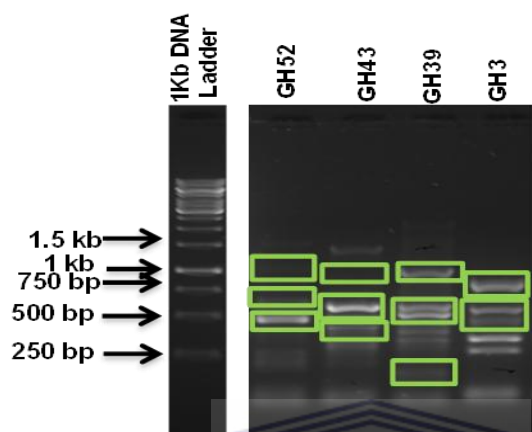


FIGURE 21: Agarose gel electrophoresis image showing low stringency PCR amplicons. Purified HMM was used as template. Amplicons were viewed on 1% agarose gel which was electrophoresed at 100V for 45 minutes.

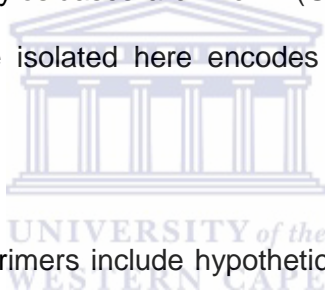
While expected amplicon size ranges could be identified, the multiple amplicons suggest that there is non-specific priming which is typical of low stringency primers when metagenomic DNA is used as template. Three amplicons per GH family primer set (in light green boxes) were excised, purified from the gel, cloned and transformed in *E. coli* to create twelve amplicon mini libraries (Library A-C per GH family) each containing between 52 and 82 clones (average of 73 clones per library). Plasmid DNA from two clones in each library (Library A-C, six clones per GH family) were isolated and the sequence determined by the Sanger method. The resulting sequences were analysed to identify the closest homologues (Table 10).

Table 10: NCBI blastx results of sequences obtained from PCR-based screening

Primer family	Highest similarity in NCBI database	Query coverage (%)	Identity percentage (%)	Accession
GH3	Glyco_hydro_3_C uncultured <i>Bacteroides</i> sp.	66	92	AIA93952.1
GH39	TonB-dependent receptor <i>Bacteroides fragilis</i>	72	96	AKA50915.1
GH39	putative cell wall-associated hydrolase <i>Acinetobacter baumannii</i> 1428368	63	90	EYU52371.1
GH39	conserved hypothetical protein <i>Escherichia coli</i> APEC O1	62	92	ABJ01621.1
GH43	glycosyl hydrolases 43 family protein, partial <i>Bacteroides fragilis</i> str. 3397 T10	63	94	EXY32071.1
GH43	hypothetical protein PROSTU_00109 <i>Providencia stuartii</i> ATCC 25827	66	92	EDU61869.1
GH43	cell wall-associated hydrolase <i>Kosakoniara dicincitans</i> DSM 16656	70	96	EJI92551.1
GH43	conserved hypothetical protein <i>Persephonella marina</i> EX-H1	62	93	ACO04432.1
GH52	TPR repeat-containing protein <i>Chitinophaga pinensis</i> DSM 2588	63	94	ACU59818.1
GH52	putative membrane protein <i>Burkholderia</i> sp. RPE67	71	96	BAO90347.1
GH52	hypothetical protein CRE_03576 <i>Caenorhabditis remanei</i>	70	96	EFO92753.1
GH52	hypothetical protein <i>Clostridium cellobioparum</i>	62	90	WP_027629854.1

GH3 primers were able to isolate sequence with highest similarity to Glycosyl hydrolase family 3 from an uncultured *Bacteroides* sp. GH43 primers could amplify a sequence with highest similarity to glycosyl hydrolases family 43 protein from *Bacteroides fragilis* str. 3397 T10. It was shown in section 2.3.1.4 that GH3 and GH43 sequences were amongst the most abundant hydrolases in the HMM sequence. This could be an indication of the efficiency of this screening method in isolating abundant sequences.

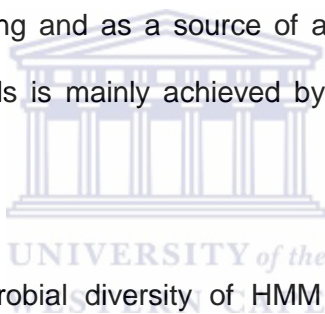
Although GH39 primers did not amplify sequence that could be annotated as beta-xylosidase or glycoside hydrolases for that matter, they were able to amplify a putative cell wall associated hydrolase. The GH39 family is represented by only two hydrolases based on sequence similarities: alpha-L-iduronidase (EC 3.2.1.76) and beta-xylosidase (EC 3.2.1.37). While cell wall associated beta-xylosidases are known (Goujon *et al.* 2003), it is difficult to conclude whether the sequence isolated here encodes for beta-xylosidase or glycoside hydrolase from family 39.



Sequences amplified by GH52 primers include hypothetical proteins which requires further investigation to understand if they encode proteins with hydrolase activity. As mentioned above, the GH52 family is one of the smallest families of hydrolases which is represented by beta-xylosidases only. Enzymes in this family are difficult to identify through a PCR method since there are few annotated sequences available. The inefficient isolation of GH39 and GH52 sequences could be associated with their abundance in the sample or the high sequence variation in these families of proteins. These two families were not identified in the shotgun sequence data of uncloned metagenomics DNA. Gonzalez *et al.* (2012) also showed that PCR-based method biases towards the most abundant sequences, while rare sequences are likely to be left unidentified.

2.3. Conclusion

Metagenomic analysis is an extremely powerful approach for investigating the microbial ecology of diverse environments, and a useful tool for accessing genetic diversity for applications in biotechnology (Diaz-Torres *et al.* 2006). The diversity of mixed microbial genomes can provide an important source for novel enzymes with different properties (Singh & Macdonald 2010). Environments where biomass turnover rates are high at elevated temperatures, such as compost, represent ideal sources for novel lignocellulose degrading enzymes. Lignocellulose constitutes the major component of plant biomass and is the most abundant renewable organic resource globally (Lange & Solutions 2007). This biomass is used in various bioconversion processes with potential applications in the production of ethanol, for paper manufacturing and as a source of animal feed (Howard *et al.* 2003). Biodegradation of these materials is mainly achieved by enzymes produced by microbial populations (Hess *et al.* 2011).



This chapter assessed the microbial diversity of HMM and their metabolic capacity in degrading lignocellulose using Miseq sequencing platform. While the Hiseq technology could have improved the sequencing depth, the objective of this chapter was to assess microbial and GHs diversity of HMM, which can be achieved through Miseq platform. However, when metagenomic bioprospecting is the objective, the Hiseq platform is most suitable since the high depth is important to identify rare genes.

Bacteroidetes was found to be the most abundant phylum. This is in line with other studies, although some studies showed *Proteobacteria* as the most abundant phylum. This difference is most probably due to the type of compost samples, where organic compost manure is dominated by *Proteobacteria* while animal manure, as used in this study, is dominated by *Bacteroidetes*.

Assessment of GHs in this metagenome revealed a wide diversity of GH reads. The most abundant reads were associated with GH2, GH3, GH9 and GH97 families, suggesting that the microbial communities of composting horse manure are well adapted for polysaccharide degradation, particularly hemicellulose and cellulose. Eight hundred and seventeen reads with homology to beta-xylosidase encoding genes were predicted to be present in this metagenome, which translates to 1.01 beta-xylosidase read per megabase of metagenome. Twenty-seven ORFs with homology to beta-xylosidase genes were predicted from assembled reads. While most of these ORFs were from the *Bacteroidetes* phylum, ORFs homologous to proteins from thermophiles and actinobacteria were also identified.

None of the beta-xylosidase ORFs matched the sequences amplified using the newly designed degenerate primers. However, PCR screening was able to isolate GH3 and GH43 proteins, showing that the primers were specific to conserved regions from these families, while a hydrolase isolated by GH39 primers could not be assigned to this family with absolute certainty. GH52 primers selected mainly sequences annotated as hypothetical proteins. Given the scarcity of protein from this family, sequence based identification is not ideal since few proteins from this family have been sequenced. The hypothetical and putative sequences which were isolated by GH52 primers can only be confirmed through activity screening to assess whether they encode for active beta-xylosidase or glycoside hydrolase proteins.

This general evaluation of composting horse manure adequately demonstrates that this material is suitable for functional assessment of lignocellulase encoding genes. The following chapters employ classical functional based screening and a novel high throughput cell-free methods to isolate beta-xylosidases encoding genes from this metagenome, and compare to what was identified through a sequence-based screening.

CHAPTER THREE: Traditional functional screening for beta-xylosidases from a horse manure metagenomic fosmid library.

3. Introduction

Functional screening of metagenomic libraries is the most widely used technique for novel gene discovery. This method has been applied in the discovery of many novel enzymes and other secondary metabolites of medical and industrial importance. Common target enzymes in metagenomic studies are predominantly biocatalysts such as acylases, phosphatases, proteases, oxidoreductases, glycosyl hydrolases and lipases/esterases. Compared to cellulases and esterases, the discovery of novel beta-xylosidases through function-based metagenomic screening are limited. Perhaps the most interesting beta-xylosidase from function-based screening is that reported by Ferrer *et al.* (2012). After screening of a cow rumen metagenome, Ferrer and co-workers identified 15 hydrolases, including a GH43 family protein, the first multi-functional enzyme to exhibit beta-1,4 xylosidase; alpha-1,5 arabinofur(pyr)anosidase; beta-1,4 lactase; alpha-1,6 raffinase; alpha-1,6 amylase; beta-galactosidase and alpha-1,4 glucosidase activities. This enzyme has great potential in the food and biofuel industries.

One of the major limitations of function-based metagenomics is that the overall hit rates are usually low. Hit rates have been shown to range from one positive hit per 2.7Mb to one hit per 3979.5Mb of DNA screened (Uchiyama & Miyazaki 2009). Based on the assessment of the most popular activity screens described in the literature, Ferrel *et al.* (2016) estimated the mean incidence rate of positive clones when performing a naïve screen in the environmental clone libraries. These estimations are summarized in Figure 22. The hit rate for acylases was estimated to be 1 active clone per 333 total clones, 1 active clone in 2843

clones for phosphatases, 1 oxidoreductases in 6670 clones, 1 proteases in 9388 clones, 1 esterase/lipases in 17 320 clones, and 1 glycosidases in 31 190.

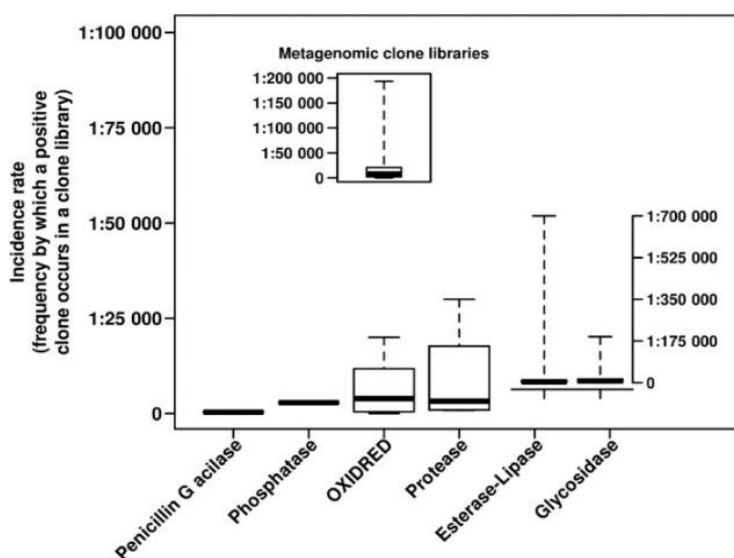


FIGURE 22: Incidence rate of the positive according to the enzyme activity. The results are based on the values for the metagenomic studies related to the top six activities commonly identified by naïve screens independent of the substrate used. The inset represents the mean incidence rate for all enzymes. Adapted from Ferrer *et al.* (2016).

The low hit rate in metagenomic screening is due to several factors including the source of the metagenomic DNA, the abundance of the gene of interest in the metagenome and subsequently the library, the vector system and host of choice, the screen itself, and the ability of the host to successfully express the gene (Uchiyama & Miyazaki 2009).

Studies elsewhere employed NGS-based screening prior to functional screening to determine the abundance of a specific gene or gene family. For instance, Hess *et al.* (2011) first sequenced and analysed 268Gb of cow rumen metagenomic DNA for carbohydrate active enzymes. They identified 27,755 putative carbohydrate-active genes and expressed 90 candidate proteins, of which 57% were enzymatically active against cellulosic substrates.

Like Hess and co-workers, the second chapter of this study evaluated the abundance of glycoside hydrolases with specific interest to beta-xylosidase encoding sequences in HMM.

Here, functional screening of beta-xylosidases was done to determine the efficiency of this method in isolating novel beta-xylosidase enzymes from composting horse manure. The ultimate objective was to compare both sequence-based and classical function-based screening platforms with the mIVC-FACS method.

3.1. Materials and methods

3.1.1. Metagenomic library construction

A metagenomic DNA library was constructed using the CopyControl™ FosmidLibrary production kit (EpicentreR Biotechnologies) according to the manufacturer's guidelines. Briefly, extracted high-molecular weight metagenomic DNA (2.2.3) was end-repaired using End-It™ DNA endrepairkit (EpicentreR). End-repaired DNA was extracted by Phenol:Chloroform:Isoamyl followed by EtOH/Na-Acetate precipitation overnight at -20°C. The end-repaired DNA was further purified as indicated in 2.2.3. Following purification, the DNA was ligated into the CopyControl pCC1Fos™ vector. Fosmid clones were packaged *in vitro* using MaxPlax™ Lambda phage. For transfection in *E. coli* EPI300-T1R, 5ml of an overnight culture was inoculated into 50mL LB medium, supplemented with 10mM MgSO₄ and incubated at 37°C shaking to an OD₆₀₀ of 0.8. MaxPlax™ Lambda packaging extract was thawed on ice and 25µl transferred to a pre-chilled sterile micro centrifuge tube. Ten micro litres of the ligation reaction was added to the thawed extract and mixed carefully by pipetting, followed by incubated at 30°C for 90 min. A further 25µl of packaging extract was added and the reaction mixture was incubated at 30°C for 90 min. Phage dilution buffer (10mM Tris-HCl, pH 8.3; 100mM NaCl; 10mM MgCl₂) was added to a final volume of 1ml. Finally, 25µl of chloroform was added and stored at 4°C. Transfection was initiated by adding 10µl of various dilutions of the packaged phage to 100µl of *E. coli* EPI300-T1R cells. The reaction was incubated at 37°C for up to 2 hours. Transformants were selected by plating the mixture onto LB agar containing chloramphenicol (12.5µg/ml) with overnight

incubation at 37°C. The library titre was determined from transformation dilutions and aliquots of 1ml cells with 1000 transformants were stored at -80°C until use.

3.1.2. Evaluation of insert size of the metagenomic library

A total of 20 clones were randomly selected from the fosmid library. Fosmid DNA was extracted using the Qiagen miniprep kit according to manufacturer's recommendations. Isolated fosmids were digested with *HindIII* and *EcoRI* (Fermentas) for 16 hours at 37°C in a 20µl reaction containing 2µl of Buffer R, 1U *HindIII*, 1U *EcoRI*, and 1µg of fosmid DNA. The digestion reaction was electrophoresed in a 1% agarose gel at 80V to estimate the average insert sizes. DNA fragments were visualised with an Alphamager 3400 imaging system (Alpha Innotech, USA).



3.1.3. High-throughput library picking and screening

One thousand clones were plated per Q-tray (Genetix) containing 300ml of LB amp agar supplemented with 12.5µg/ml chloramphenicol. The plates were incubated overnight at 37°C. A Genetix QPix2 automated colony picker was used to inoculate individual clones into 96well microplates, with each well containing 50µl LB broth supplemented with 12.5µg/ml chloramphenicol, 2.5mg/ml pNPX (4-Nitrophenyl beta-D-xylopyranoside) and 0.02% L-arabinose for fosmid copy number amplification. The microplates were sealed with breathable sealing membrane (Sigma, USA) and cultured at 37°C overnight with shaking. Positive clones were identified by change of LB broth colour to orange. After identification of positive clones, up to 20% glycerol was added and the plates were stored at -80°C.

3.1.4. Secondary screening of positive clones

Positive clones were re-inoculated into LB agar supplemented with 12.5µg/ml chloramphenicol and grown overnight at 37°C. A single colony from the overnight culture was inoculated into 5ml of LB broth supplemented with 12.5µg/ml chloramphenicol and 0.02% L-arabinose to increase fosmids copy number. The culture was grown overnight at 37°C, followed by transferring 10µl (adjusted to OD₆₀₀=2.5 using sterile distilled water) into 100µl LB broth supplemented with 12.5µg/ml chloramphenicol, 0.02% L-arabinose and 2.5mg of pNPX, 4-Nitrophenyl-beta-D-glucopyranoside (pNPG), 4-Nitrophenyl alpha-L-arabinofuranoside (pNPA) or 4-Nitrophenyl beta-D-cellobioside (pNPC) in a 96 well plate. Positive activities were identified by reading the plate on a spectrophotometer at the wavelength of 405nm.

3.1.5. Fosmids extraction, sequencing and analysis

Fosmid DNA was extracted using the Qiagen mini prep kit as per manufacturer's instruction. Sequencing of selected fosmids and assembly was done as described in 2.1.4. Sequence reads were referenced against pCC1FOS™ vector before assembly to remove backbone sequence. Identification of beta-xylosidase ORFs was done in the NCBI blastx tool. Sequence analysis was done in NCBI using the Smartblast tool. This tool processes the query in such a way that presents the three best matches from the non-redundant protein sequence database along with the two best protein matches from well-studied reference species. Sequence alignments were done using Aliview sequence alignment viewer (Larsson 2014). All gene sequences were translated into protein ORFs using the ExPasy bioinformatics resource portal (Artimo *et al.* 2012).

3.2. RESULTS AND DISCUSSIONS

3.2.1. Verification of metagenomic library

Titration of the library revealed that it contained approximately 20000 clones. Restriction digestion of 20 randomly selected fosmid clones revealed that the library has an average insert length of over 23kb (Figure 23), totalling 460Mb (± 100 bacterial genomes) of DNA contained in the library.

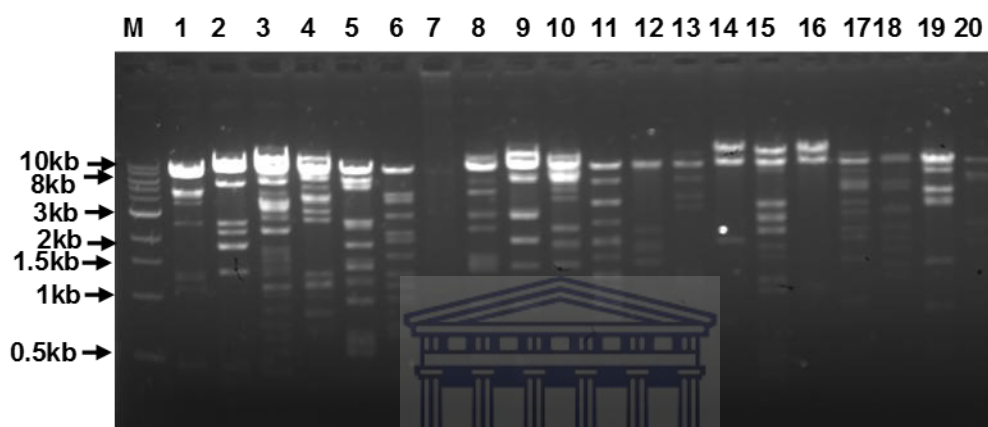


FIGURE 23: Restriction digest of 20 randomly selected fosmid clones. Fosmids were digested with *HindIII* and *EcoRI* (Lanes 1-20). Lane designated **M** contain 1kb DNA marker (NEB). The Fosmid vector backbone is clearly present in each clone at approximately 8.1 kb.

3.2.2. Screening for beta-xylosidase activity

A total of 26 positive hits were identified from the plate screening for beta-xylosidase activity after screening 20000 colonies, corresponding to an average hit rate of one hit per 769 clones, or one hit per 17.7Mb (one hit per 4 bacterial genomes). These hits were further exposed to a secondary screening with three additional substrates (Figure 24). Bastien *et al.* (2013) reported extremely high global hit rates in two fosmid libraries of whole termite abdomens and fungal-comb material. They screened for the presence of xylan-degrading enzymes and found 101 positive clones, corresponding to an average hit rate of 3 hits per 1000 clones. Most of these clones displayed either beta-xylosidase or alpha-L-arabinofuranosidase activity. While that study was not exclusively on beta-xylosidase, the hit

rate presented in our study only shows clones with beta-xylosidase activity, which may suggest a better hit rate than that obtained in the Bastien *et al.* study.



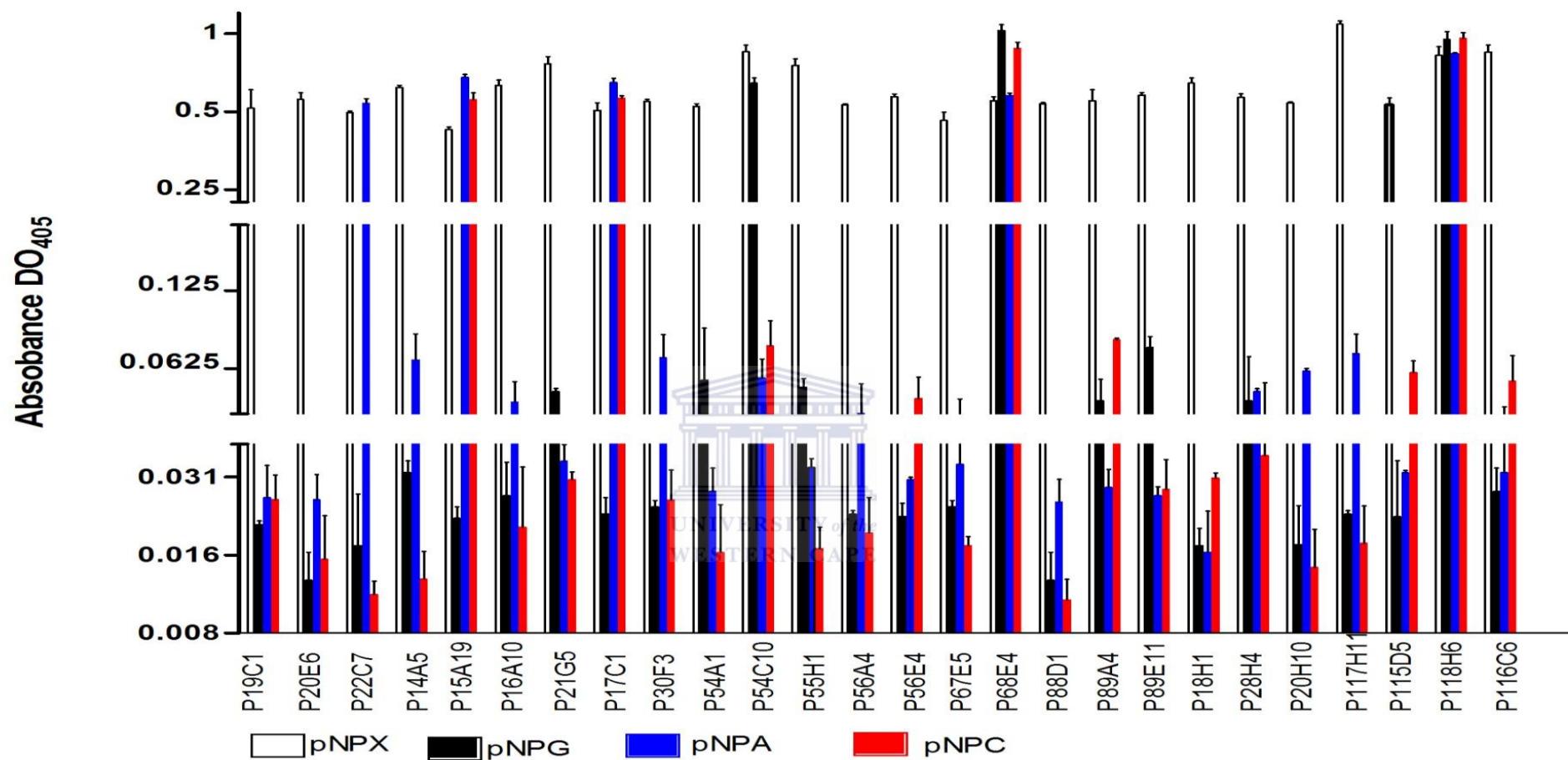


FIGURE 24: Activity analysis of positive fosmid clones on four chromogenic substrates. The plotted values were first adjusted by subtracting values obtained from the negative controls. The substrates used for this assay were: 4-Nitrophenyl β -D-xylopyranoside (pNPX), 4-Nitrophenyl β -D-glucopyranoside (pNPG), 4-Nitrophenyl β -D-cellobioside (pNPC) and 4-Nitrophenyl α -L-arabinofuranoside (pNPA).

In order to establish the degree to which these fosmids may represent duplicated inserts, eleven fosmids of the single activity clones were digested and compared by agarose gel electrophoresis (Figure 25). All the digested clones showed different restriction patterns, suggesting that this library has a low- to no duplication of inserts. However, it should also be noted that randomly sheared metagenomic DNA fragments from up- or downstream regions of a single beta-xylosidase would generate different restriction profiles. While this restriction profiling offers some insight of the extent of duplicates insert in the library, sequencing of these clones would ultimately confirm this.

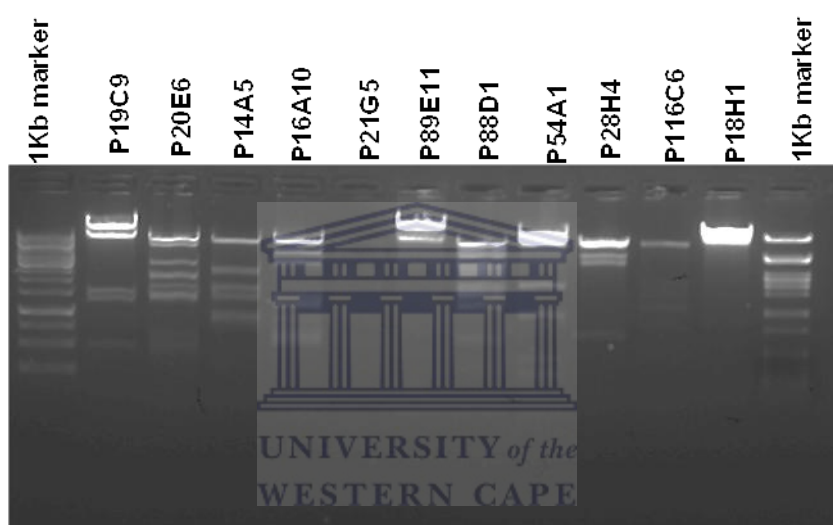


FIGURE 25: Restriction analysis of positive hits. Fosmids were isolated and digested using *HindIII* and *XhoI*. Restriction digests were viewed on 1% agarose gel which was electrophoresed at 80V for 45 minutes.

3.2.3. Sequence analysis of selected positive fosmids showing beta-xylosidase activity

Clone P15A19, P17C1, and P118H6 were three of six clones which showed multiple activities in the secondary assays. These clones were selected for sequencing and further characterised through *in silico* analysis. Although clone P117H11 only showed activity against one substrate, it showed the highest pNPX activity and was therefore also selected

for sequence analysis. The complete table of genes identified from these clones are shown in appendix A.

3.2.3.1. Clone P17C1

Clone P17C1 showed convincing activity on three substrates (pNPX, pNPA and pNPC). Two contigs, P17C1.1 (18kb) and P17C1.2 (15kb) were identified after removal of fosmid backbone. Contig P17C1.1 contained mainly hypothetical protein and a gene cluster which seems to be involved in bacterial redox homeostasis and antioxidant (defence) (Figure 26) (Jones *et al.* 2014). In particular, ergothioneine biosynthesis protein and ornithine carbamoyltransferase have been reported in *Mycobacterium tuberculosis* as an antioxidant gene cluster (Napolitano *et al.* 2008). None of the proteins encoded in this contig seemed to be responsible for any of the activity observed.

Contig P17C1.2 contains ORFs encoding proteins with similarities to hypothetical proteins, metal transport genes and carbohydrate metabolism genes. Three ORFs encode proteins with similarity to alpha-glucuronidases, glycoside hydrolase and alpha-glucosidase. The alpha-glucuronidase in this contig belongs to family 67 glycosidases, which hydrolyse the alpha1,2-glycosidic bond between 4-O-methyl-d-glucuronic acid (4-O-MeGlcA) and the xylan or xylooligosaccharide backbone via the inverting mechanism. Since this enzyme has no known beta-xylosidase, alpha-L-arabinofuranosidase or cellulose activity, it is unlikely to be responsible for the activities observed in Figure 24. The glycoside hydrolase (ORF P17C1-18) and alpha-glucosidase (ORF P17C1-19) found in this contig also show 35% and 41% similarity to beta-xylosidase/alpha-L-arabinofuranosidase (EC: A5JTQ2.1) and probable beta-D-xylosidase (EC: Q9SGZ5.2) respectively when analysed through blastp on the curated Swissprot database. The beta-xylosidase/ alpha-L-arabinofuranosidase which has 41% similarity to ORF P17C1-19 is a bi-functional protein which was characterised by Xiong

et al. (2007), and may explain the activity reported in Figure 24. While the pNPX and pNPA activity of this clone can be attributed to ORF P17C1-18 and 19, the ORF responsible for the pNPC activity is not clear and requires further investigation.



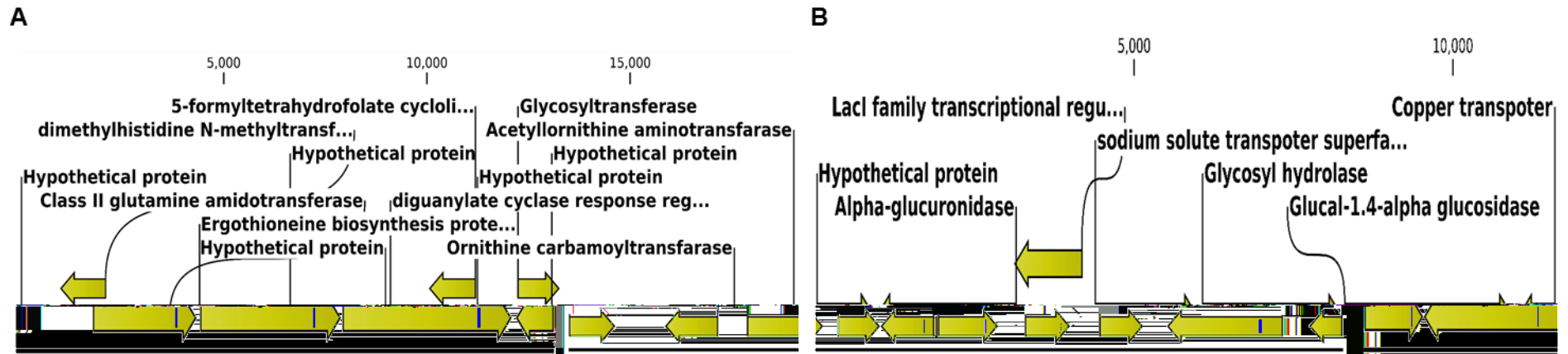


FIGURE 26: Gene organisation of contig P17C1.1 (A) and contig P17C1.2 (B). The gene map was generated and annotated using CLC workbench. The genes were annotated base on the hits obtained from NCBI blastx tool.



3.2.3.1.1. Analysis of conserved residues of ORFs 18 and 19

Both ORF P17C1-18 and P17C1-19 belong to glycosyl hydrolase family 3. This family contains enzymes which catalyse the hydrolysis of a glycosidic bond through a general acid catalysis that requires a proton donor and a nucleophile/base (Chir *et al.* 2002). This reaction involves a pair of residues that act as a catalytic nucleophile (D286) and a catalytic acid/base (E473) (Chir *et al.* 2002). The catalytic nucleophile aspartame was reported to be located within the conserved sequence GFVISDW (Hrmova *et al.* 1998) or VMSDW (Bause & Legler 1974).

Analysis of the catalytic residues of ORF P17C1-18 and P17C1-19 show a typical nucleophilic-acid base mechanism that is characterised by conserved aspartic acid and glutamic acid. However, phylogenetic analysis of these two ORFs suggests that ORF P17C1-19 is genetically different when assessed using the SmartBlast tool in the NCBI server. The server compares the input sequence to three best matches in NCBI blastp databases and to two best matches from well-studied reference species within the database. This could suggest that this ORF encodes a novel protein with beta-xylosidase activity. Further characterisation of this protein, including secondary structure analysis could provide more information about the novelty of this protein.

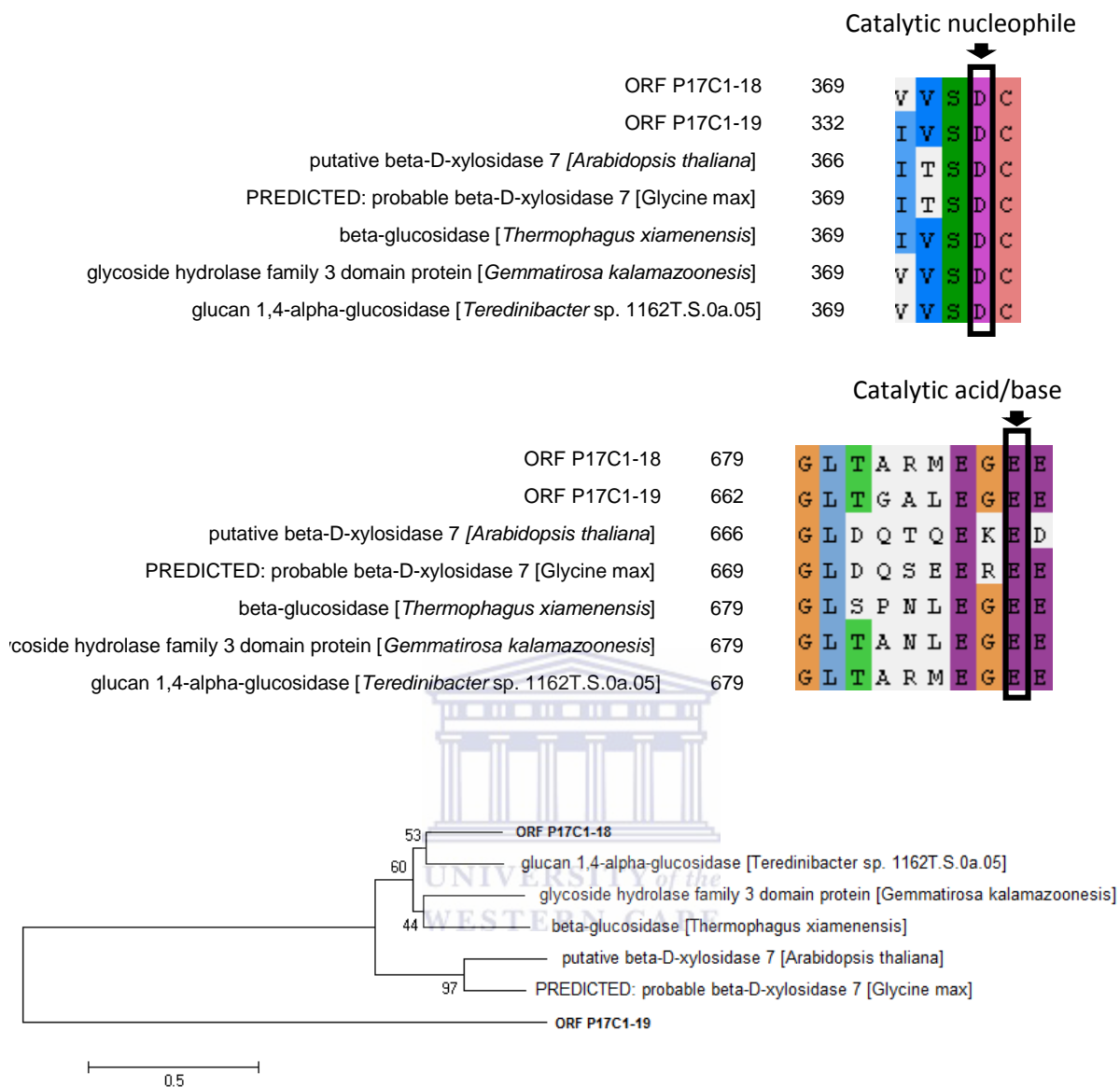


FIGURE 27: Phylogenetic and catalytic analysis of ORF P17C1-18 and P17C1-19. Highlighted blocks show known conserved residues of GH3 proteins. The automated SmartBlast tool compares three best matches in the NCBI proteins database together with the two best matches from well-studied reference species within the database.

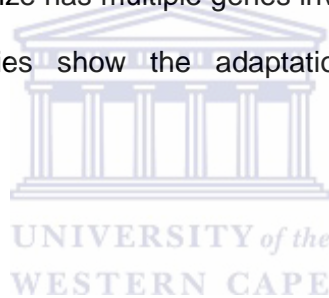
3.2.3.2. Clone P15A9

Two contigs were also identified in clone P15A9 (Figure 28). Contig 15A9.2 has similar ORFs and genetic structure as contig P17C1.2, suggesting that these two clones carry DNA

fragments from the same organism. However, ORF P17C1-18 which was identified in clone P17C1 was not present in clone P15A9. Clone P15A9 has an ORF (ORF P15A9-7) with 38% similarity to a beta-glucosidase (EC: P27034.1) when analysed with the Swiss-prot database. This ORF is absent in clone P17C1.

Beta-glucosidase catalyses the hydrolysis of the glycosidic bonds to terminal non-reducing residues in beta-glucosides and oligosaccharides (Dabek *et al.* 2008). This enzyme is the key component required to complete the final step of cellulose hydrolysis by converting the cellobiose to glucose, the activity which was observed from this clone in Figure 24.

Genetic structure of the consensus fragment of clone P15A9 and P17C1 show that this fragment, which is over 30kb in size has multiple genes involved in carbohydrates hydrolysis (Figure 28C). These capabilities show the adaptation of the source organism in polysaccharides hydrolysis.



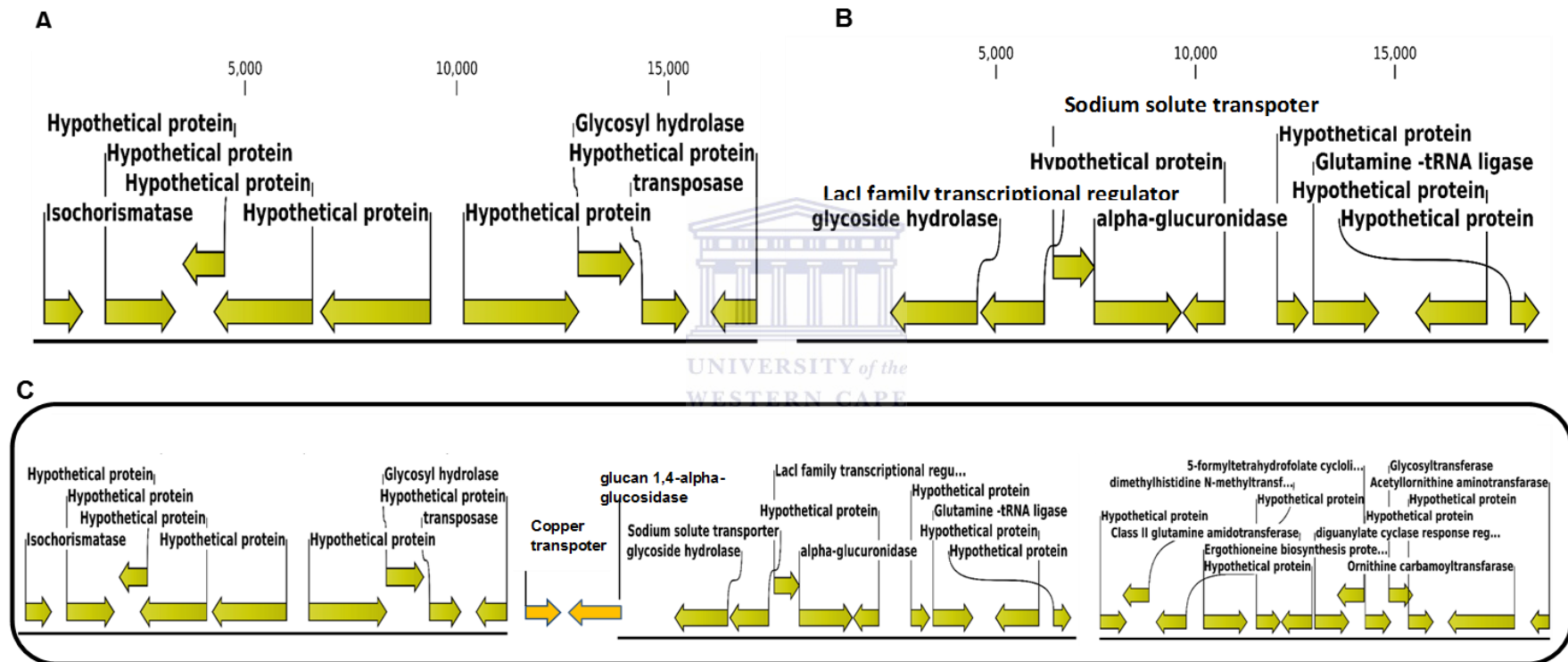


FIGURE 28: Gene organisation of contig P15A9.1 (A), contig P15A9.2 (B) and proposed genetic structure of original insert DNA fragment of clone P17C1 and P15A9 (C). The gene map was generated and annotated using CLC workbench. The genes were annotated base on the hits obtained from NCBI blastx tool.

3.2.3.2.1. Analysis of conserved residues of ORF P15A9-7

Analysis of the catalytic residues of ORF P15A9-7 show the expected residues that has been reported for GH3 proteins (Figure 29) when compared to the closest related proteins.

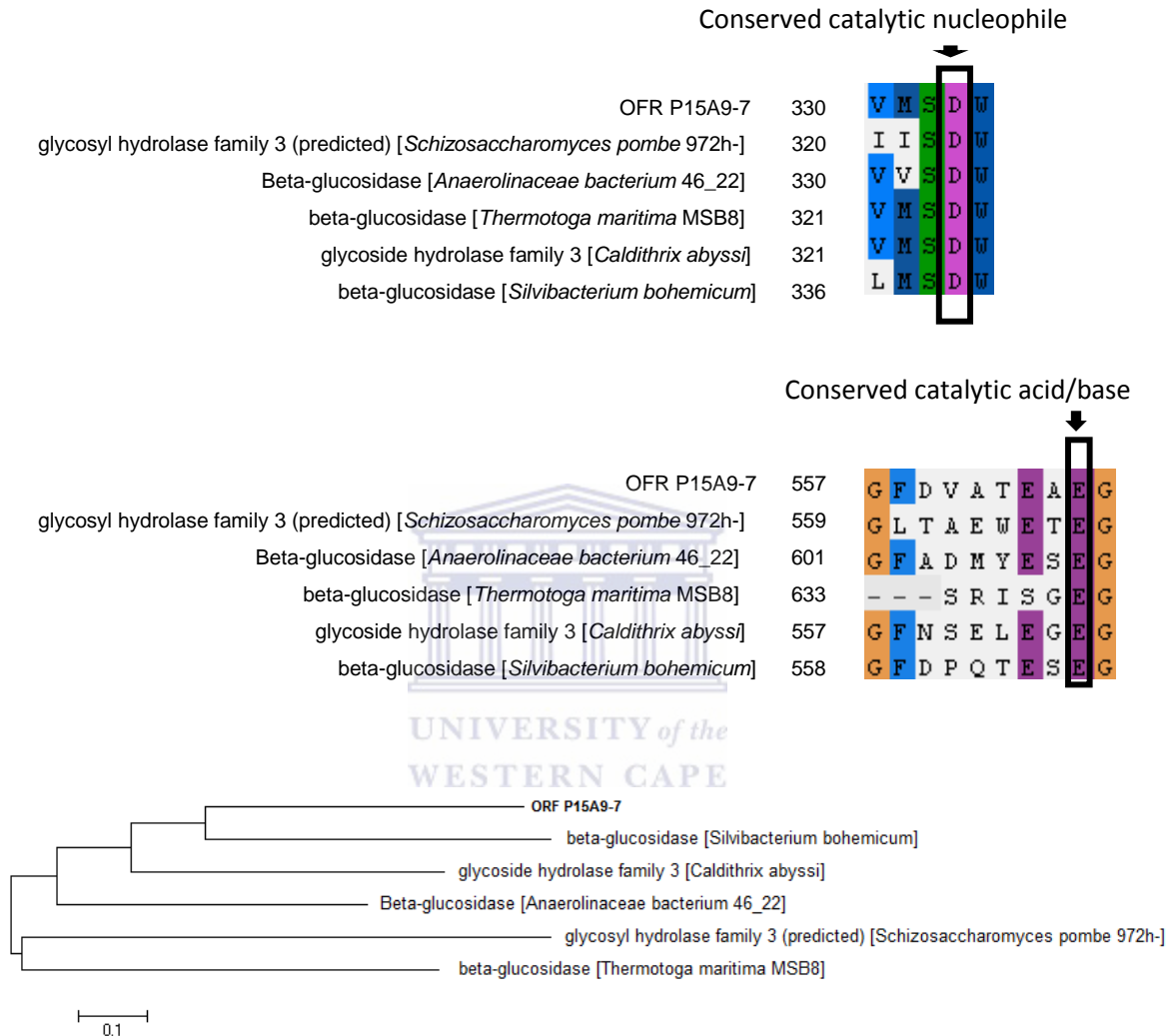


FIGURE 29: Phylogenetic and catalytic analysis of ORF P15A9-7. Residues in black blocks represent conserved catalytic residues of GH3 proteins. The automated SmartBlast tool compares three best matches in the NCBI proteins database together with the two best matches from well-studied reference species within the database.

3.2.3.3. Clone P118H6

Clone P118H6 showed activity on all four substrates tested. However, analysis of contigs obtained from this clone show that it carries one ORF (ORF P118H6-1) with similarity to glycoside hydrolase encoding genes (Figure 30). This ORF is 98% identical to a beta-glucosidase from *Thermobifida fusca* (Appendix A). Although beta-glucosidase which shows xylosidase activity has been reported (Kimura *et al.* 1999) few beta-glucosidases have been shown to act on multiple substrates. In natural environments, beta-glucosidases act synergistically with endo-1,4-beta-D-glucanases in the hydrolysis of the native cellulose into glucose units. The complete degradation of lignocellulose requires the consortium of microorganisms that produce not only cellulases but also hemicellulases and ligninases. In industrial processes, this is achieved through enzymatic cocktails which contain different enzymes that act on different components of lignocellulatic substrates (Zhou *et al.* 2012). The presence of a multi-substrate enzyme which can tolerate industrial processes may greatly improve the efficiency of this process. Indeed, further characterisations of this protein particularly on natural substrates, are required to explore its suitability for industrial processes are necessary.

Most proteins identified from this clone have high similarity to *Thermobifida fusca*, a moderately thermophilic soil bacterium that belongs to *Actinobacteria*. This was not surprising since the metagenome used here was derived from thermophilic material. Moreover, *T. fusca* is a major degrader of plant cell walls and has been used as a model organism for the study of secreted, thermostable cellulases (Gomez del Pulgar & Saadeddin 2014). Analysis of the complete genome sequence revealed the existence of multiple cellulases and xylanases. The glycosyl hydrolases include enzymes predicted to exhibit mainly dextran/starch- and xylan-degrading functions (Lykidis *et al.* 2007).

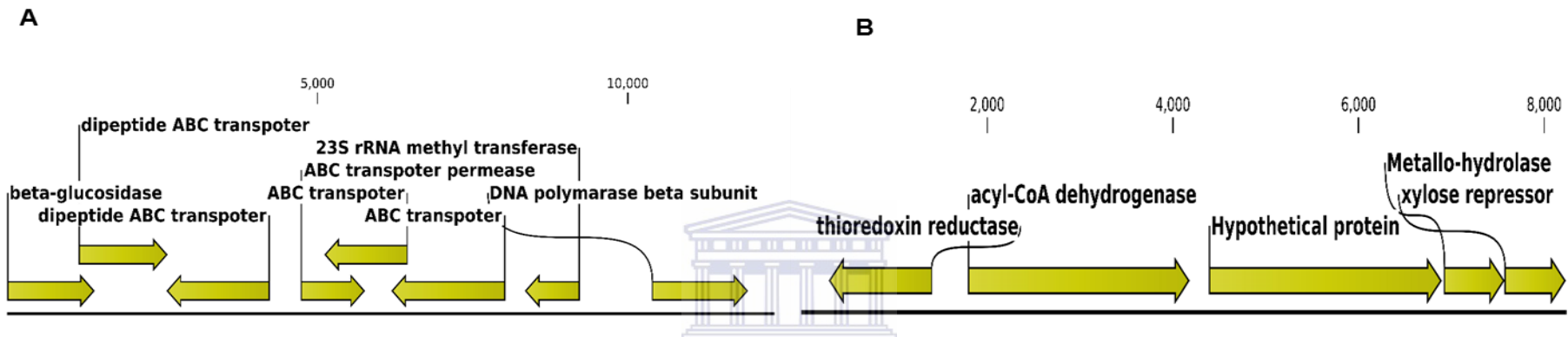


FIGURE 30: Geneorganisation of contig P118H6.1 (A) and contig P118H6.2 (B). The gene map was generated and annotated using CLC workbench. The genes were annotated base on the hits obtained from NCBI blastx tool.

3.2.3.3.1. Analysis of conserved residues of ORF P118H6-1

ORF P118H6-1 harbours a multi-substrate GH1 protein. Yuan *et al.* (2015) also reported a multi-substrate hyperthermophilic glycosidase from *Thermococcus kodakarensis* KOD1 that shows beta-glucosidase, beta-mannosidase, beta-fucosidase and beta-galactosidase activities. Sequence alignment of this protein with other beta-glycosidases from hyperthermophilic archaea showed two unique active site residues, Q77 and D206, which were represented in all other hyperthermophilic beta-glycosidases. When these two sites were mutated to Q77R, D206N and D206Q the secondary structure of all mutants did not change when compared to the wild type. However, Q77R and D206Q mutations were shown to affect the catalytic activity of the enzyme while D206N altered the catalytic turn over rate for glucosidase and mannosidase activities (Yuan *et al.* 2015).

In P118H6-1, the catalytic Q77 was found to be replaced by R77 (Figure 31), suggesting that the catalytic activity of this protein is different to that reported by Yuan *et al.* (2015). Moreover, the catalytic residue D206 was replaced by N206, suggesting that the catalytic turn over rate of this protein is altered compared to that of *T. kodakarensis* KOD1. Although ORF P118H6-1 was only analysed for activity on four substrates, it is possible that this protein might also show beta-mannosidase, beta-fucosidase and beta-galactosidase activities based on the active site residues. However, this is also susceptible to the presence of correct substrate recognition residues.

In another study, Suzuki *et al.* (2012) characterized three GH1 proteins encoded by the thermophile *Geobacillus kaustophilus* HTA426 (GK1856, GK2337, and GK3214). GK3214 was extremely thermostable and retained most of its activity during 7 days of incubation at 60 °C.

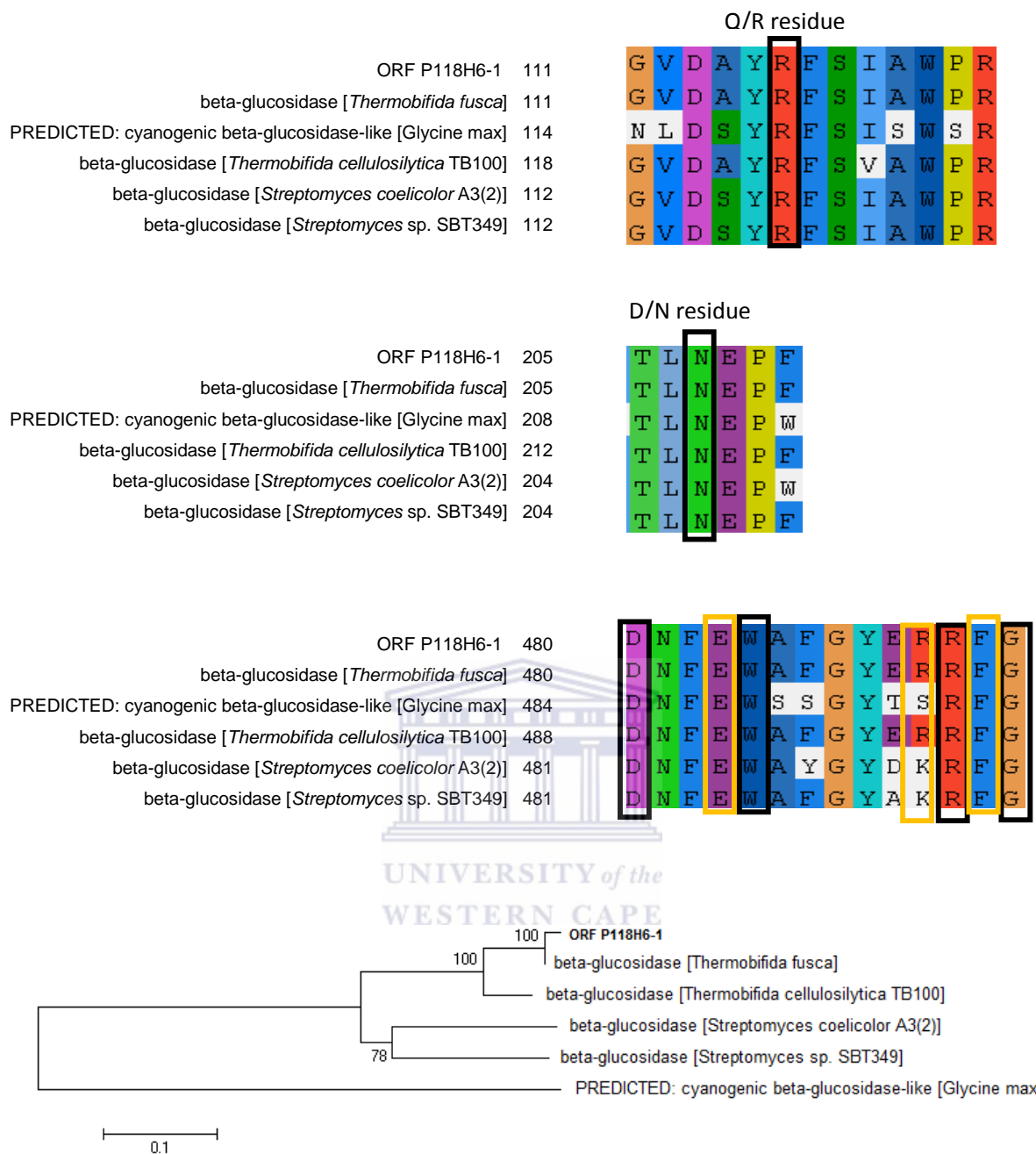


FIGURE 31: Phylogenetic and catalytic analysis of ORF P118H6-1. Residues in black blocks represent catalytic residues of GH1 proteins. Residues in orange blocks show substrate specific residues. The automated SmartBlast tool compares three best matches in the NCBI proteins database together with the two best matches from well-studied reference species within the database.

This protein was also found to have transglycosylation activity, a dimeric structure, and a possible motif that governed its substrate specificity. Substitution of this substrate specific motif (DLLSWLNGYQ-KRYG) with that of a beta-glucosidase from *Paenibacillus polymyxa*

(DNFEWAEGYN-MRFG) resulted in the unexpected generation of a thermostable protein with high specificity to beta-fucosidase and large increases in beta-glucosidase, beta-cellobiosidase, alpha-arabinofuranosidase, beta-mannosidase, beta-glucuronidase, beta-xylopyranosidase, and beta-fucosidase activities but a dramatic decline in 6-phospho- β -glucosidase activity (Suzuki & Okazaki 2013).

The substrate specific residues of ORF P118H6-1 (DNFEWAFGYE-RRFG) are similar to that of *P. polymyxa* (DNFEWAEGYN-MRFG), except that methionine was replaced with arginine. The effect of this substitution needs further analysis of ORF P118H6-1, including expanding the number of substrates to be tested and the effect on thermostability.

3.2.3.4. Clone P117H11

Analysis of the sequence of P117H11 shows that this clone carries an ORF (P117H11-13) with 63% identity to a GH52 beta-xylosidase from *Paenibacillus mucilaginosus* (Figure 32). GH52 is one of the smallest families of hydrolases, with only nine characterised proteins and only two structures available (Cazy accessed December 2015). So far, only beta-xylosidases have been reported in this family, almost all of which are thermophilic. This includes beta-xylosidases isolated from a deep-sea thermophilic bacterium *Geobacillus stearothermophilus* (Huang *et al.* 2014), and *Thermoanaerobacterium saccharolyticum* (Currie *et al.* 2014). These proteins displayed maximum activity at a temperature of 70 °C and pH 5.5.

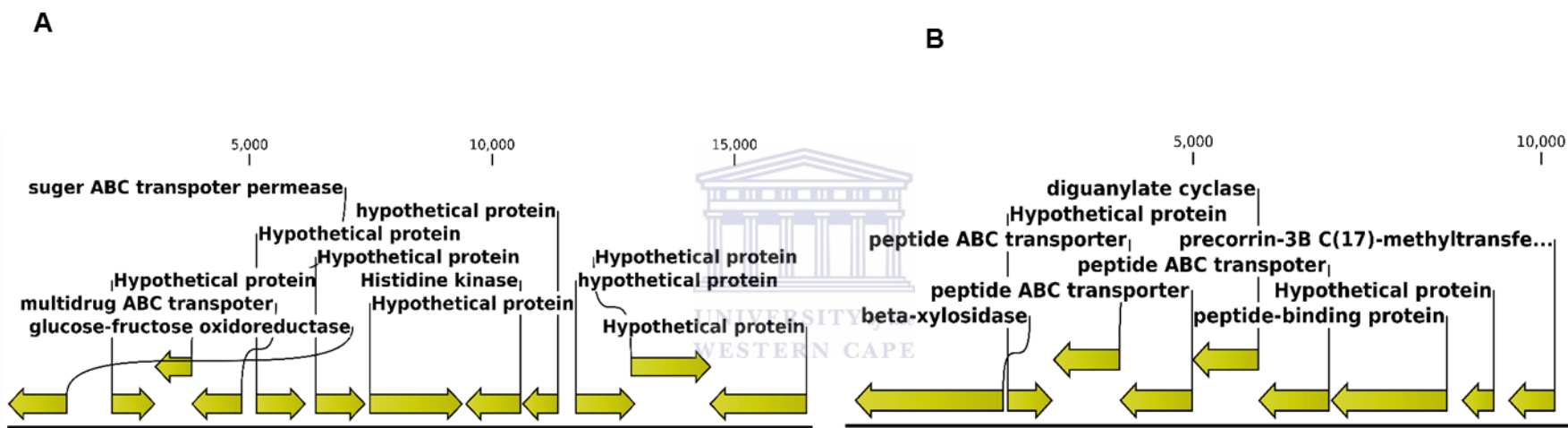


FIGURE 32: Gene organisation of contig P117F11.1 (A) and contig P117F11.2 (B). The gene map was generated and annotated using CLC workbench. The genes were annotated base on the hits obtained from NCBI blastx tool.

3.2.3.4.1. Analysis of conserved residues of ORF P117H11-13

Huang *et al.*(2014) mutated a GH52 xylosidase gene, *gsxyN*, from the deep-sea thermophilic *Geobacillus stearothermophilus* to introduce an exo-xylanase activity while retaining beta-xylosidase activity by substituting Y509 to E509. The optimum xylanase activity of the Y509E mutant displayed at pH 6.5 and 50 °C, and retained approximately 45 % of its maximal activity at 55 °C, pH 6.5 for 60 min. Analysis of ORF P117H11-13 show that this ORF harbours the wildtype tyrosine (Figure 33), which might suggest that it does not have exo-xylanase activity. However, phylogenetic analysis shows that this protein is significantly different to the best related proteins based on sequence similarity.

Analysis of the crystal structure of a novel beta-xylosidase structure from *Geobacillus thermoglucosidasius* showed three catalytic residues: D517, W654 and R715. Whilst the fold of the *G. thermoglucosidasius* beta-xylosidase is completely different from xylosidases in other CAZy families, the enzyme surprisingly shares structural similarities with other glycoside hydrolases, despite having no more than 13% sequence identity. The catalytic residues identified in P117H11-13 were similar to those of *G. thermoglucosidasius* beta-xylosidase (Figure 33). Since only few structures of GH52 beta-xylosidases have been solved, it is difficult to state if all GH52 proteins use these residues for catalysis.

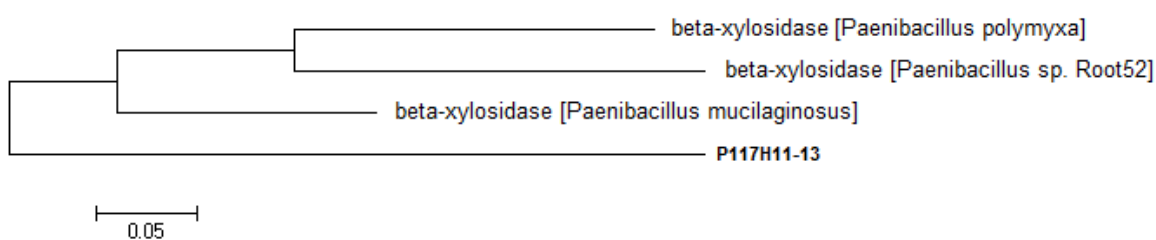
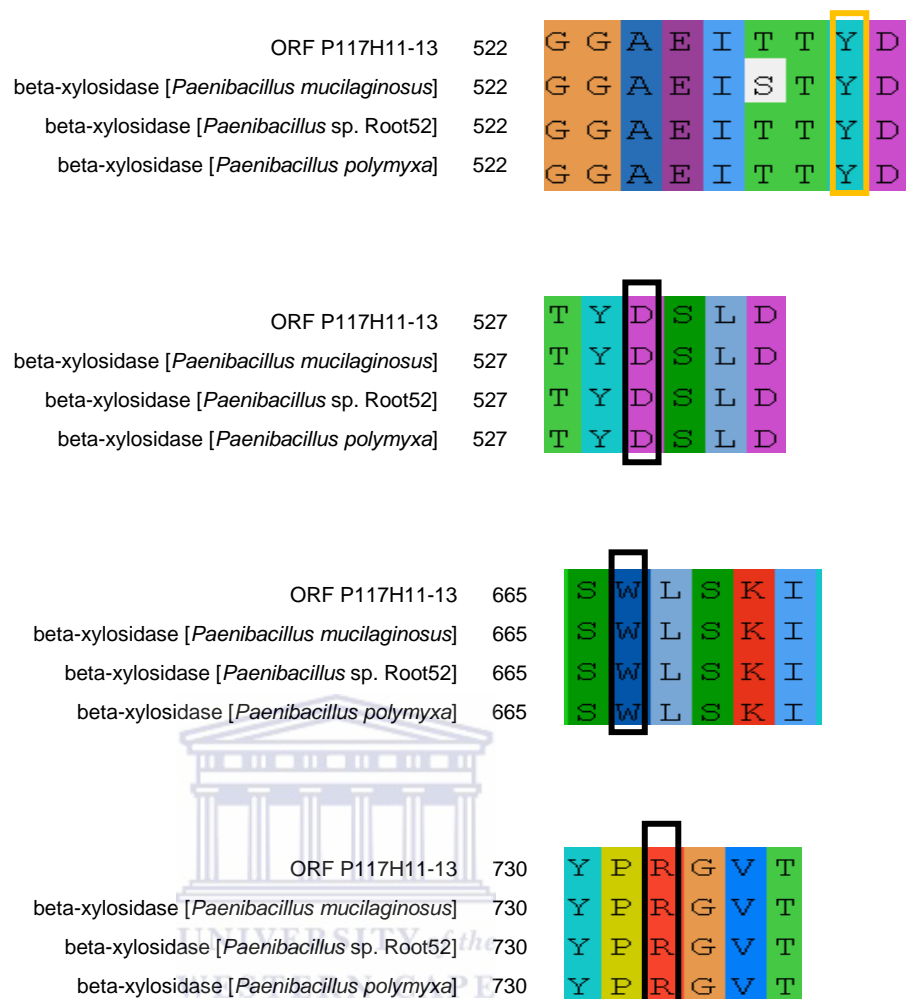
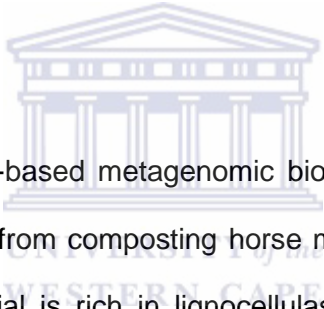


FIGURE 33: Phylogenetic and catalytic analysis of ORF P117H11-13. Residues in black blocks represent catalytic residues of GH52 proteins. Residues in orange blocks show substrate specific residues. The automated SmartBlast tool compares three best matches in the NCBI proteins database together with the two best matches from well-studied reference species within the database.

3.3. Conclusion

The uncertain future of petroleum sources as well as rising cost of fuels have shifted global efforts to utilize renewable resources for the production of greener energy. Significant potential exists for bioconversion of lignocellulose, one of the most abundant and most renewable biomaterials on our planet. However, the requirements of enzyme complexes which act synergistically to unlock and saccharify polysaccharides from the lignocellulose complex to fermentable sugars incur major costs in the overall process and present a great challenge. Therefore, more potent and efficient enzyme preparations need to be developed for the enzymatic saccharification process to be more economical. Approaches like enzyme engineering, reconstitution of enzyme mixtures and bio-prospecting for superior enzymes are gaining importance.



In this chapter classical function-based metagenomic bio-prospecting was used to isolate beta-xylosidase encoding genes from composting horse manure. It has been shown in the previous chapter that this material is rich in lignocellulase degrading enzymes, including beta-xylosidase. Using agar plate activity screening, twenty six hits which showed beta-xylosidase activity were identified from the 20000 clone library. Analysis of these clones in three additional substrates show that some of these hits are capable of degrading multiple substrates. Characterisation of four clones at the sequence level revealed that some of these clones harbour multiple lignocellulase degrading genes while others harbour genes which encode for proteins with multiple substrate activity. Table 11 summarises the characteristics of the ORFs obtained from sequenced clones.

Two GH3 proteins identified from clone P17C1 are most probably responsible for two activity observed from activity assay (pNPX and pNPA).

Table 11: Summary of predicted GH ORFs identified in this chapter

ORF	Family	Characteristic of encoded protein
P17C1-18	GH3	Beta-xylosidase
P17C1-19	GH3	Most possibly bi-functional beta-xylosidase/ alpha-L-arabinofuranosidase encoding gene Phylogenetically different from three highly similar proteins and two well characterised GH3 proteins
P15A9-7	GH3	Possible bifunctional beta-xylosidase/cellobiose encoding gene
P118H6-1	GH1	Possible multiple substrate encoding gene (pNPX, pNPG,pNPC and pNPA)
P117H11-13	GH52	Possible thermophilic protein

However, it is not certain which of the two clones were responsible for pNPC activity since no cellobiose encoding gene was identified from this clone. Interestingly, P17C1-19 that is a possible bi-functional beta-xylosidase/alpha-L-arabinofuranosidase encoding gene is phylogenetically distinct from three of the most similar proteins at the sequence level. Whether the product of this ORF has additional activity such as cellobiose activity requires further investigation. ORF P15A9-7 showed highest similarity to beta-glucosidase. This ORF is most likely responsible for the cellulose activity observed in activity assay (Figure 24). In addition to this ORF, clone P15A9 harbours similar genetic structure as clone P17C1, suggesting that the cloned fragments of these clones are from the same organism. Indeed, sequence alignments of these two fragments show that they are 100% identical at nucleotide level. This is in line with three activities observed from this clone. The genetic structure of the cloned fragments from P15A9 and P17C1 suggest that the source organism is adapted for carbohydrate degradation.

ORF P118H6-1 from clone P118H6 harbours a possible multifunction GH1 glycoside hydrolase. Sequence analysis of conserved motifs of the product of this gene shows similar substrate specific residues as that of most thermophilic GH1 proteins, except that methionine was replaced with arginine (Figure 31). The effect of this substitution needs further analysis. A protein with a GH52 domain was also identified from ORF P117H11-13 of clone P117H11. This protein has 63% sequence identity to a thermophilic beta-xylosidase from *Paenibacillus mucilaginosus*. GH52 beta-xylosidases are rare and the identification of this gene demonstrates the diversity of hemicellulose degrading microorganism in the sample.

None of the sequenced clones from the library corresponds to those isolated through the PCR-based method or ORFs identified in the shotgun data (Chapter 2). Moreover, while GH52 proteins have been identified, no GH43 protein was identified from sequenced clones. This could be because of the few number of clones which were sequenced and functionally screened. Moreover, the shotgun data was shown to represent only a fraction of environmental species, leaving the majority of the genes unsampled. The lack of overlap between the two datasets shows that these methods have to be used in concert to identify the full complement of a particular enzyme class from an environmental sample.

The genes isolated in this chapter show great potential for biotechnological exploitation. The gene products of these clones require further characterisation to evaluate their suitability in industrial production of biofuel and other metabolites.

CHAPTER FOUR: Development of a cell-free metagenomic screening platform using *E. coli* S30 CFPS system as transcription-translation machinery.

4. Introduction

The exploitation of the metagenome for novel biocatalysts through functional screening relies on the ability to express the respective genes in a surrogate host (Gabor *et al.* 2004). Despite the potential for mining genetic novelty, function-based metagenomic studies is often hampered by methodological challenges (Beloqui *et al.* 2008; Singh & Macdonald 2010). The major reasons for this have always been associated with heterologous expression in surrogate hosts such as *E. coli* (Schlegel *et al.* 2013) and low throughput of screening methods. Although a lot has been done to overcome these challenges, most advances thus far still depend on this classical way of screening metagenomic libraries for novel gene discovery. One advancement to overcome some of the challenges is the SIGEX system (Uchiyama & Watanabe 2007) which was, in addition to screening for gene products encoded by an operon, developed to overcome the low throughput capacity of screening procedures associated with classical functional metagenomics. The SIGEX system is an ultra-high throughput screening system based on FACS (Fluorescence Activated Cell Sorting) to screen *E. coli*-based metagenomic libraries. Other high throughput fluorescence based screening systems such as METREX and PIGEX have been discussed in Chapter 1. Although these screening systems were successful, the construction of a metagenomic library results in cloning biases and as a consequence the loss of genes, particularly those which are present in low abundance. In addition, low expression level and toxicity of other environmental proteins in *E. coli* and the need for translocation of other recombinant proteins to the surface of the cell present a challenge.

Cell-free protein synthesis has become a convenient tool mainly because it overcomes some of the barriers of *in vivo* protein expression (Liu *et al.* 2005). This chapter describes the development of a second generation metagenomic screening platform. This ultra-high throughput system is based on cell-free protein synthesis and FACS. Beta-xylosidase was used as a model enzyme to demonstrate the effectiveness of this system in bio-prospecting genes from uncloned metagenomic DNA. The use of uncloned environmental DNA should ensure optimal gene screening since there would be no loss of genes during cloning and library constructions. The general overview of the method described in this chapter is represented in Figure 34.

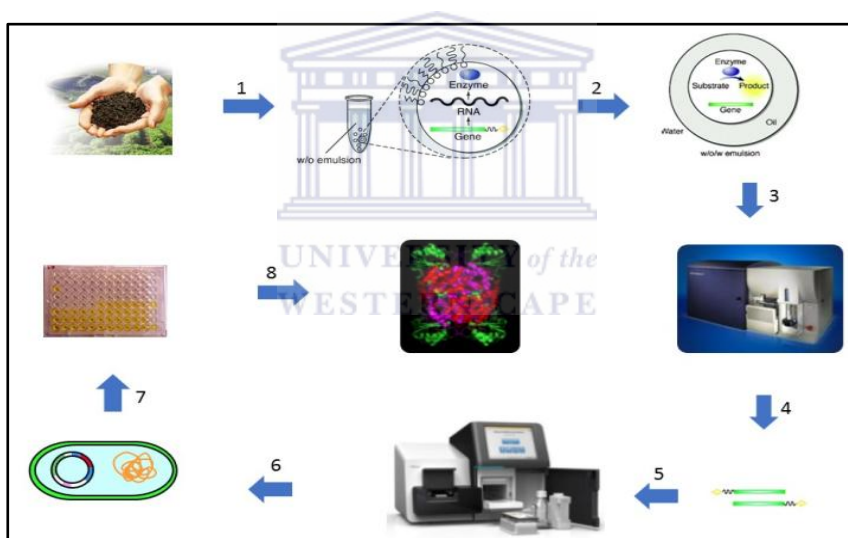


FIGURE 34: General overview of a novel IVC-FACS for screening uncloned environmental DNA. 1: A single molecule of DNA from the environmental sample is extracted and compartmentalised in a W/O droplet together with *in vitro* CFPS mixture and fluorescence substrate, 2: The W/O emulsion is incubated at the relevant temperature to allow transcription and translation of the environmental genes which are recognized on the DNA molecule and then re-emulsified to form a W/O/W emulsion, 3: W/O/W emulsions are screened on a FACS and positive emulsions are sorted and collected, 4: Positive emulsions are broken to release the DNA fragment, 5: The DNA fragment is next generation sequenced and the gene encoding enzyme is identified bioinformatically, 6: The identified gene is cloned and expressed *in vivo*, 7: Activity assay is carried out to confirm activity, 8: The resulting protein can be characterised for novelty.

In the previous chapters, thermophilic composting horse manure was shown to be rich in genes encoding for hemicellulose and cellulose degrading enzymes. In particular, it was shown in shotgun sequence data that beta-xylosidase encoding reads were present in the abundance of 1.01 reads per megabase, making these genes suitable as screening model genes. Beta-xylosidase encoding genes were isolated using classical function-based screening in Chapter 2, confirming that this material contains functional genes which encode beta-xylosidases.

4.1. Materials and methods

4.1.1. Bacterial strains and plasmids

Bacterial strains and plasmids used in this chapter are shown in Table 12.

Table 12: Bacterial strains and plasmids used in this chapter

Strain/Plasmid	Description	Source
<i>E. coli</i> BL21 (DE3)	F ⁻ ompThsdSB (rB ⁻ mB ⁻) gal (λ cl857 ind1 Sam7	Novagen
<i>E. coli</i> JM109	recA1 relA1 thi-1 (lac-proAB) gyrA96 hsdR17 endA1	Stratagene
pET21a		Novagen

4.1.2. Development of double emulsion (dE)

dE were prepared according to a modified method of Mastrobattista *et al.* (2005). Generally, a 50 μ l aqueous phase reaction mixture for IVC-FACS double emulsions was prepared consisting of commercial *E. coli* S30 (Promega) cell-free protein synthesis (CFPS), 10ng of uncloned metagenomic DNA of approximately 24kb (2.3.1.), 50 μ M of 4-Methylumbelliferyl- β -D-xylopyranoside (4-MUX). Double emulsions were prepared according to a modified method of Mastrobattista *et al.* (2005). Briefly, a solution of 1% (w/v) Span 60 and 1% (w/v)

cholesterol in decane was prepared at 45°C, and divided into 200µl aliquots and placed in a heat block at 37°C. A hand-extruding device (mini extruder; Avanti Polar Lipids, Alabaster, AL) was fitted with a 19mm Track-Etch polycarbonate filter with average pore size of 12µm (Sigma) inside the mini extruder. Two gas-tight 1ml Hamilton syringes (Gastight 1001; Hamilton, Reno, NV) were used for extrusion. The extruder was pre-rinsed with 3 x 1ml of decane. For emulsification, 50µl of aqueous phase was added to 200µl of the preheated decane/surfactant mix and loaded into one of the syringes. Aqueous phase was forced through the filter into the alternate syringe and directly forced back into the original syringe to complete one round of extrusion. In total, 7.5 rounds of extrusion were completed. The w/o emulsion formed was transferred to a 2ml microtube and incubated at 30°C for 2 hours to allow expression and enzymatic hydrolysis of 4-MUX within micro-droplets.

For the second emulsification step, a 19mm Track-Etch polycarbonate filter with an average pore size of 8 mm was fitted inside the extruder. The extruder was pre-rinsed with 3 x 1ml PBS. Two hundred and fifty microliters of the primary w/o emulsion was added to 750 ml of PBS containing 0.5% (w/v) Tween 20 and loaded into one of the syringes. The CFPS mix was forced through the filter into the alternate syringe and directly forced back into the original syringe to complete one round of extrusion. In total, 3.5 cycles of extrusion were performed. The double emulsion (dE) droplets were collected on ice.

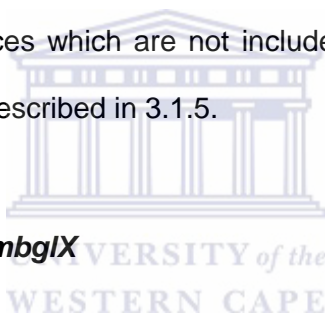
4.1.3. Activity screening using a fluorescence activated cell sorter (FACS)

dE droplets were diluted 100-fold in sterile-filtered PBS and run on a DB FACSaria II using PBS as sheath fluid. The cytometer was fitted with a 100mm nozzle, an ion laser emitting at 488 nm, and an argon ion laser tuned to 350nm. The applied rate of sorting was 20,000 events per second and single positive events were collected in a sterile 96 well plate. Positive events, when using 4-methylumbelliferyl-β-D-xylopyranoside (excitation. 360nm; emission. 449nm) as substrate, were detected using the DAPI fluorescence channel

(excitation. 345nm and emission. 455nm). Collected events were re-suspended in 30µl PBS and store at 4°C. To break the emulsion, collected events in PBS were placed in an oven at 75°C for 3 hours to evaporate PBS. DNA was re-suspended in 2.5µl of sterile ddH₂O and used in a 50µl multiple displacement amplification (MDA) using the Genomiphi kit according to manufacturer's specification (GE Health care life sciences).

4.1.4. Sequencing and Identification of beta-xylosidases

Sequencing of MDA-amplified DNA and sequence assembly was done as described in 2.1.4. Identification of beta-xylosidase ORFs was done as described in 2.1.6. In addition, identified ORFs were also analysed using the blastx tool on the NCBI database to determine if they have high similarities to sequences which are not included in the local library (see 2.1.6). Analysis of *mbgIX* was done as described in 3.1.5.



4.1.5. Amplification of *mbgIX*

PCR amplifications were done in a 50µl reaction mixture consisting of 1x Phusionbuffer, 200µM dNTPs, 0.5µM reverse primer, 0.5µM forward primer, 50ng template DNA, and 1U high fidelity Phusionpolymerase. Primer pair GGATCCATGCTGCCCTTCGAGATT and CTCGAGCTCTGCGTTTCGAGGTAAG were designed based on the sequence flanking the *bgIX* gene (approximately 100bp before the ATG) and used to amplify *mbgIX* using the following conditions: 98°C for 30sec, 35 cycles (98°C for 10sec, 56 for 30sec, 72°C for 1min), 72°C for 2 min. Reactions were performed using an Applied Biosystems thermocycler Gene AmpR2700.

4.1.6. Cloning of PCR products

The PCR products were purified using a Nucleospin PCR clean-up kit (MACHEREY-NAGEL, Germany) according to manufacturer's recommendations. Cleaned PCR products were cloned in pJET 1.2/ blunt (Fermentas) according to the manufacturer's recommendations. The reaction mixture contained 10µl of 2 X reaction buffer, 2µl PCR product (55ng/µl), 1µl pJET vector, and 5U of T4 ligase per µg of DNA and 6µl ddH₂O. The mixture was incubated at room temperature for 20 minutes and then used to transform electro-competent *E. coli* JM109 cell.

4.1.7. Preparation of electro-competent cells

E. coli JM109 or BL21 (DE) from glycerol stocks were plated on LB agar plates and incubated overnight at 37°C. A single colony from an overnight culture was transferred to 5ml LB broth in 50ml flask and incubated overnight at 37°C with shaking at 150rpm. This culture was used to inoculate pre-warmed 50ml LB broth in a 250ml flask. The flask was incubated at 37°C with shaking at 150rpm until the optical absorbance (OD) at 600nm reached about 0.6. The cells were transferred to a 500ml centrifuge tube, incubated on ice for 15 minutes and harvested by centrifugation at 3500g (Beckman J-26 XP, USA) for 10 minutes. The pellet was gently re-suspended in 10ml of ice cold 0.1M CaCl₂, incubated on ice for 30 minutes and harvested by centrifugation as above. The pellet was again gently re-suspended in 10ml of ice cold solution of 0.1M CaCl₂ containing 15% glycerol. Aliquots of 100µl were transferred into 1.5 ml tubes and stored at -80°C.

4.1.8. Electroporation

Electrocompetent *E. coli* JM109 cells were removed from -80°C storage and thawed on ice. Fifty microliters cell suspension was mixed with 3µl plasmid DNA (~10ng) in a pre-chilled Eppendorf tube and incubated on ice for 15 min. The mixture was transferred to a pre-chilled 0.1cm electroporation cuvette (BioRad). Electroporation was performed using a

BioRad Gene PulserR at 1.8KV, 25 μ F and 200 Ω . Following electroporation, 500ml LBbroth (10g/l tryptone, 5g/l yeast extracts and 10g/l NaCl) was added tothe cuvette and the cells were gently re-suspended. The cell suspension was transferred to 2ml eppendorf tube and incubated at 37°C horizontal shaking for 1 hour. Transformed cells were platedonto LB plates (LB broth containing 15g/l agar and100 μ g/ml ampicillin) followed by incubation at 37°C overnight.

4.1.9. Plasmid isolation

A glycerol stock of transformed *E. coli* culture was streaked on LB agar plate containing 100 μ g/ml of ampicillin and grown overnight at 37°C. A single colony of overnight cells was inoculated in 2ml LB broth containing 100 μ g/ml of ampicillin in a 50ml tube. The culture was incubated overnight at 37°C with shaking at 150rpm. Plasmid isolation was done using the Qiagen miniprep kit according to manufacturer's recommendations. The concentration was quantified using the Nanodrop ND-1000 spectrophotometer at 260nm (Nanodrop, Delaware USA).

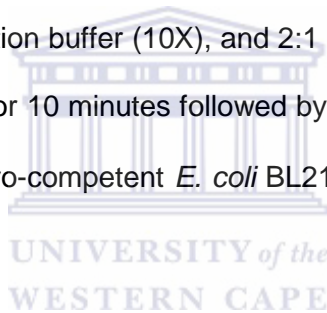
4.1.10. Plasmid digestion and gel electrophoresis

Isolated pJET2.1 clones weredigested with *Xho*I and *Bam*HI (Fermentas) in a 20 μ l reaction at 37°C for 16 hours. The digestion reaction contained 1U of restriction enzyme per 100ng of plasmid DNA, 2 μ l of reaction buffer (Buffer R), and 500ng of plasmid DNA. The insert fragment was separated from pJET2.1 backbone by agarose gel electrophoresison 0.7% [w/v] of agarose gel preparedin 0.5 X TAE buffer (0.2% [v/v] Tris base; 0.5% [v/v] glacial acetic acid and 1% [v/v] 5MEDTA, pH 8.0) and 10 μ l/L of 0.5 μ g/ml ethidium bromide. DNA was mixed with a loading dye (60% [v/v] glycerol and 0.25% [w/v] Orange G) and loaded into the wells of cast gel. DNA molecular marker (1kb Fermentas, λ *Hind*III or λ *Pst*I) was used as a standard. The samples were electrophoresed at 100V in 0.5 X TAE buffer.The insert

fragment was excised from the gel and purified using the Nucleospin gel purification kit. Purified insert fragment was confirmed by sequencing using an automated DNA sequencer 373 and fluorescein labelled primers (Applied Biosystems, USA) at the University of Stellenbosch sequencing facility.

4.1.11. Sub-cloning

pET21a vector (Novagen) was linearized by digesting with *Xho*I and *Bam*HI (Fermentas). Digestions were done in a 20µl reaction as indicated above. Linearised plasmid was purified using Nucleospin PCR product purification kit per manufacturer's recommendations. Purified insert DNA (4.1.11) was ligated with linearised pET21a in a 20µl reaction containing 1U T4 DNA ligase (Fermentas), 2µl ligation buffer (10X), and 2:1 molar ratio of insert to vector. The reaction was incubated at 22°C for 10 minutes followed by enzyme inactivation at 70°C for 5 minutes. Transformation of electro-competent *E. coli* BL21 (DE3) was done as described in 3.2.10.



4.1.12. *In vivo* expressions

E. coli clones from glycerol stock were streaked on LB agar containing 100µg/ml ampicillin and grown overnight at 37°C. A single colony of overnight cells was inoculated in 5ml LB Amp broth in a 50 tube and grown overnight. These cells were used to inoculate 100ml LB Amp broth and grown until $OD_{600}=0.5-0.6$. IPTG was added to a final concentration of 1mM followed by further incubation at 37°C for 3 hours with shaking. Cells were harvested by centrifugation at 2000g for 5 minutes. The cell pellet was re-suspended in 1ml of PBS buffer, subjected to 3 cycles of sonication at 50% power for 10 second using a Sonoplus HD-070 sonicator (Bandelin, Germany). Lysed cells were centrifuged at 30000g for 30 minutes; supernatant was carefully transferred to a 2ml sterile tube and stored at -20°C until use.

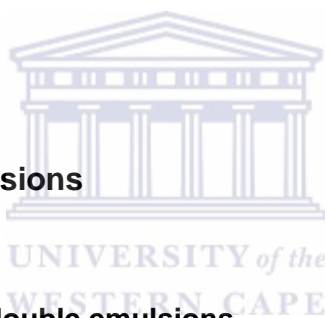
4.1.13. Enzyme essays

Enzyme assays were done in a 250µl reaction in PBS buffer using 4-Methylumbelliferyl- beta-D xylopyranoside (4-MUX, Sigma, USA), 4-Nitrophenyl beta-D-xylopyranoside (pNPX, Sigma, USA) and 4-Nitrophenyl beta-D-glucopyranoside (pNPG, Sigma, USA). The reaction contained 500µl of cell-free extracts and 50µM 4-MUX, pNPX or pNPG in PBS. The reaction was incubated at 30°C for 16 hours. The reaction was stopped by adding 600µl of 1MNa₂CO₃ (for chromogenic substrates) or quantified directly. The release of nitrophenol (pNP) was measured as absorbance at 405nm in a plated reader. Hydrolysis of 4-Methylumbelliferyl (4-MUX) was measured at the excitation wavelength of 355nm and emission wavelength of 488nm.

4.2. Results and discussions

4.2.1. Development of double emulsions

In order to use FACS to screen uncloned environmental DNA, DNA molecules, together with cell-free protein synthesis (CFPS) components need to be compartmentalised in emulsion droplets. Mastrobattista *et al.* (2005) have demonstrated that double emulsions (dE) can be used as a cell like compartment to screen for enzyme activity. Therefore, dEs containing DNA molecules and commercial *E. coli* CFPS system were developed and analysed immediately through FACS. While an in-house CFPS preparation could be easily developed and used for this purpose, the commercially available S30 preparation was used to obtain optimal activity since the product is fortified with various components to ensure optimum transcription-translation (Carlson *et al.* 2011).



One milliliter of developed dEs was analysed at a time after dilution with PBS while the rest of the sample was kept on ice. A general decrease in the number of dE (double emulsion) droplets per millilitre of solution over time was observed during FACS analysis. Samples which were analysed immediately after the development of dEs contain more dE droplets than samples which were stored on ice. The decrease was directly related to time of storage as shown in Figure 35. This observation was also reported by Wu *et al.* (2011) who showed that the stability of the emulsion droplets containing the in-house S30 extracts decrease over time during storage. Wu *et al.* also showed that replacing the in-house S30 system with the commercial kit did not stabilize the droplets significantly but replacing it with deionized water did prevent them from increasing in size over a 2h observation period. They suggested that the instability was likely due to the relative high concentration of the IVTT (*in vitro* transcription-translation) reagents in the internal aqueous phase, which presumably created an osmotic pressure difference between the internal droplets and the surrounding environment.

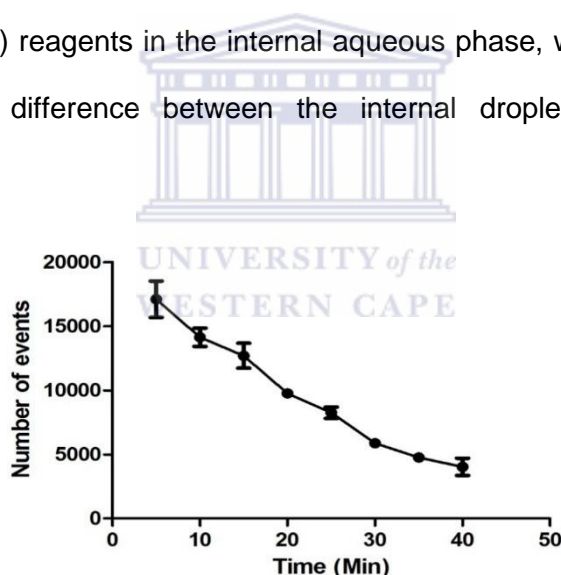
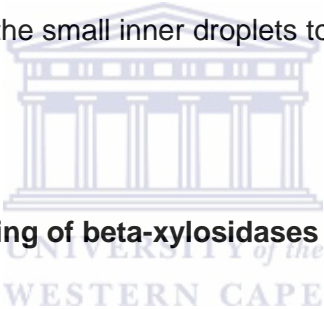


FIGURE 35: Number of events observed vs storage time. The number of droplets which appear during FACS sorting decreased over time. Error bars represent standard deviation of three independent experiments.

Indeed, Florence & Whitehill (1981) also indicated that the diffusion of the external aqueous phase through the oil layer into the highly concentrated internal aqueous phase can result in the expansion of the droplet and eventually lead to rupture. Recent studies have demonstrated that double emulsions containing a NaCl isotonic solution as the external water phase were stable for at least two months (Zhang *et al.* 2014), confirming that a

difference in osmotic pressure contributes to droplet rupture. However, this was avoided in this study in order to prevent enzymatic inhibition due to high salt concentrations.

Chávez-Páez *et al.* (2012) described another process which results in emulsion expansion called coalescence. This is the process by which two or more emulsion droplets merge during contact to form a single emulsion. When this happens the resulting daughter droplet is bigger in size, and the solutes from both parent droplets are mixed in a daughter droplet. An earlier study by Ficheux *et al.* (1998) identified two types of instabilities that are responsible for the disruption of double emulsions: (i) coalescence of the small inner droplets with the oil interface and (ii) coalescence between the small inner droplets within the dE (double emulsion) droplets which has multiple inner droplets. The first type of instability leads to a complete delivering of the small inner droplets toward the external phase whereas the second one does not.



4.2.2. IVC-FACS screening of beta-xylosidases activity

A commercial *E. coli* S30 CFPS system was used as transcription-translation machinery in the IVC-FACS screening of beta-xylosidase activities using uncloned metagenomic DNA as template. Metagenomic DNA was prepared as described in section 2.1.2. The aim was to incorporate one uncloned metagenomic DNA fragment of approximately 24kb per dE. Ten nanograms of 24Kb DNA molecules were used in a reaction which generates approximately 10^{10} dE droplets. This in principle translates to 3.86×10^8 DNA molecules per 10^{10} dE droplets, resulting in a ratio of 26:1 dE droplets to DNA molecule. According to Poisson statistics, the distribution of DNA molecules within the dE droplets should result in 26 droplets without DNA molecule for every dE which contains a DNA molecule. This should ensure that each dE droplet would contain no more than one DNA molecule.

Gating was based on DAPI fluorescence and the size of dE droplets (FSC-H). Figure 36 shows a fluorograph of one million events after *in vitro* expression of metagenomic DNA in the presence of MUX. Three populations of positive dE events were observed, as shown in N1, M1 and R1 in figure 36. Gate M1 in the experiment represents the most fluorescing population which has 8 events per million. These events are not present in either of the controls. Gate R1, which represents the second most fluorescing population, has 150 events per million in the experiment compared to 15 and 12 in the no DNA and no substrate controls respectively. Gate N1, which represents the least fluorescing events close to background fluorescence, has 57 events in the experiment compared to 42 and 12 in the no DNA and no substrate controls respectively.

Highly fluorescing dE droplets in gate M1 were extremely rare, and in principle should represent genes which are efficiently expressed using the *E. coli* extract system under conditions used are rare in this metagenome. On the contrary, genes which are expressed in R1 make up the majority of the positive events. Low fluorescence intensity of this population may indicate that enzymes in this gate have low affinity to MUX (Herrmann *et al.* 1997). Alternatively, this population might represent enzymes that have optimum activity at different conditions other than those used in this study. Other factors, such as codon bias, weak promoter recognition, incorrect protein folding etc may have also contributed to this low fluorescence intensity (covered in section 1.2.4). This is also most likely to apply to the enzymes in population N1. dE droplets from gate M1 were sorted while R1 and N1 events were avoided to ensure that only true positive events were selected for further analysis.

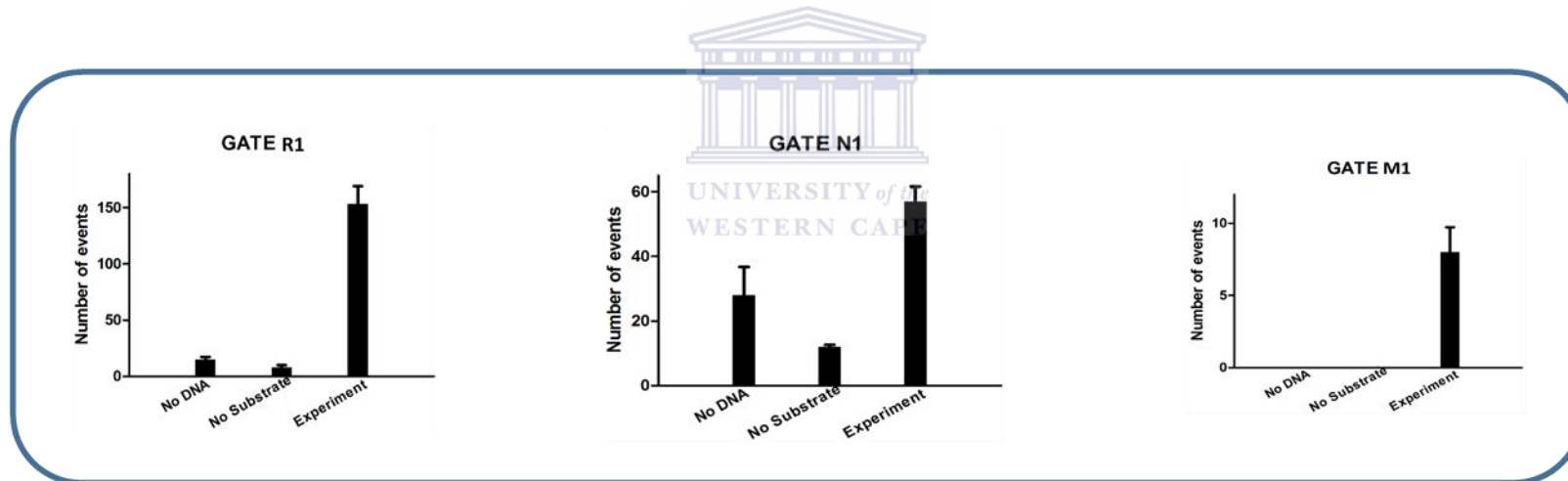
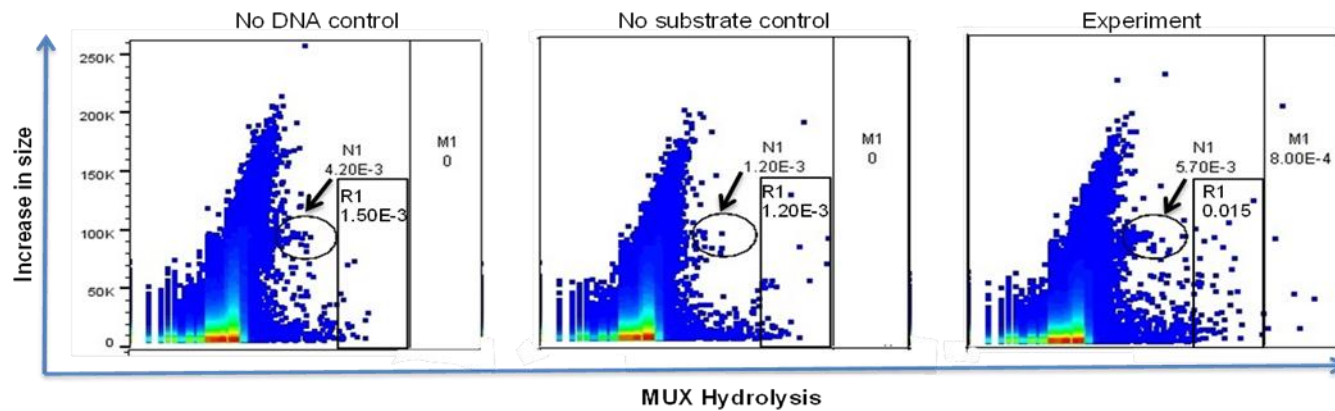


FIGURE 36: FACS recording produced during beta-xylosidase activity screening. 4-MUX was used as fluorogenic substrate and *E. coli* CFPS system was used as transcription translation system. The image shows one million events. Three gates, M1, N1 and R1 represent the average of three rounds of one million events. Error bars are the standard deviation of three rounds of one million events screening.

4.2.3. DNA recovery from dE droplets and amplification

DNA molecules from sorted dE droplets ingate M1 (Figure 36) were recovered and amplified in a MDA (multiple displacement amplification) reaction. DNA from nine individual dE droplets was recovered and amplified. Five dE droplets were combined and DNA was also recovered and amplified from these combined droplets. This was done to maximize the number of genes recovered per sequencing run. The resulting DNA was electrophoresed on an agarose gel (Figure 37).

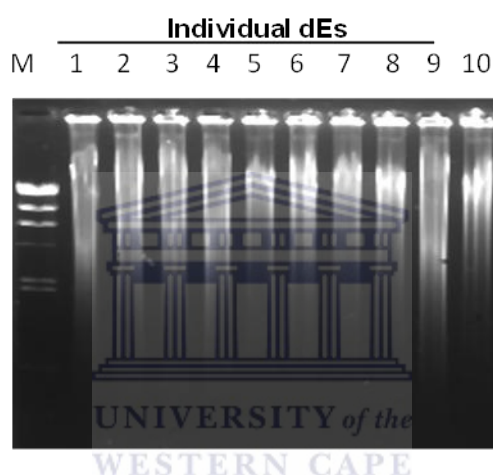


FIGURE 37: Image of agarose gel electrophoresis of DNA recovered and MDA-amplified from dEdroplets. Lane M is a marker, lane 1-9 is DNA from single dE droplets, lane 10 is the DNA recovered and amplified from combined five double emulsions droplets.

4.2.4. Sequencing and reads assembly

Amplified DNA from dE droplets were sequenced as described in 2.1.4. An average of 11 Mb of DNA was obtained per dE droplet after sequencing (Table 13). This was higher than an expected single DNA molecule of about 24 kb in size (the amount of DNA molecules which was expected to be incorporated per dE droplet was indicated in section above), likely indicating that there was significantly more DNA than just one DNA molecule incorporated per dE droplet. Another possibility could have been that droplets ruptured and released DNA molecules into the sheath fluid, resulting in the contamination of the sheath fluid and

consequently increased amount of DNA per dE droplet. Previous studies have also shown that recovery of a single DNA molecule per emulsion droplet is a challenging process even when the minimum copy numbers of DNA molecules were used (Sepp & Choo 2005).



Table 13: Amount sequence data generated per dE droplet

	C5dE	dE1	dE2	dE3	dE4	dE5	dE6	dE7	dE8	dE9
Total number of reads	2934692	120139	38869	543780	434940	218826	445292	1162350	500542	337220
Average read length (bp)	228	244	239	223	294	219	232	226	246	268
Number of de-replicate reads	2406447	84097	26820	500278	404494	188190	374045	1115856	460499	296754
Number of reads after de-replicate removal	528245	36042	12049	43502	30446	30636	71247	46494	40043	40466
Number of assembled reads	21646	3235	2760	3209	2315	3138	3137	3675	2835	2563
Number of contigs	6295	886	789	816	762	826	799	958	873	771
Average contig length (bp)	784	891	836	877	893	832	911	867	799	891
Number of un-assembled reads	506599	32806	9290	40293	28131	27498	68109	42819	37208	37903
Total DNA(bp)	120439860	8794175	2879804	9701035	8951065	6709205	16529239	10507644	9850667	10844995



For instance, Griffiths & Tawfik (2003) compartmentalized the phosphotriesterase (PTE) encoding gene (*opd*) attached to microbeads in a ratio of 0.3:1 beads with no genes attached to them (negative). When they analysed these emulsions using flow cytometry, they find found that positive beads (attached to *opd*) were 74.3% rather than the expected 33%, about 41% higher. They concluded that this is almost certainly because some compartments contain more than one bead. Furthermore, they mixed high fluorescence beads (with *opd* gene attached) with negative population (beads attached to Δ *opd*) in a ratio of 1:10, 1:100 or 1:1000 and sorted positive beads based on fluorescence using flow cytometry and cell sorting. Analysis of the sorted beads indicated up to 200-fold enrichment of the *opd* gene following selection. However, although this was a significant enrichment, for every positive event (*opd*) they sorted, they also get four negative events (Δ *opd*) per emulsion droplet. This shows that even after diluting the gene of interest significantly, compartmentalisation of a single DNA molecule was still a challenge.

Bernath *et al.* (2004) prepared two separate primary emulsions with two genes of different length. The positive emulsion contained the *foIA* gene and fluorescent substrate, while the negative emulsion contained the *M.HaeIII* gene without any fluorescent substrate. The two primary emulsions were mixed in a ratio 1:100 *foIA* to *M.HaeIII* and then re-emulsified to give dE. After FACS screening for positive genes on the basis of fluorescence, the positive genes isolated from the sorted droplets appeared at a ratio of 1:3 *foIA*:*M.HaeIII*, indicating a 30 fold enrichment from a starting ratio of 1:100. Since only the positive droplets were sorted, this also indicates that there was more than one primary emulsion per dE which resulted in the incorporation of *M.HaeIII* genes in positive droplets. Indeed, the authors admitted that there was an average of five primary emulsions per dE droplet which resulted in the presence of negative genes in dE.

During the early development of ePCR, Kojima *et al.* (2005) mixed equal amount of six different DNA mutants. They diluted the DNA mixture before compartmentalisation in w/o

emulsions in such a way that each droplet contains one molecule on average. Single DNA molecules were immobilised on beads and compartmentalised in the w/o emulsions together with PCR reagents and specific primers. After solid-phase single molecule ePCR, the beads-DNA complexes were recovered and extensively diluted to less than one bead per tube on average, and each DNA fragment on individual beads was separately re-amplified. After sequencing the amplicons, they found that of 29 samples, 22 samples were sequence-amplified from a single mutant DNA, and the other 7 samples were from multiple mutant DNA templates. This meant that 24% of emulsion droplets had multiple DNA templates. Although the authors were uncertain of the origin of multiple DNA templates, the most probable source could be compartmentalisation of more than one template per emulsion.

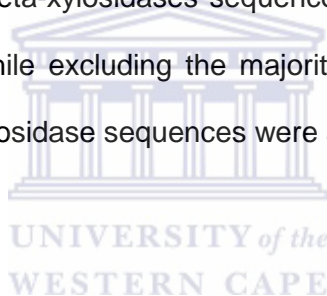
It is also worth mentioning that estimating the size of high molecular weight metagenomic DNA is not as accurate as quantifying small DNA fragments. This is mainly because it is easier to estimate sizes of small linear DNA fragments (>20 kb) using agarose gel electrophoresis than a mix of high molecular weight environmental DNA (Fangman 1978). Environmental samples contain DNA in circular and linear, relaxed and supercoiled forms, which although they are different sizes, might appear in a band of seemingly one size on gel electrophoresis (Garner & Chrambach 1992; Åkerman & Cole 2002). This makes it difficult to make precise estimations of the distribution of DNA molecules per dE droplet.

This suggests that although the distribution of DNA in emulsions is expected to follow Poisson distribution, there is enough experimental evidence that shows that generating a single DNA molecule per emulsion droplet is a challenge even when the number of DNA molecules is significantly lower than the number of droplets. This aspect of the method still requires vigorous optimisation and improvement, particularly when the goal is to compartmentalise single un-cloned DNA fragments to facilitate the identification of a gene sequence that may be completely unknown encoding the activity for which an event is sorted and selected. In particular, background DNA presents a serious challenge in the

development of this technique since identification of genotype is less complicated without background DNA.

4.2.5. Identification of sorted genotype

In the presence of a high amount of background DNA, the identification of the DNA sequence responsible for the activity for which the positive droplet was selected becomes challenging. However, since each droplet was sorted based on the phenotypical characteristic expressed (beta-xylosidase activity) and thus the gene that encodes for that characteristic, it should be possible to screen for the sequence of interest *in silico* within the high background of data generated for each dE droplet. All assembled contigs were referenced to a local library of beta-xylosidases sequences (section 2.1.6) to identify beta-xylosidase containing contigs while excluding the majority of background DNA sequence. Contigs which contained beta-xylosidase sequences were analysed using the NCBI database to identify their closest hit.



Seven of nine individual dE droplets contained a contig with sequence identity to a beta-xylosidase encoding gene in the local database while dE1 and dE8 does not show any sequence with significant homology to beta-xylosidase. Interestingly, a NCBI blastx analysis of seven contigs with sequence similarity to beta-xylosidase genes showed that they all have 54% identity to a glucoside hydrolase family 3 protein from *Burkholderia* sp. A1 (Accession number: WP_025101624.1). In the well that contains five dE droplets, four contigs also showed identity to the same protein, totalling eleven dE droplets out of fourteen which contain the same glucoside hydrolase gene (*mbgIX*). It has been shown in Chapter 2 and 3 that the sample used in this study contains diverse beta-xylosidase encoding genes. The identification of *mbgIX* in the majority of dE droplets suggests that mIVC-FACS enriched for this specific gene and that the gating selected adds another level of selection.

Interestingly, this gene (*mbglX*) could not be identified in the shotgun data. This could be because *mbglX* was not abundant enough to be sampled in the 830Mb of sequence which was generated for the shotgun metagenome. According to the rarefaction curve (section 2.2.2), more sample needed to be sequenced to generate data which reflect the entire sample diversity. In principle, an equivalent of about 930Mb of DNA is screened in less than 1 minute in mIVC-FACS screening (assuming the DNA:dE ratio discussed). This represents a very high throughput method of screening which analyses exponentially higher amount of DNA to increase the chances of identifying rare genes. This underlines one of the advantages of mIVC-FACS over other screening methods, including *in vivo* expression of clone libraries in heterologous hosts.

The *mbglX* enrichment could also be due to the bias of the *E. coli* S30 CFPS system. However, the exclusive selection of this gene is most likely due to the stringency of the gate used, which was purposefully designed to select only a subset of the population of positive dEs. A range of gatings would need to be analysed in future, which may reveal a larger diversity of actively expressing beta-xylosidases.

4.2.6. MBglX confers beta-xylosidase activity

While the identification of *mbglX* in the sequence for the majority of the screened dE droplets strongly suggests that it was selected for during FACS screening, it was also important to confirm if it encodes for a protein with beta-xylosidase activity, particularly because it was annotated as a glucoside hydrolase/beta-glucosidase (and not specifically a beta-xylosidase) by the NCBI BLAST analysis. Beta-glucosidases catalyse the cleavage of the glycosidic bonds existing in disaccharides, oligosaccharides and alkyl or aryl beta-glucosides (Wallecha & Mishra 2003). Beta-xylosidases on the other hand are exotype glycosidases that hydrolyse short xylooligomers into single xylose units (Shallom & Yuval 2003).

While beta-glucosidases conferring beta-xylosidase activity have been reported, they represent a handful of dual activity hydrolases. Kimura *et al.* (1999) reported a beta-glucosidase (EC 3.2.1.21) which was purified as an electrophoretically homogeneous protein from a solid culture of *Aspergillus sojae*. The purified enzyme was shown to hydrolyse beta-D-xylopyranosides as well as beta-D-glucopyranosides; the K_m and V_{max} values on rho-nitrophenyl beta-D-glucopyranoside were 0.14mM and 16.7 micromol/min/mg protein, and on rho-nitrophenyl beta-D-xylopyranoside 0.51mM and 12.2 micromol/min/mg protein, respectively. Zhou *et al.* (2012) also demonstrated that a beta-glucosidase gene (*RuBGX1*) isolated from a yak (*Bos grunniens*) rumen metagenome was able to hydrolyse p-nitrophenyl-beta-d-glucopyranoside (pNP-Glc) and p-nitrophenyl-beta-d-xylopyranoside (pNP-Xyl).

To confirm if BglX also exhibits both xylosidase and glucosidase activity, it was amplified from DNA recovered from the combined five dE droplets, cloned into pET21a (Figure 38) and expressed in *E. coli* under control of the T7 promoter.

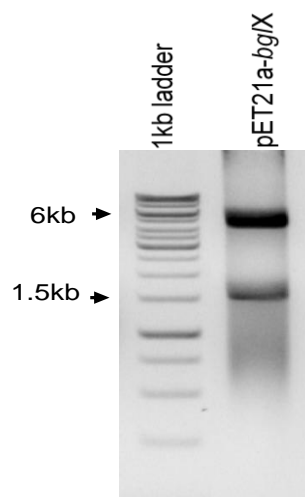


FIGURE 38: Agarose gel image of double digested pET21a harbouring *mbglX*. The plasmid was double digested with *Bam*H1 and *Xho*I and run in 1% agarose.

After protein expression, cells were lysed and crude extract was used for activity assays using pNPX and pNPG as substrates. Figure 39 shows that MBgIX is able to hydrolyse pNPX and pNPG. This protein shows higher activity on pNPG compared to pNPX, suggesting that it is a multi-substrate beta-glucosidase.

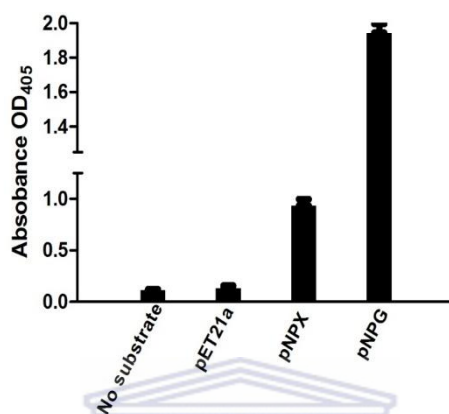


FIGURE 39: Activity obtained from crude extracts expressing MBgIX, assayed on pNPX or pNPG. The assay was done with crude extracts of induced BL21(DE3) cells harbouring pET21a-*mbgIX* or pET21a. One microgram of total cellular protein was used for the assay. No substrate indicates activity of cells harbouring pET21a-*mbgIX* in a buffer without pNPX or pNPG. pET21a activity indicates activity of cells harbouring pET21a only while pNPX and pNPG represent activity of cells harbouring pET21a-*bglX* and assayed on pNPX and pNPG respectively.

4.3.7. Catalytic amino acids residues and active sites of MBgIX

MBgIX has been shown to hydrolyse both pNPX and pNPG with preference for pNPG. In order to understand the catalytic mechanisms involved in this bi-functional activity, multiple sequence alignments were performed and the catalytic residues were analysed (Figure 40). This alignment revealed that MBgIX has some unique differences based on conserved domains and catalytic residues. The most obvious difference is the substitution of the catalytic acid/base residue glutamic acid with aspartic acid. This was also observed in a bi-functional beta-glucosidase/beta-xylosidase (RuBGX1) as reported by (Zhou *et al.* 2012).

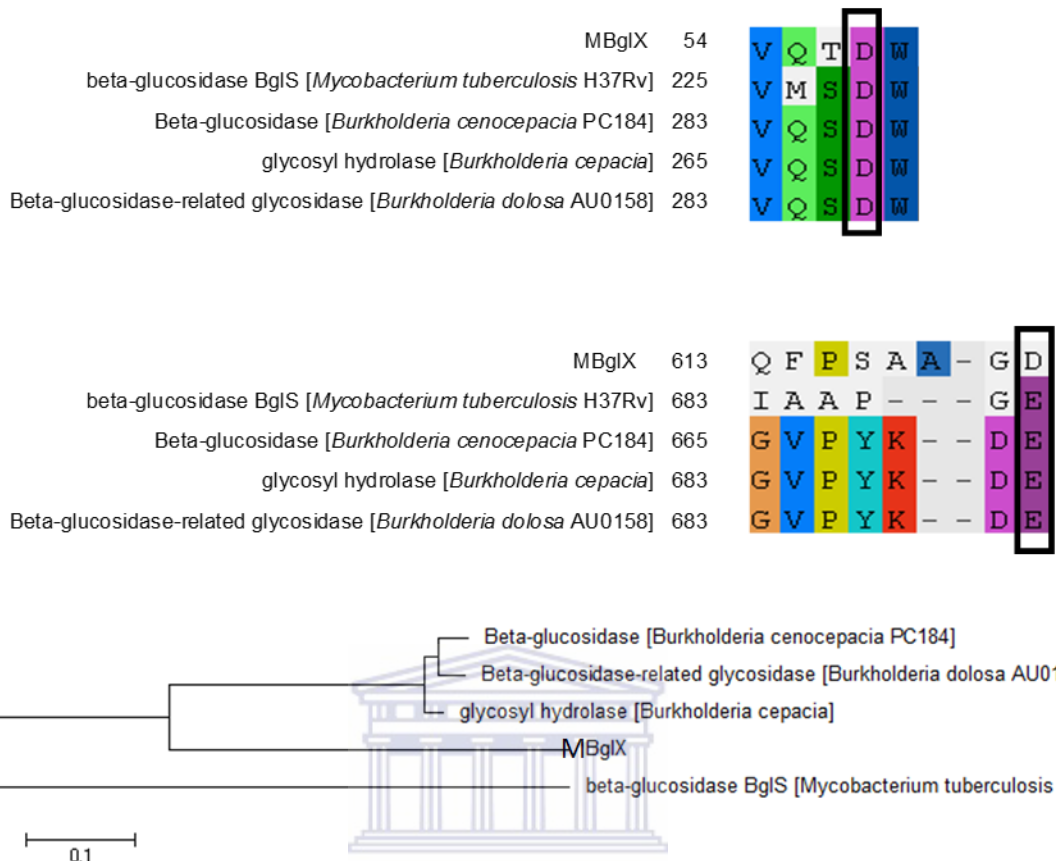


FIGURE 40: Multiple sequence alignments of MBglX with four GH3 beta-xylosidases. Residues in black blocks show known conserved residues. Two best matches in the NCBI proteins database together with the two best matches from well-studied reference species were selected for comparison. Analyses were done using the SmartBlast tool in the NCBI server.

Dodd *et al.* (2010) conducted site-directed mutagenesis studies of the conserved residues of GH3 enzymes. They reported that substitution of the catalytic acid/base glutamic acid with aspartic acid resulted in a 14-fold decrease in pNPX hydrolysis and a concurrent 1.1-fold increase in pNPG hydrolysis. They concluded that glutamic acid may therefore contribute to the discrimination between xylooligomers and beta-glucosides, which might be the case with MBglX.

4.3. Conclusion

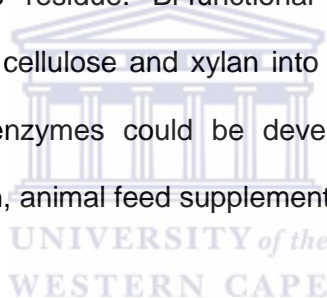
This chapter describes the development of a novel method for screening uncloned metagenomic DNA. The *in vitro* transcription translation system was compartmentalised in mini dE droplets to allow the synthesis of protein from uncloned metagenomic DNA. Although the droplets were observed to be losing stability over time, enzymatic activity within these mini reactors was successfully screened and sorted through FACS. This instability may have resulted in the high amount of DNA sequences obtained than what was expected per droplet. This background DNA was a major challenge in the development of this method. Understandably, this problem is under emphasized in most studies where cloned genes are screened in expression vectors. In such studies, background DNA does not have a major effect on gene enrichment since a desired fragment is separated from the background through gene specific amplification (Mastrobattista *et al.* 2005).

The accurate DNA quantification process is also a very critical aspect for successful application of this method. The complexity of metagenomic DNA makes it difficult to correctly quantify DNA molecules based on the size obtained from agarose gel electrophoresis. Uncloned environmental DNA exists in different forms which migrate differently based on whether it's circular or linear. Therefore, a band which appears to be of the same size in agarose gel can be significantly different, resulting in mis-quantification of the DNA loaded in a reaction.

FACS screening of positive droplets based on the hydrolysis of MUX revealed three distinct populations that fluoresce above the background population. Analysis of the DNA recovered from nine highly fluorescing dE droplets revealed that the *mbgIX* gene could be identified in eleven of fourteen dE droplets. Together with the finding that expression of *mbgIX* confers beta-xylosidase activity, these findings strongly suggest that *mbgIX* was responsible for the phenotypic characteristic observed during FACS screening. Moreover, these findings

demonstrate that mIVC-FACS is a very sensitive and highly selective technique. Indeed, flow cytometry gives a direct relationship between gene expression and fluorescence intensity. Genes which are highly favoured by the transcription-translation machinery and assay conditions also show high fluorescence intensity over genes which are poorly or not expressed under the same conditions (Rosano & Ceccarelli 2014). While this is a shortcoming when genes encoding diverse enzymes are of interest, selection of activity in different gates could most likely improve the variety of selected genes.

Further analysis of MBgIX shows that this protein is a bi-functional enzyme which acts on both pNPX and pNPG but with preference for pNPG. Sequence analysis of MBgIX showed that this bi-functional activity might be due to the substitution of glutamic acid with aspartic acid in the acid/base catalytic residue. Bi-functional enzymes are important in the simultaneous saccharification of cellulose and xylan into fermentable glucose and xylose (Christel *et al.* 2011). These enzymes could be developed for applications in biofuel production, detergents production, animal feed supplements, etc.



Isolation of *mbgIX* also demonstrates that bio-prospecting of novel genes through shotgun sequence data analysis can be limiting. While this method can provide information on relative abundance of general family of genes, genes such as *mbgIX* which confers multiple activities are most likely to be left unidentified. In addition, mis-annotation of genes in the databases also makes bio-prospecting of genes based on shotgun data challenging. mIVC-FACS is high throughput, and was able to isolate *mbgIX* which was not identified in shotgun sequence reads, suggesting that this method can be used for the discovery of genes which occur in less abundance.

This study is the first to describe a method for screening uncloned metagenomic DNA and it offers a lot of potential in metagenomic studies, and represents a significant paradigm shift. Construction of metagenomic libraries is a costly and laborious process which also reduces

the efficiency of the screening process (Daniel 2005). The use of a CFPS system for screening metagenomic DNA eliminates the majority of the processes associated with library construction. FACS is a robust technique which can routinely screen over 10^7 clones per hour (Bernath *et al.* 2009). The use of this technique for metagenomic screening offers a revolutionary opportunity to screen DNA molecules in a very high throughput manner.



CHAPTER FIVE: mIVC-FACS screening of beta-xylosidase activity using a *Rhodococcus erythropolis* CFPS system

5. Introduction

Chapter 4 reported on the use of the commercial *E. coli* CFPS system in the development of a novel method for screening for genes from uncloned metagenomic DNA using beta-xylosidase as the model protein. Challenges and opportunities of this technology were also highlighted. In particular, the amount of background DNA stands out as one of the concerning challenges. Possible sources of this background DNA were hypothesised to be the rupture of dE droplets resulting in DNA molecules in the sheath fluid as well as high input DNA. To assess if the high amount of starting DNA was the source of this background, the amount of starting DNA was reduced onethousand fold.

Moreover, the selection of *mbgIX* in the majority of the droplets sorted from gate M1 (4.2.2) suggests that this gene is efficiently expressed using an *E. coli* based cell-free expression under the conditions used. In order to understand if a different CFPS system will also select for the same gene at the same level of expression, a novel in house actinobacteria based CFPS system was used instead of the commercial *E. coli* system. The method of developing CFPS extracts are well established (Kim *et al.* 1996). These extracts contains all the catalytic components necessary for energy generation and protein synthesis from crude lysates of microbial, plant, or animal cells (Shin & Noireaux 2010). This includes the necessary elements for transcription, translation, protein folding, and energy metabolism (Tawfil & Graffiths 1998). While any organism can potentially provide a source of crude lysate, the only commercially available CFPS systems are derived from *E. coli*, rabbit reticulocytes (RRL), wheat germ (WGE), and insect cells (Kim & Kim, 2009) with the *E. coli* system being the only one derived from bacterial cells.

Actinobacteria was specifically chosen because this phylum represents one of the largest taxonomic units within the domain Bacteria (Clarridge & Zhang 2002) and exhibits diverse physiological and metabolic properties (Hosaka *et al.* 2009). In particular, *R. erythropolis* was chosen for its ability to recognise promoters different to those that *E. coli* can, and fold proteins differently due to different codon usage.

5.1. Material and methods

5.1.1. Bacterial strains and plasmids

Table 14: Bacterial strains and plasmids used in this chapter

Strain/Plasmid	Description	Source
<i>R. erythropolis</i> PR4		Nakashima & Tamura (2004)
<i>E. coli</i> BL21 (DE3)	F ⁻ ompThsdSB (rB ⁻ mB ⁻) gal (λcl857 ind1 Sam7	Novagen
<i>E. coli</i> JM109	recA1 relA1 thi-1 (lac-proAB) gyrA96 hsdR17 endA1	Stratagene
pET21a		Novagen
pET21a-xyIB	pET21a with xyIB between BamHI and XhoI	This study
pET21a-xyIT	pET21a with xyIT between BamHI and XhoI	This study
pET21a-bgICX	pET21a with bgICX between BamHI and XhoI	This study
pTip-RC1	P _{tipA} ChI ^r rep (pRE8424), MCS type 1	Nakashima & Tamura (2004)
pTip-RC4	pTip-RC1 without P _{tipA}	This study
pTip- xyIA	pTip-pRC4 with xyIA fragment between BamHI and XhoI	This study
pTip- bgAX	pTip-pRC4 with bgAX fragment between BamHI and XhoI	This study
pTip- bgCX	pTip-pRC4 with bgCX fragment between BamHI and XhoI	This study
pTip- xyIB	pTip-pRC4 with xyIB fragment between BamHI and XhoI	This study
pTip- xyIT	pTip-pRC4 with xyIT fragment between BamHI and XhoI	This study
PTip-bglX	pTip-pRC4 with bglX fragment between BamHI and XhoI	This study
Fos5902	pCC1FOS vector harbouring ~ 40Kb insert with beta-xylosidase encoding gene	University of the Western Cape, IMBM

5.1.2. Development of an actinobacteria derived cell-free protein synthesis (CFPS) system

A novel *R. erythropolis* CFPS system was prepared in the IMBM lab by Ms Whitney Takalani Maake using a protocol adopted from Kim *et al.* (2006). Generally, cells were grown in a laboratory bioreactor at 30°C and 500rpm agitation in 1.5L of TGP medium. The cells were harvested in the mid-log phase ($OD_{600} \sim 2.6$) by centrifugation at 4000g at 4°C for 30 minutes and washed three times by re-suspending in 20ml ice-cold buffer A (10mM Tris-acetate buffer at pH 8.2, 14mM magnesium acetate, 60mM potassium glutamate, and 1mM dithiothreitol containing 0.05% of 2-mercaptoethanol). The thawed cells were re-suspended in 1ml buffer B (buffer A without 2-mercaptoethanol). Cell suspensions were lysed using the Bandelin Sonoplus sonicator at 50% power setting for 6 X 30 sec rounds on ice. The lysate was centrifuged for 30 min at 20000g at 4°C and the supernatant was retained. The supernatant was incubated at 30°C for 1h followed by a second round of centrifugation at 20000g at 4°C. The supernatant was retained and the extract was flash-frozen in liquid nitrogen and stored at – 80°C until use.

5.1.3. Validation of the *R. erythropolis* CFPS system

The integrity of *R. erythropolis* cell-free extracts was tested according to a method of Kim *et al.* (1996). Typically, a 50µl cell-free protein synthesis (CFPS) reaction mixture consists of 57 mM Hepes-KOH (pH 8.2), 1.2mM ATP, 0.85mM each of ATP, CTP, GTP, and UTP, 2mM DTT, 0.17mg/ml total *E. coli* tRNA mixture, 0.64mM cAMP, 90mM potassium glutamate, 80mM ammonium acetate, 12mM magnesium acetate, 34µg/ml folinic acid, 1.5mM of each 20 amino acids, 2% PEG (8000), 67mM creatine phosphate, 3.2µg/ml creatine kinase, and 50% of cell extract. *E. coli* tRNA was (Roche) were used to supplement *R. erythropolis* tRNA in order to improve protein synthesis. Three microgram of Fos5902 was used as template for the transcription- translation reaction. Fos5902 was isolated from a metagenomic library in

our lab based on its beta-xylosidase activity. The reaction was incubated at 30°C for two hours and then used for a beta-xylosidase activity assay. The activity assay was carried out as described in 3.2.15 using 4-MUX as substrate.

5.1.4. Development of double emulsion and IVC-FACS screening

Double emulsions were developed as described in 4.1.2 except that the commercial *E. coli* CFPS system was substituted by the in-house *R. erythropolis* cell-free reaction mixture (section 5.1.2). The amount of DNA used for *in vitro* transcription-translation was reduced to 0.01ng in a reaction to understand if high DNA input was the source of background DNA (section 4.2.4). Expression of uncloned metagenomic DNA within the emulsion droplets was carried out at 30 °C for 2 hours prior to sorting.

5.1.5. PCR amplifications and cloning of genes isolated through IVC-FACS

PCR amplifications were conducted as described in section 4.1.7. The thermocycle conditions were: 98°C for 30sec, 35 cycles (98°C for 10sec, annealing for 30sec, 72°C for 1min), 72°C for 2 min. Table 15 shows all the primers used and their annealing temperatures. PCR products were purified using the Nucleospin PCR cleanup kit according to manufacturer's recommendations and ligated in pJet 1.2/blunt (4.1.7) and the resulting clone was used to transform *E. coli* JM109 as described in sections 4.1.9 and 4.1.10. Subcloning in pET21a was done as described in section 4.1.12. The same cloning method was used for cloning in pTip-RC1 or pTip-RC4 (Table 14).

Table 15: Primers used in this chapter

Primer name	Sequence	Annealing	Product size	Source
BGCF1	TATTGTCTCAATTTGGGACA	56 °C	1.6kb	This study
BGCF2	TGGATGTTACAGGAAGGTTA	56°C	1.6kb	This study
BGCR	GGATCCTTATTTAAATTCTA	56°C	1.6kb	This study
XYL1	TAGTGTTTGT TTCAGGTGGG	52°C	900bp	This study
XYL2	GTGAGTATTG TTTCCGAGGG	52°C	1.1kb	This study
XYLTR	GGATCCTGGATCAAATAGGA	52°C	1.1kb	This study
XYLB1	CGATAAAACG GCTCCCTCTA	54°C	1kb	This study
XYLBR	GGATCCATGGATCATATTTA	54°C	1kb	This study
BGAXF	GGATCCATGAAGAACTATCAATT	56°C	2.4kb	This study
BGAXR	CTCGAGAAAAGCAGCTTTTCAGCT	56°C	2.4kb	This study
XYLAF	TAACCTCTCT ACATCCAGAG	56°C	1.9kb	This study
XYLAR	GGATCCTTATTCTGATTTC A	56°C	1.9kb	This study
BGLX1	AGCAGGAATGTCTTATAATG	56°C	1.6 kb	This study
BGLX2	ATGCTGCCCTTCGAGATTGC	56°C	1.5kb	This study
BGLXR	GGATCCGATCTTCTCCATCA	56°C	1.5kb	This study



5.1.6. Development of pTip-RC4

To express genes from their native promoter, the *Rhodococcus erythropolis* pTip-RC4 vector was prepared by removing the *p_{tipA}* promoter region of the pTip-RC1 vector (Nakashima & Tamura 2004) by digesting the plasmid DNA with *Bsr*GI and *Nco*I restriction enzymes. One microgram of pTip-RC1 was digested with 2U of *Bsr*GI and 1U of *Nco*I (Fermentas) for 16 hours at 37°C. Digested DNA was electrophoresed in 1% agarose at 60V for 1 hour. The 8.5kb plasmid backbone was excised and purified using the Nucleospin gel purification kit (Macherey-Nagel) according to the manufacturer's recommendations. Purified linear plasmid was blunted using the Klenow Fragment of DNA polymerase I (Fermentas) according to manufacturer's recommendations. Fifty nanograms of blunted plasmid was self-ligated in a 25µl reaction containing 5µl of 10X T4 ligase buffer, 5U of T4 DNA ligase and sterile ddH₂O added to final volume.

5.1.7. Preparation of *R. erythropolis* electro-competent cells

R. erythropolis electro-competent cells were prepared according to a modified method of Treadway *et al.* (1999). Briefly, cells from glycerol stock were plated on a LB agar plate and incubated overnight at 30°C. A single colony was transferred to 5ml LB medium and grown overnight at 30°C. This culture was used to inoculate 100ml of MB medium (5g/l yeast extract, 15 g/l tryptone, 5g/l soytone, and 5g/l NaCl), supplemented with 1.5% glycine and 1.8% sucrose. Cells were grown until OD₆₀₀ reached 1.6, harvested by centrifugation at 8000g for 5 minutes and washed twice with 30ml ice cold Buffer 1 (20mM HEPES pH 7.2 and 15% glycerol). Washed cells were suspended in 2ml ice cold Buffer 2 (20mM HEPES pH 7.2 and 15% glycerol) and stored at -80°C. Electroporation of electro-competent cells was carried out as described in section 3.1.9.

5.1.8. *In vivo* expressions

Expression in *R. erythropolis* was carried out according to the modified protocol of Nakashima & Tamura (2004). Briefly, recombinant cells were grown overnight at 30°C in 5 mL LB broth containing 50mg/ml ampicillin. These cells were used to inoculate 100ml LB broth containing 50mg/ml ampicillin and grown until OD₆₀₀ of 0.6. Where pTip-RC1 was used, one microgram per millilitre of thiostrepton was added to induce expression and cells were harvested at OD₆₀₀ of 2.5, and then lysed by sonication (section 4.1.13). Total cellular proteins were quantified using the Bradford assay (Fermentas) according to the manufacturer's recommendation. Expression in *E. coli* BL21 DE3 was done as described in section 4.1.13.



5.1.9. DNA recovery, sequencing and sequences analysis

Recovery of DNA from dE droplets was described in section 4.2.3. Sequencing and reads assembly was described in section 2.1.4. Sequences analysis and alignments was described in section 3.1.5.

5.2. Results and discussions

5.2.1. Development of actinobacteria based CFE

Although the development of this CFPS system was based on a simple and effective cell-free extract preparation protocol for *E. coli* (Kim and co-workers), disruption of *Rhodococcus* cells was challenging. This is expected for any gram positive bacteria given the structural complexity of their cell wall (Silhavy *et al.* 2010). Higher total protein concentrations were obtained by harvesting and disrupting cell membranes at the mid-log phase of growth contrary to stationary phase as shown in other studies (Nakashima & Tamura 2004).

The ability to use a crude preparation of *R. erythropolis* cell-free extract for *in vitro* transcription-translation was demonstrated by expressing the beta-xylosidase encoding gene from Fos5902 (Table 15). Fos5902 was isolated from a metagenomic library in our lab based on its activity on pNPX. This system was able to express functional beta-xylosidase as shown by activity when using MUX as the substrate (Figure 41). Other proteins, including fosmid encoded xylanase and cellulase were transcribed and translated to active proteins in our lab using this system (results not shown).

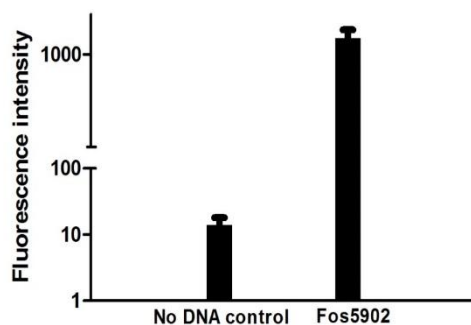


FIGURE 41: MUX hydrolysis by expressing a beta-xylosidase using a novel *R. erythropolis* CFPS system. Transcription-translation was carried out at 30°C for 2 hours using recombinant fosmid DNA carrying a beta-xylosidase encoding gene as template. A reaction without fosmid DNA was used as control. Three microgram of Fos5902 was used in the cell-free transcription- translation reaction.

5.2.2. IVC-FACS using the actinobacteria based CFPS system

The *R. erythropolis* CFPS system was used as transcription-translation machinery in IVC-FACS screening of beta-xylosidase activities using uncloned metagenomic DNA as template. The screening was done using the same principle as described in section 4.2.2. The amount of DNA in the reaction was reduced 1000 fold to investigate if high amount of starting DNA was the source of background DNA experienced in the screening presented in the previous chapter.

Figure 42 shows FACS recording of a million events. Only one event per million was observed in gate M1 of the experiment compared to two events in the no DNA control, suggesting that this event might be a false positive. Gate R1 also had fewer positive events, over six fold when compared to when the *E. coli* system was used. This could indicate that genes in this gate are expressed better in *E. coli* than in *R. erythropolis*. On the contrary, gate N1 has 4110 events in the experiment compared to 2 and 5 in the no DNA and no substrate controls respectively. This is significantly higher than 57 positive events observed when the *E. coli* system was used, suggesting that N1 represents proteins which are difficult to express or inactive in the *E. coli* system under the conditions used. However, these

positive events show low fluorescence intensity, suggesting that their expression is not optimal under conditions used. This may also represent proteins which might have optimal activity at different temperatures and pH. As indicated in the previous chapter, this could also suggest low affinity of these proteins to 4-MUX. While false positive events are present in gate N1, the majority of events in this gate are most possibly true positive.



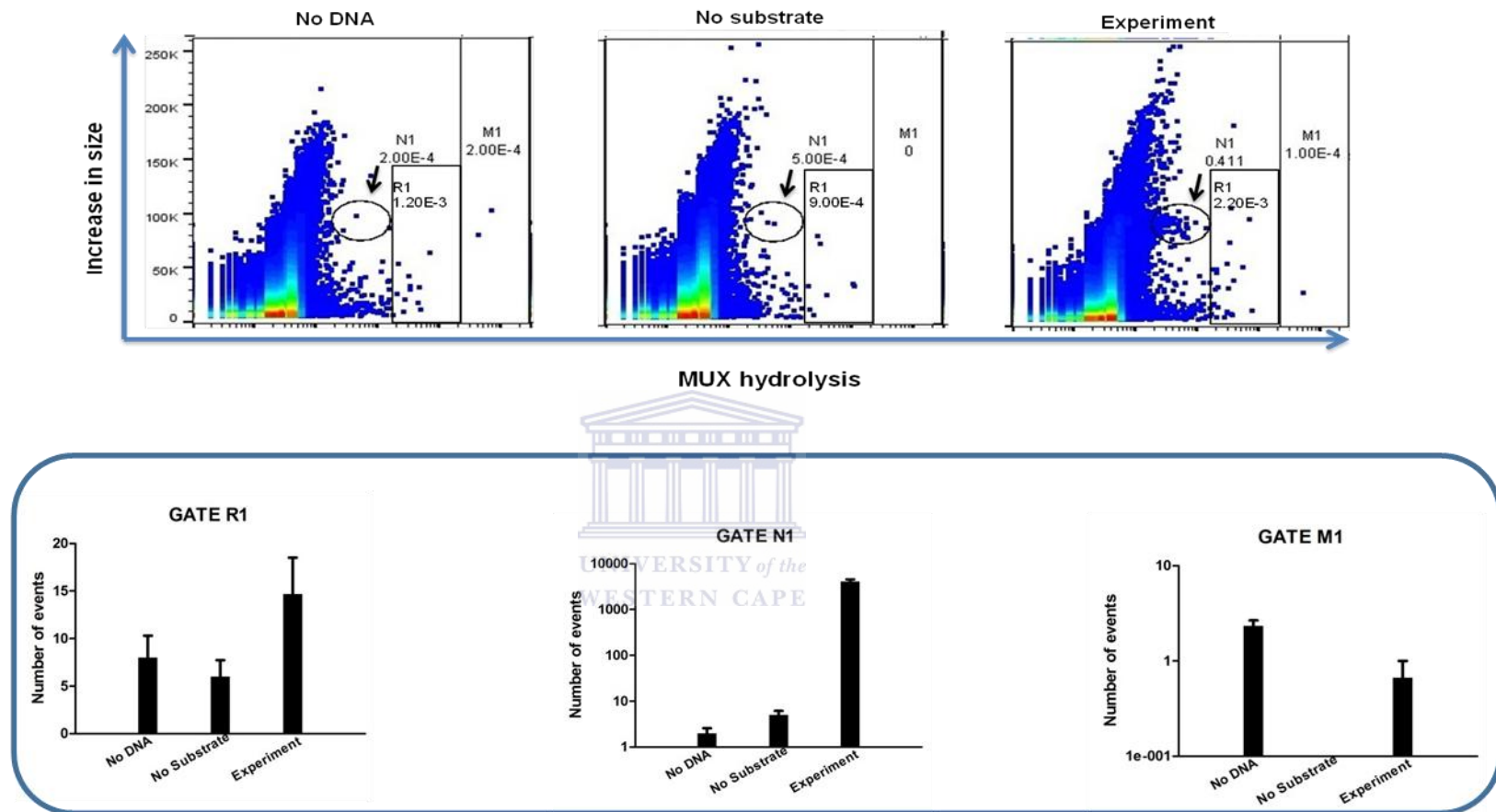


FIGURE 42: FACS recording produced during beta-xylosidase activity screening. 4-MUX was used as fluorogenic substrate and *R.erythropolis* CFPS system was used as transcription translation system. The image shows one million events. Three gates, M1, N1 and R1 represent the average of three rounds of one million events. Error bars are the standard deviation of three rounds of one million events screening.

Since these events are significantly higher than in the *E. coli* system, they were sorted individually for further analysis (Table 16). Nine individual events (dE1-9) were collected and five more events were combined for sequencing and analysis (C5dE). This was done to minimise the costs of sequencing and time required for *in silico* analysis.

5.2.3. DNA recovery and sequencing

DNA was then recovered from the droplets, amplified and sequenced as described in 4.2.3 and 3.2.4. Table 16 summarises the amount of DNA obtained per dE droplet. The amount of DNA recovered per dE droplet was still higher than expected, suggesting that background DNA was not exclusively due to high input DNA. However, this background was reduced by 2.5 fold than what was obtained without dilution. This suggests that dilution of DNA reduced the amount of background DNA. Moreover, quantification of DNA to be used in the reaction needs further optimisation. However, a 2.5 fold decrease in background DNA does not correspond to a 1000 fold dilution of DNA in the reaction, suggesting that other factors are contributing in the generation of background DNA. It should also be noted that the use of different extracts (*R. erythropolis* other than *E. coli*) does not give precise comparison on the level of DNA within dE droplets.

Table 16: Amount of DNA obtained per dE droplet

	C5dE	dE1	dE2	dE3	dE4	dE5	dE6	dE7	dE8	dE9
Total number of reads	3766957	224002	310388	511594	218826	445460	153091	194501	193023	277745
Average read length (bp)	228	285	224	219	220	234	221	216	207	212
Number of de-replicate reads	3503270	206082	282453	439971	203508	423187	145436	175051	175651	236083
Number of reads after de-replicate removal	263687	17920	27935	71623	15318	22273	7655	19450	17372	41662
Number of assembled reads	132550	4962	3201	6623	4111	5812	6531	3609	4143	3918
Number of contigs	3356	998	973	1083	716	990	1445	2103	1513	1215
Average contig length (bp)	988	869	781	778	660	660	713	708	662	629
Number of un-assembled reads	131137	12958	24734	65000	11207	16461	1124	15841	13229	37744
Total DNA(bp)	33214962	4560338	6300311	15077609	2938060	4505274	1278590	4910602	3740023	8765910



5.2.4. Identification of possible beta-xylosidase genes

Identification of possible beta-xylosidase genes from recovered DNA was done as described in section 3.3.5. Interestingly, the majority of sequences (represented on 10 of 14 contigs) identified from different dE droplets were identical (Table 17).

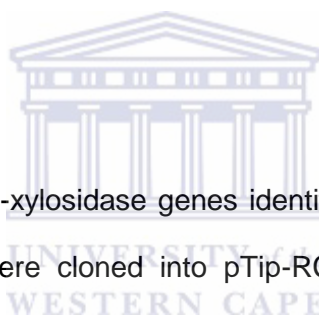
Table 17: GH hits identified from NCBI

dE droplet	ORF name	Closest sequence identity when compared to NCBI blastx search	Identity %	Accession number
dE1	xyIA	xylosidase [<i>Elizabethkingia anophelis</i>]	100	WP_009084683.1
dE2	bgAX	glycosyl hydrolase family 3 [<i>Elizabethkingia anophelis</i>]	100	WP_035589458.1
dE3	xyIT	glucosyltransferase [<i>Pseudomonas fluorescens</i>]	99	WP_003218785.1
dE4	bgAX	glycosyl hydrolase family 3 [<i>Elizabethkingia anophelis</i>]	100	WP_035589458.1
dE5	bgCX	glycoside hydrolase [<i>Elizabethkingia anophelis</i>]	100	WP_059344858.1
dE6	xyIB	Hypothetical protein [<i>Elizabethkingia meningoseptica</i>]	100	WP_021348989.1
dE7	bgAX	glycosyl hydrolase family 3 [<i>Elizabethkingia anophelis</i>]	100	WP_035589458.1
dE8	bgAX	glycosyl hydrolase family 3 [<i>Elizabethkingia anophelis</i>]	100	WP_035589458.1
dE9	bgAX	glycosyl hydrolase family 3 [<i>Elizabethkingia anophelis</i>]	100	WP_035589458.1
C5dE	bgAX	glycosyl hydrolase family 3 [<i>Elizabethkingia anophelis</i>]	100	WP_035589458.1
C5dE	bgAX	glycosyl hydrolase family 3 [<i>Elizabethkingia anophelis</i>]	100	WP_035589458.1
C5dE	bgAX	glycosyl hydrolase family 3 [<i>Elizabethkingia anophelis</i>]	100	WP_035589458.1
C5dE	bgAX	glycosyl hydrolase family 3 [<i>Elizabethkingia anophelis</i>]	100	WP_035589458.1
C5dE	bgAX	glycosyl hydrolase family 3 [<i>Elizabethkingia anophelis</i>]	100	WP_035589458.1

These contigs have high sequence identity to a glycosyl hydrolase family 3 from *Elizabethkingia anopheles* when compared to the NCBI protein database. Five of these contigs were from the combined 5dE events (C5dE) and the other five from individual emulsions. Although dE1, dE3, dE5 and dE6 contain different genes, identification of the *bgAX* in the majority of dE droplets once again demonstrates that the screening is highly selective, as was observed with the *E. coli* system where the *mbgIX* gene was found in all the dE droplets (section 4.2.5). The fact that two different gene expressions were selected for using the 2 different CFPS extracts, suggests that the IVC-FACS system can be used to isolate different genes based on their expression level.

When these hits were traced back to the shotgun metagenome sequence data (Chapter 2), they were not identified from assembled contigs. Some of the possible reasons for this were explained in 4.2.5. Although only nine individual dE droplets and a combined five dE droplets were sequenced, it was interesting to see that BgAX makes almost a quota of all the selected activities. Since the selected and analysed positive activities did not correspond to any of those observed using the *E. coli* system, we hypothesised that *bgAX* might not be expressed into functional proteins or its expression is less efficient with the *E. coli* system, which would result in the events falling elsewhere other than gate N1. This hypothesis was investigated in the following sections.

5.2.5. Cloning of five ORF's



To confirm that the putative beta-xylosidase genes identified from the dE events conferred xylosidase activity the genes were cloned into pTip-RC4 in which the 200bp inducible promoter region (P_{tipA}) was removed (Figure 43), which allows genes to be expressed into active proteins from their native promoters to mimick the circumstances under which the IVC-FACS selection was performed. In the absence of this promoter, it is possible to establish whether selected ORFs can be transcribed and translated from their native promoters.

pTip vectors were developed by Nakashima & Tamura (2004) as *E. coli-Rhodococcus* shuttle vectors (Figure 43). These vectors can mediate heterologous protein expression, as they contain inducible promoters, a multiple cloning site with 11 restriction enzyme sites and a hexahistidine (six-His) tag sequence. They also work over a wide temperature range, from 4 to 35°C (Nakashima & Tamura 2004). These vectors carry the *repAB* operon of the cryptic plasmid pRE2895, which is necessary for the autonomous replication of the plasmids

in *Rhodococcus*. The *repAB* operon encodes the replication proteins RepA and RepB, which are characteristic of pAL5000-type plasmids (Nakashima & Tamura 2004; Stolt & Stoker 1996). The replication mechanisms of pAL5000 and pRE2895 are unknown, but RepA proteins of pAL5000 and pRE2895 are similar to Rep proteins of ColE2 plasmids (Hiraga *et al.* 1994), suggesting that they may replicate by a θ -type mechanism (Stolt & Stoker 1996). ColE2 is a known plasmid replication protein for *E. coli*, allowing pTip vectors to replicate in both *E. coli* and *Rhodococcus*.

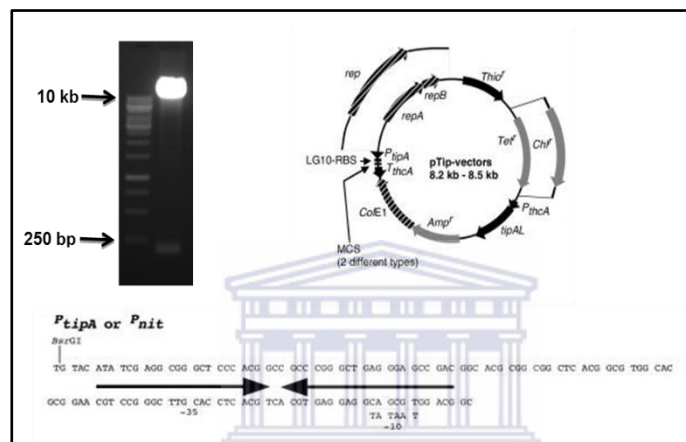


FIGURE 43: Agarose gel electrophoresis images showing pTip-RC1 after digestion with *BsrGI* and *NcoI* restriction enzymes (left). Schematic representation of pTip vectors with the promoter region and ribosomal binding site (right).

DNA recovered from dE1, dE3, dE5, dE6 and C5dE droplets was used as template for amplification of respective ORFs with their promoter region. The amplicons were cloned into pTip-RC4 between the *Bam*HI and *Xho*I region for expression in *E. coli* and *R. erythropolis* (Figure 44).

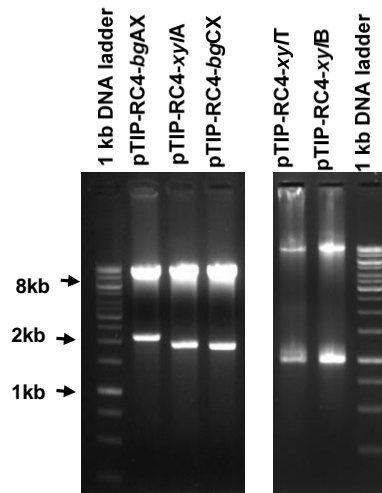


FIGURE 44: Agarose gel image of double digested pTIP-RC4 harbouring genes recovered from dE droplets. The plasmids were double digested with *Bam*HI and *Xho*I.

5.2.6. *In vivo* expression in *E. coli* and *R. erythropolis*

ORF's (XylA, XylB, XylIT, BgCX and BgAX; Table 17) cloned into pTip-RC4 were expressed in both *E. coli* and *R. erythropolis*. Expression was performed at 30°C to allow these genes a chance to fold correctly. Figure 45 shows that *R. erythropolis* is able to express all genes into active proteins under their native promoter. However, in *E. coli* only XylA and XylB show appreciable activity. Moreover, as hypothesised above, BgAX only shows low activity when using an *E. coli* system compared to a *R. erythropolis* system under the conditions used. This suggests that the activity of this gene would not fall in gate N1. Although the reason for low activity of these proteins in *E. coli* is not clear, it might be due to the difference in cellular environments such as turnover rates of proteins and/or folding machinery differences between *E. coli* and *R. erythropolis* cells (Nakashima & Tamura 2004). Moreover, it could be that the promoter region of these genes is weakly recognised by the *E. coli* system.

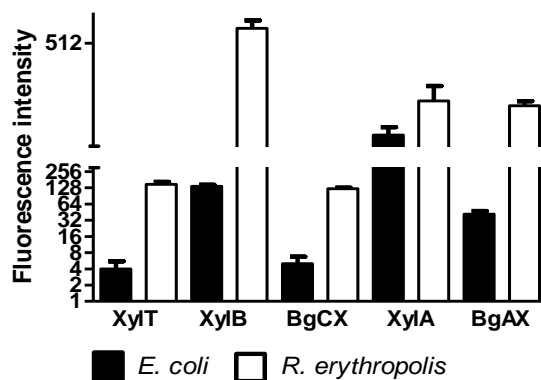


FIGURE 45: Bar charts showing increase in fluorescence intensity as a result of MUX hydrolysis. Five beta-xylosidases ORFs were expressed under their native promoter in *E. coli* and *R. erythropolis*. One hundred micrograms of whole cellular protein was used for enzyme assay. Cells grown with pTIp-RC4 were used as negative control. The plotted values were first adjusted by subtracting values obtained from the negative controls.

To differentiate between promoter recognition and/or promoter strength vs protein folding and activity as the reasons for low activity of these genes, *bgCX*, *bgAX* and *xyIT* were amplified without their promoter regions and cloned in pET21a vector for expression from the T7 promoter. BgAX showed some improved activity when expressed under the control of the T7 promoter (Figure 46), suggesting that weak promoter recognition was partly the reason for lack of activity in gate M1.

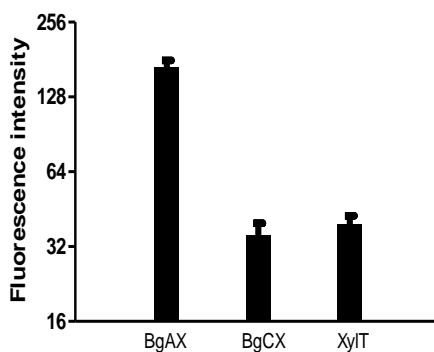


FIGURE 46: Fluorescence intensity as a result of MUX hydrolysis for BgAX, BgCX and XylT when expressed from the T7 promoter in *E. coli*. One hundred micrograms of whole cellular protein was used for enzyme assay. Cells grown with pET21a were used as negative control. The plotted values were first adjusted by subtracting values obtained from the negative controls.

Although there is no clear reason for some genes to express in *R. erythropolis* and not in *E. coli* and *vice versa*, there is evidence in the literature which shows that codon usage influences the expression of genes in heterologous hosts (Sastalla *et al.* 2009). When the codon bias of the gene to be expressed differs significantly from that used by the host strain, the concentration of the host's tRNAs for the lesser used codons are insufficient to optimally translate the RNA. Consequently, translational errors such as stalling, termination, amino acid substitution and possibly frame shifting may occur resulting in low expression of target protein and protein folding issues (Chen & Inouye 1994). Competition for rare tRNA may also adversely affect the expression of host genes or elicit a stringent response (Stoletzki & Eyre-walker 2007).

Furthermore, codon preference has been reported to be influenced by GC content of the host organism. Zyllicz-Stachula *et al.* (2014) have demonstrated that high GC content negatively effects expression efficiency of the *taqIIIRM* gene in *E. coli*. Sastalla *et al.* (2009) also optimised the GC content of several fluorescence protein genes in order to achieve high expression in low GC gram positive bacteria.

In this study, analysis of the nucleotide content of these five ORF's showed that they have an average GC content of 37.58, lower than that of *R. erythropolis* strains which have an average GC content of 62.31. Translation of mRNA to active protein through this CFPS system solely depends on the native *Rhodococcus* tRNAs anticodons and their respective transferases, which could have resulted in a bias towards the expression of high GC content genes. However, because *E. coli* tRNAs were added to the reaction (4.2.2), this may have counter-balanced the bias and even resulted in a bias towards low GC-genes. As a result, rare codon substitution may occur in low GC content genes, altering protein folding and subsequently reduced activity. This was also demonstrated by Komar *et al.* (1999) who showed that the removal of rare codons can reduce the specific activity of chloramphenicol acetyltransferase.

In addition, XylIT has a rare start codon GTG (Appendix) which may have affected translation of this gene in both *R. erythropolis* and *E. coli*. Reddy *et al.* (1985) investigated the effect of varying the initiation codon on the expression of the adenylate cyclase (*cya*) gene. Using oligonucleotide-directed mutagenesis, they changed the UUG initiation codon to GUG and the more common initiator AUG and assayed for *cya* gene expression. They found that the GUG initiation codon, in place of UUG, doubled *cya* expression when *cya* was expressed from the dual *cya* P1/P2 promoters. The corresponding AUG codon construct was nonviable. When the *cya* gene was placed under the transcriptional control of the thermos-inducible phage A PL promoter, the relative amounts of *cya* gene product were 1:2:6 for the UUG, GUG, and AUG initiation codons, respectively.

Annotation of ten different *E. coli* strains revealed that 82.5% of the start codons were ATG, 12.3% were GTG and 5% were TTG, with CTG, ATT and ATC used at lower frequencies (Villegas & Kropinski 2008). Translational efficiency has been shown to decrease in *E. coli* when a start codon other than ATG is used, with an eight-fold reduction in translation seen with GTG or TTG start codons. Myronovskyi *et al.* (2011) studied the efficiency of translation of the *gusA* gene initiating at ATG, GTG, TTG and CTG start codon in actinobacteria. Surprisingly, constructs using a TTG start codon showed the best activity, whereas those using ATG or GTG were approximately one-half or one-third less active, respectively. This may suggest that the preference of start codon in actinobacteria might be different to that of *E. coli*, which might explain the expression of XylIT in *R. erythropolis* and not in *E. coli*.

5.2.7. MBglX cannot be expressed by the *R. erythropolis* system

Compared to the *E. coli* system screening, the *R. erythropolis* system does not generate as many positive events in gate M1 (a frequency of 8 per million were observed in this gate

when *E. coli* system was used and no clear positive was observed in *R. erythropolis* experiment). This suggests that *mbglX*, which was sorted in this gate through the *E. coli* screening cannot be expressed into functional protein using *R. erythropolis* system. To confirm this hypothesis, *mbglX* was amplified and cloned in pTIP-RC4 to allow expression of this gene under its native promoter in a *R. erythropolis* cellular background (Figure 47).

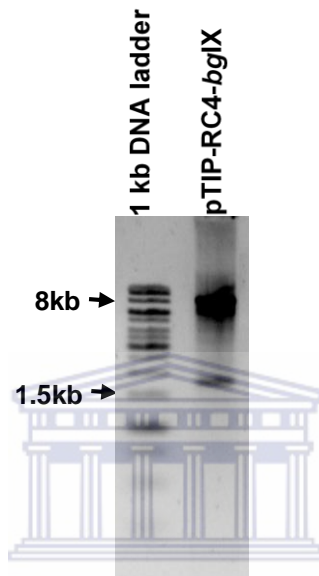


FIGURE 47: Agarose gel image of double digested pTIP-RC4 harbouring *mbglX*. The plasmid was digested with *Bam*HI and *Xho*I.

The resulting clone was used to transform *E. coli* and *R. erythropolis* for expression in both strains. Figure 48 demonstrates that *BglX* can be expressed in significant quantity in *E. coli* while the activity of this protein is negligible in *R. erythropolis*. This could be due to either *R. erythropolis* not recognising the promoter of this gene, or the production of non-functional protein.

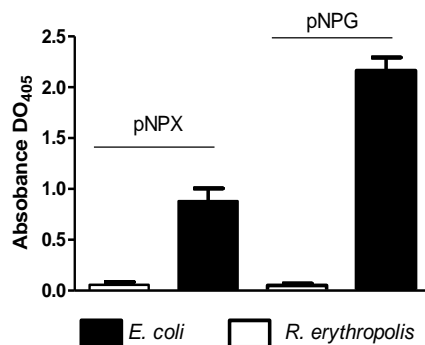


FIGURE 48: *In vivo* expression of *bgIX* in *E. coli* and *R. erythropolis* under its native promoter and assessment of enzyme activity. One hundred micrograms of whole cellular protein was used for enzyme assay. Cells grown with pTip-RC4 were used as negative control. The plotted values were first adjusted by subtracting values obtained from the negative controls.

5.2.8. Sequence analysis of genes isolated through the *R. erythropolis* based CFPS system

To get an insight on the nature of genes isolated in this chapter, all five genes were analysed at the sequence level and compared with other published genes. In particular, it was important to establish if mIVC-FACS is able to isolate novel genes which cannot be identified through traditional function-based screening. In principle, this should be possible given the high throughput nature of this screening technique compared to traditional agar plate or liquid-based screening methods. In IVC-FACS, as has been shown in this study and elsewhere (Bernath *et al.* 2009), 20 000 genes can be screened in one minute. In traditional screening methods using robotic systems, this can take up to days. In addition, DNA fragments occurring in low abundance, which are likely to be lost during cloning and library construction, are more likely to be screened through IVC-FACS since there is no significant loss of DNA. This increases the probabilities of isolating novel hits.

5.2.8.1. Analysis of XylA and XylB

Conserved domain analysis of XylA and XylB on NCBI domain databases shows that these proteins contain a GH43 domain. The GH43 family includes enzymes with beta-1,3-xylosidase (EC 3.2.1.-), alpha-L-arabinofuranosidase (EC 3.2.1.55), arabinanase (EC 3.2.1.99), xylanase (EC 3.2.1.8), endo-alpha-L-arabinanase and galactan 1,3-beta-galactosidase (EC 3.2.1.145) activities (Matsuzawa *et al.* 2015). These are inverting enzymes that have an aspartate as the catalytic general base, and a glutamate as the catalytic general acid and another aspartate that is responsible for pKa modulation and orienting the catalytic acid (Barker *et al.* 2010). Some of the enzymes in this family display both alpha-L-arabinofuranosidase and beta-D-xylosidase activity (Brüx *et al.* 2006).

Sequence alignments of XylA with other GH43 proteins confirm that this protein has similar catalytic residues found in GH43 proteins (Figure 49). According to the enzyme–substrate complexes structure of an inverting GH43 beta-xylosidase from *Geobacillus stearothermophilus* (Brüx *et al.* 2006), the general residue, Asp15 is located between two conserved proline residues that provide the exact orientation required for the catalytic activity of this acidic residue. The location of the aspartic acid residues in proteins used for this alignment, including XylA, was not between proline residues, but between an alanine and a proline residue (Figure 49). In addition, XylB does not align with any of the sequences in the alignment, suggesting that this protein is completely different. This is also shown in the phylogenetic tree (Figure 49).

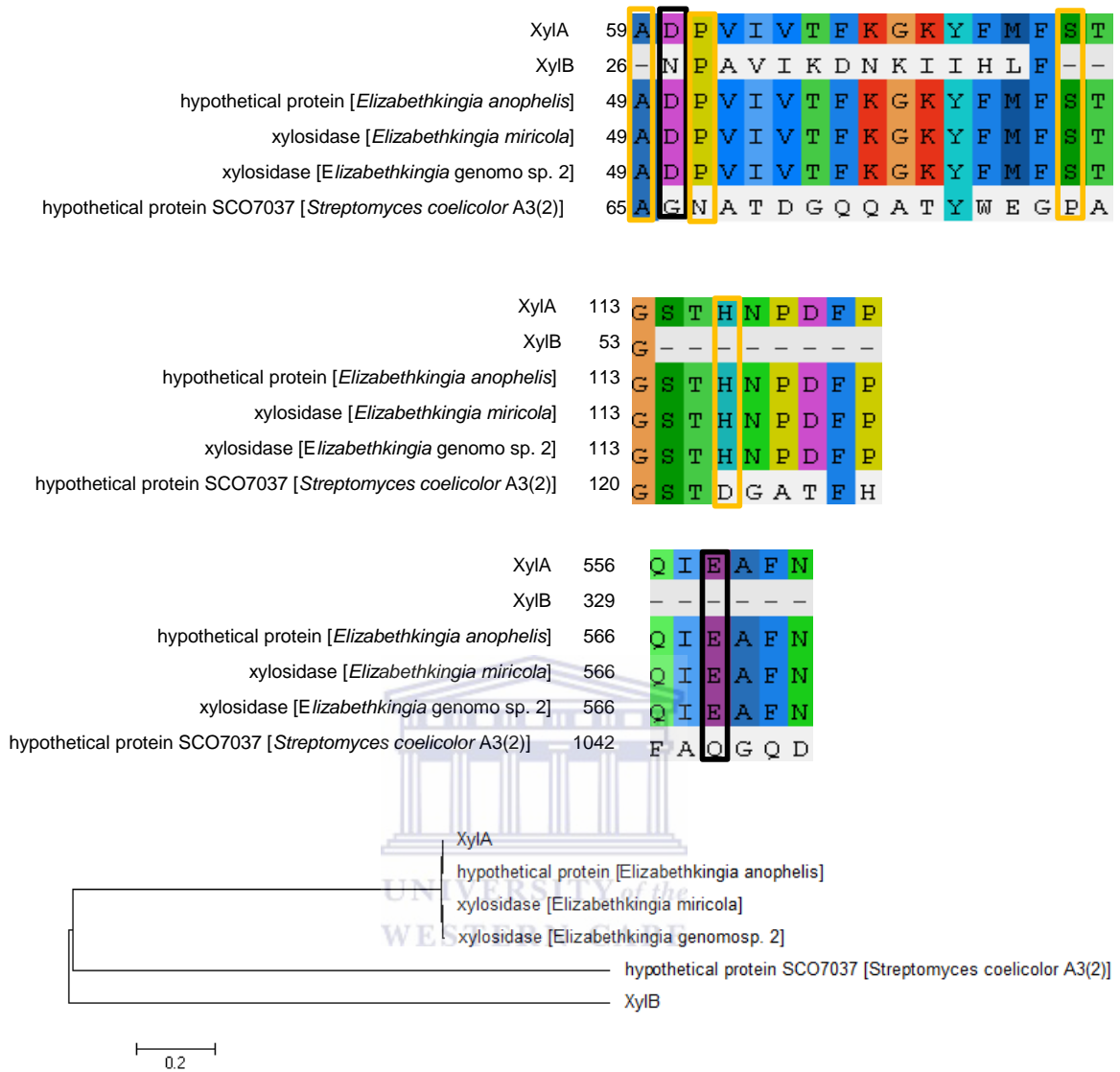


FIGURE 49: Sequence alignments of XylA and XylB with other GH43 proteins. Residues in black blocks show catalytic residues and orange blocks show known conserved residues. Three best matches in the NCBI proteins database together with the two best matches from well-studied reference species were selected. Analyses were done using the SmartBlast tool in the NCBI server.

5.2.8.2. BgAX has the general GH3 catalytic motifs

BgAX has the structural domain architecture of general GH3 glycosyl hydrolases. Unlike BgIX which has a glutamic acid substitution in the catalytic site (section 4.2.7), the catalytic residues of this protein are conserved as for other GH3 proteins (Figure 50).

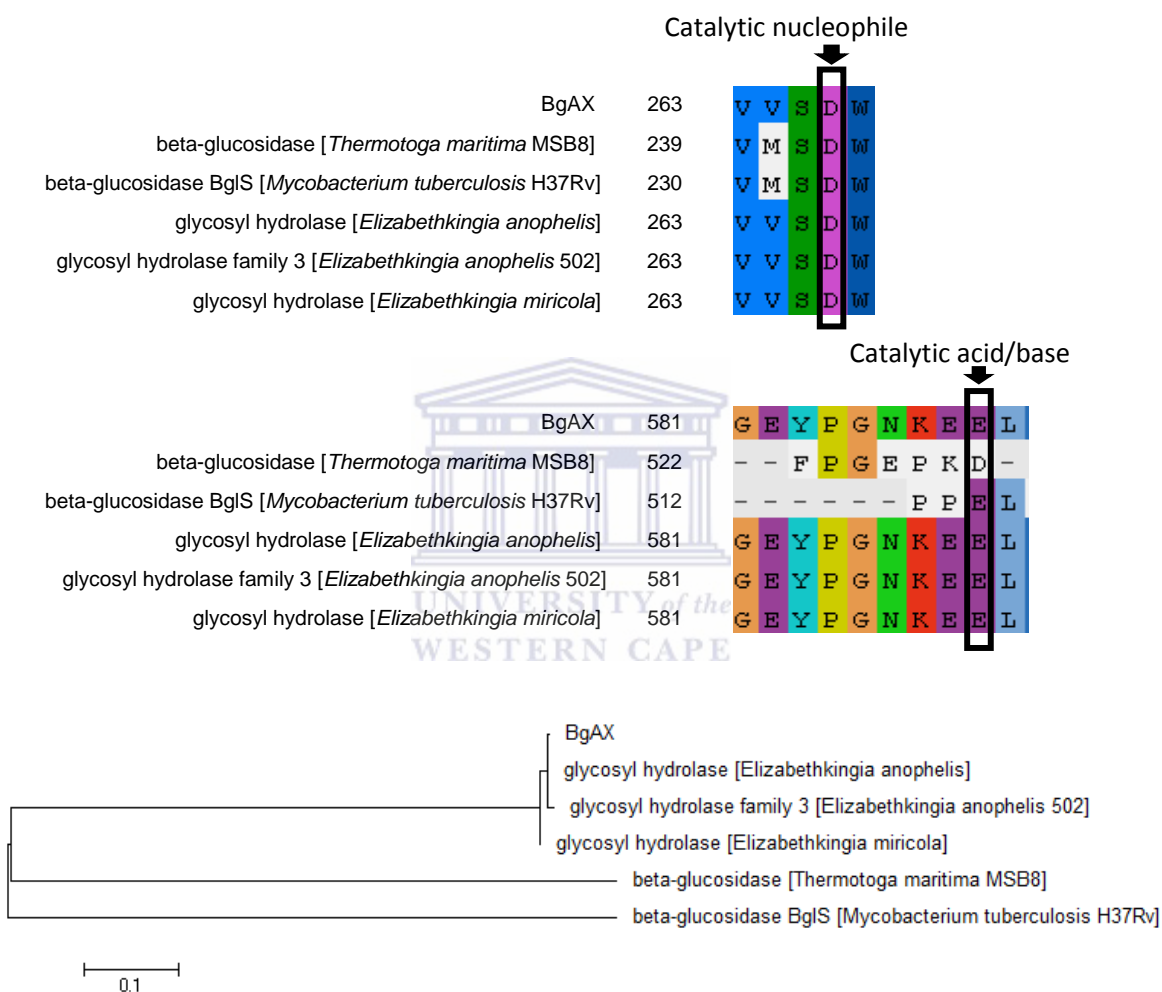


FIGURE 50: Sequence alignment of BgAX with other GH3 proteins. Residues in black blocks show catalytic residues and orange blocks show known conserved residues. Three best matches in the NCBI proteins database together with the two best matches from well-studied reference species. Analysis where done using SmartBlast tool in the NCBI server.

It has been suggested that the glutamic acid substitution inMBgIX (Chapter 4) might explain its dual activity (beta-xylosidase and beta-glucosidase). The conservation of glutamic acid in BgAX might suggest that this enzyme cannot hydrolyse glucosides.

5.2.8.3. XylIT is a transferase/beta-xylosidase

Sequence analysis of XylIT on the NCBI conserved domain server showed that this protein has a conserved domain belonging to the glycosyltransferase family A (GT-A) (Figure 51). These enzymes synthesise oligosaccharides, polysaccharides, and glycoconjugates by transferring the sugar moiety from an activated nucleotide-sugar donor to an acceptor molecule, which may be a growing oligosaccharide, a lipid, or a protein (Coutinho *et al.* 2003). The majority of the proteins in this superfamily are glycosyltransferase family 2 (GT-2) proteins, but they also include families GT-43, GT-6, GT-8, GT13 and GT-7; which are evolutionarily related to GT-2 and share structural similarities (Bourne & Henrissat 2001).

Glycosyltransferases with hydrolase activity have been reported but are extremely rare. Hudson *et al.* (1993) isolated a beta-xylosidase from the anaerobic thermophilic bacterium *Caldocellum saccharolyticum* and expressed it in *E. coli*. While this enzyme shows beta-xylosidase activity, it is most likely to be a xylose transferase rather than a true beta-xylosidase because it could use neither hydrolyse xylobiose nor triose as a substrate, although it hydrolyses aryl-xylosides. Another highly thermostable glycosyltransferase with beta-xylosidase activity and an unusually high arabinosidase activity was isolated from the thermophilic bacteria *Thermoanaerobacter ethanolicus* by Shao & Wiegel (1992).

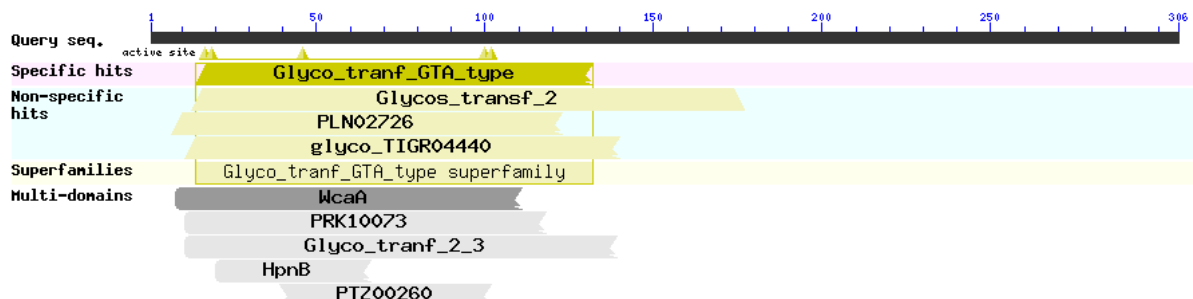


FIGURE 51: Domain architecture of XylIT. The domain architecture was determined by the NCBI blastx server using conserved domain tool.

While transferases with beta-xylosidase activity are rare, many biologically important compounds including various oligosaccharides, glycoconjugates and neoglycoproteins can readily be synthesised using the transglycosylating potency of glycosidases. Microbial beta-xylosidases have been utilized in enzymatic synthesis of high molecular weight polymers in transxylosylation reactions. For instance, Kurakake *et al.* (2005) used the *Aspergillus awamori* K4 beta-xylosidase to carry out transxylosylation with xylobiose and xylotriose and alcohol. The hydrolysis rate of xylobiose was much lower than the transxylosylation rate at the initial stage, decreasing gradually as the substrate concentration increased, whereas the transxylosylation rate increased greatly. This beta-xylosidase had broad acceptor specificity toward alcohols, hydroxybenzenealcohols, sugar alcohols and disaccharides.

The enzymatic mechanism of transxylosylation is less understood and the action of catalytic groups in the active site and acceptor specificity remain largely unknown (Kurakake *et al.* 2005). While hydrolase activity of XylIT has been demonstrated (section 4.3.8), transxylosylation experiments are required to establish if XylIT can translocate xylose from one molecule to another. The hydrolase activity of XylIT was significantly low compared to true beta-xylosidases isolated in this study (Figure 45) and elsewhere (Lama *et al.* 2004). While hydrolysis efficiency cross-comparison of XylIT with reported proteins with transxylosylation activity is not possible due to different experimental conditions, the presence of the transferase domain suggests that this enzyme is a family 2 transferase which displays hydrolase activity.

Based on mutagenesis studies, the catalytic motif of GT2 members has been shown to be aspartic acid located in the sequence DXD and the glutamic acid within the sequence ED(Y) (Saxena *et al.* 1995). However, the exact number of Aspresidues has been shown to vary. In other GT2 proteins, this DXD sequence is represented by the sequence DXDD (Charnock & Davies 1999).

Figure 52 below shows that unlike other GT2 proteins, this sequence is represented by DXDA in XylIT. Keenleyside *et al.* (2001) have demonstrated that changing a D93A and D96A mutations completely abolished the activity of the mutant, while D95A led to significantly reduced activity. However, the D104A mutation of XylIT does not abolish beta-xylosidase activity, although this activity was significantly low. Indeed, the activity demonstrated by Keenleyside and co workers was transferase rather than beta-xylosidase, which might suggest a completely different effect of beta-xylosidase activity of the protein.

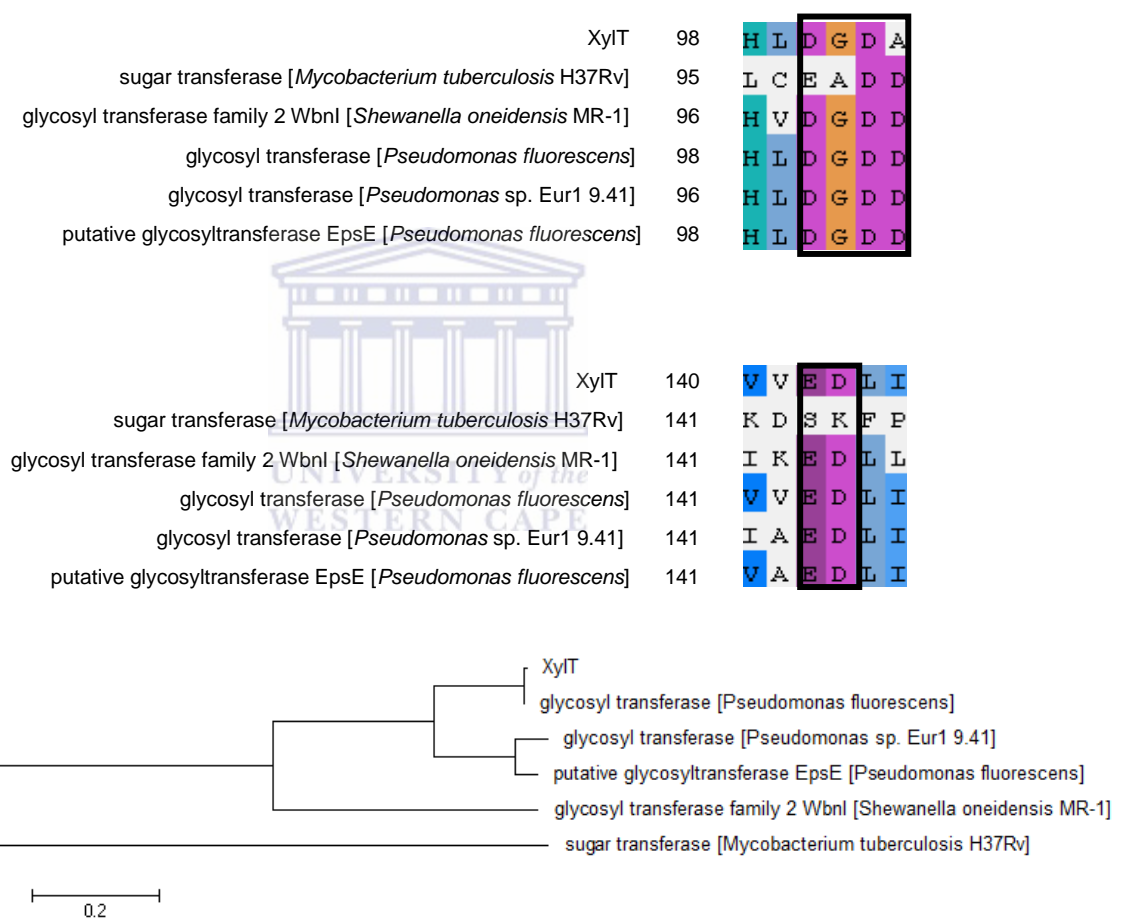


Figure 52: Multiple sequence alignment of XylIT. Highlighted block show known conserved residues of GT2 proteins. Residues in black blocks show catalytic residues and orange blocks show known conserved residues. Three best matches in the NCBI proteins database together with the two best matches from well-studied reference species. Analysis where done using SmartBlast tool in the NCBI server.

5.2.8.4. Catalytic residues of BgCX

The domain architecture of BgCX suggests that this enzyme belongs to the GH5 family. There is no known beta-xylosidase that has been reported from this family, making this protein the first GH5 protein with beta-xylosidase activity. This family includes mainly cellulose degrading enzymes such as beta-glucosidase (EC 3.2.1.21); exo-beta-1,4-glucanase (EC 3.2.1.74); 6-phospho-beta-glucosidase (EC 3.2.1.86); strictosidine beta-glucosidase (EC 3.2.1.105) etc.

GH5 enzymes use the classical Koshland double-displacement mechanism and the two catalytic residues (catalytic nucleophile and general acid/base) for catalysis. Two glutamates found at the C-terminal end of β -strands 4 (acid/base) and 7 (nucleophile) have been shown to be responsible in catalysis. In the nucleotide sequence of the *Clostridium thermocellum* gene *bglA*, coding for the thermostable beta-glucosidase A, a distinctive feature of the catalytic domain was suggested to be the sequence motif H – NEP in which the catalytic residue Glu is separated from His residue by 35–55 amino acid residues (Grabnitz *et al.* 1991). Sequence analyses and mutagenesis study of the *Erwinia chrysanthemi* endo-glucanase (EGZ) also suggested that the H98 residue is involved in the folding of the catalytic domain while the E33 residue intervenes directly in the beta 1–4 glycosidic bond cleavage (Py *et al.* 1991). The catalytic nucleophile of GH5 proteins has been reported to be E280 (Henrissat *et al.* 1995).

Analysis of a sequence alignment of BgCX with GH5 family proteins shows that the motif NEP is also present (Figure 53). However, the conserved nucleophile residue is also replaced by aspartic acid. Although aspartic acid can also act as a nucleophile residue, it is unlikely that BgCX follows the catalytic mechanism of GH5 proteins. Structural analysis and mutagenesis studies would provide a better understanding of the catalytic mechanisms and the residues involved in the catalysis of BgCX.

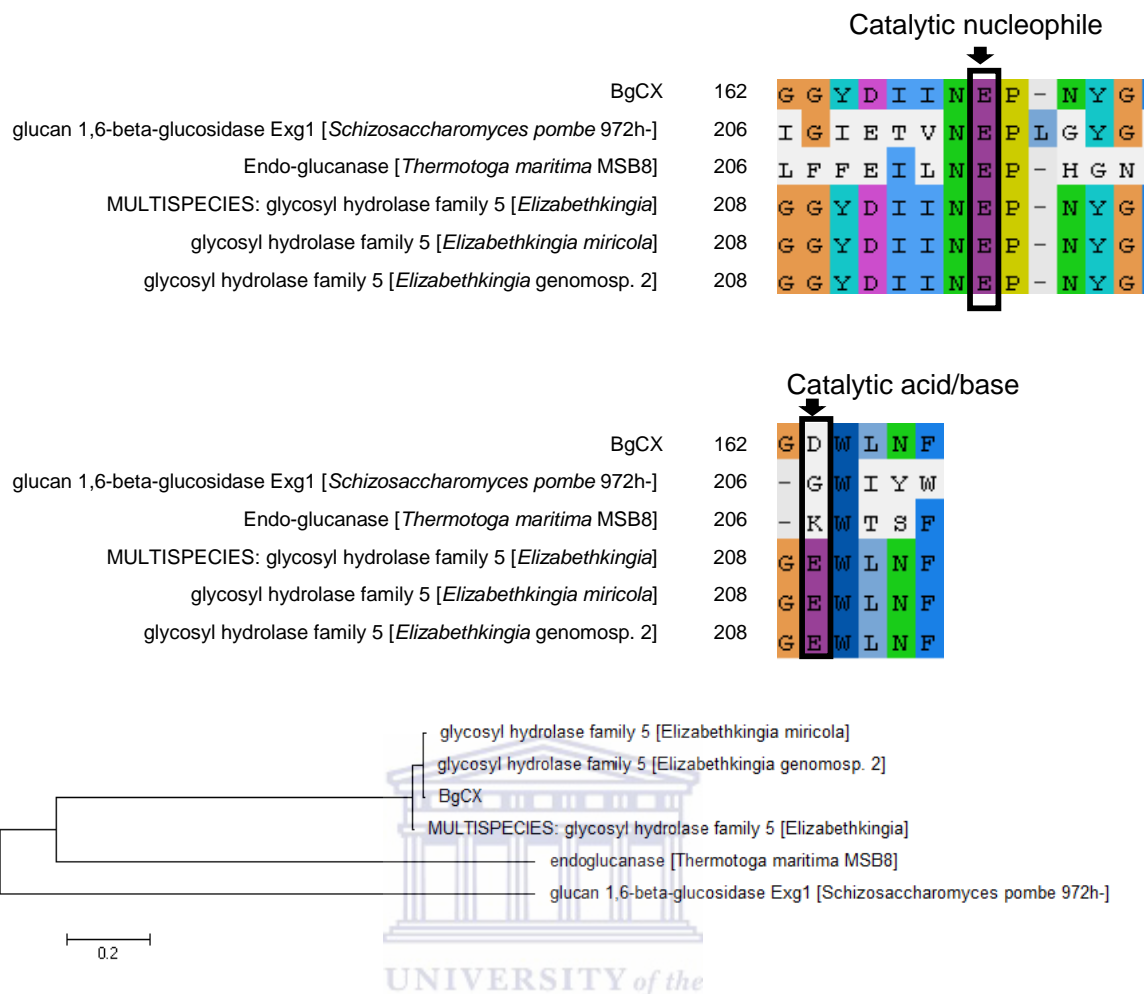
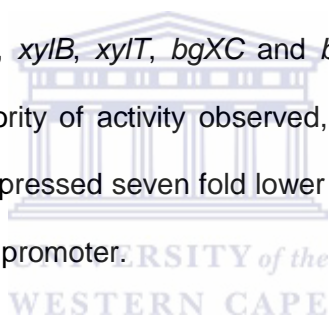


FIGURE 53: Alignments of catalytic residues of GH5 enzymes with BgCX. Highlighted block show known conserved residues of GH5 proteins. Residues in black blocks show catalytic residues and orange blocks show known conserved residues. Three best matches in the NCBI proteins database together with the two best matches from well-studied reference species. Analysis was done using SmartBlast tool in the NCBI server.

5.3. Conclusion

In this chapter, a *R. erythropolis* CFPS system was developed and used in the IVC-FACS screening of beta-xylosidases based on fluorescence activity. Although the number of analysed dE droplets were few, the results obtained here suggests that IVC-FACS shows selectivity in the beta-xylosidases isolated based on the cell-free system used. This was demonstrated in 2 different ways: i) *bgIX*, which was isolated in gate M1 through *E. coli* based expression could not be identified when using *R. erythropolis* CFPS. This was further demonstrated when BglX was subsequently shown, through *in vivo* expression under its natural promoter in both *E. coli* and *R. erythropolis*, to not be expressed into functional protein in *R. erythropolis*. ii) The *R. erythropolis* CFPS selected five different beta-xylosidase encoding genes (*xylA*, *xylB*, *xylT*, *bgXC* and *bgAX*). The *in vivo* expression of *bgAX*, which makes up the majority of activity observed, was not identified by the *E. coli* system, and revealed that it is expressed seven fold lower in *E. coli* compared to expression in *R. erythropolis* under its native promoter.



Although the evidence above strongly suggests that IVC-FACS can be used to select for a gene of interest from a mixture of microbial genes, there are challenges which still limits this technology. In particular, background DNA still needs to be addressed so that only a desired DNA fragment and its phenotype can be recovered from individual dE droplets. In an attempt to eliminate this background DNA, the amount of input DNA was reduced 1000-fold in the initial reaction mixture. While some improvement was observed, background DNA was still significantly high.

The application of ddPCR technology has been suggested in Chapter 3. Automated microdroplet generators such that used in ddPCR can be used to sort DNA molecules evenly within the droplets in such a way that only a single DNA molecule is partitioned within a droplet (Hindson *et al.* 2011). Although this technology only generates single emulsion

droplets that have been reported to be incompatible with FACS screening, the use of more stable oil might overcome this challenge. In addition, this technology is able to quantify DNA molecules more accurately and efficiently than real-time PCR and agarose gel electrophoresis quantification (Hindson *et al.* 2011). This can allow accurate quantification of DNA molecules from the original sample, thereby giving precise concentration to be distributed within the droplets.

Studies elsewhere employ the use of multiple host systems for the parallel screening of a metagenomic library to overcome limitations associated with heterologous expression in *E. coli* (Troeschel *et al.* 2010). mIVC-FACS offers an alternative that does not require development of shuttle vectors that are able to replicate in multiple hosts, which represents one of the major limitations to classic functional metagenomic screening. Since no cloning is required, and no active cells are necessary, it is conceivable that a multi species cell-free cocktail can be developed to expand the environmental genes that can be transcribed and translated in a single mIVC-FACS run. This chapter laid a significant foundation towards achieving this possibility through the successful use of the first actinobacterial and non-*E. coli* derived cell-free protein synthesis system. Genes that cannot be expressed into functional protein in *E. coli* were successfully expressed in *R. erythropolis* and vice versa, suggesting that combining transcription-translation machinery of multiple organisms can offer an efficient system which is able to express different genes into functional proteins, thereby improving the metagenomic screening hit rate.

Although further characterisation of isolated genes still needs to be done, isolation of genes that could not be identified in shotgun metagenomic sequence data suggests that mIVC-FACS can be used to identify rare genes from environmental samples. This can be attributed to a number of advantages that this platform offers over sequence-based screening and classic function-based agar plate screening systems. These include the high throughput

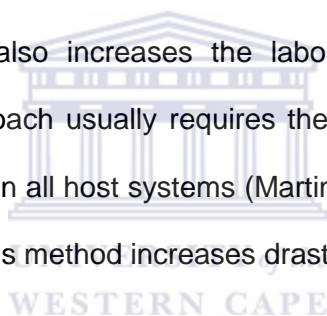
nature of this platform which allows screening of 20000 events per minute and which is independent of the construction of a clone library.

Sequence analysis of identified genes revealed interesting residues in the active site of some of these genes. In particular, the conserved nucleophile residue in BgCX was replaced by aspartic acid compared to other GH5 proteins, suggesting that a different catalytic mechanism is used by BgCX. Further studies, including structural analysis and mutagenesis studies are required to better understand catalysis of BgCX.



CHAPTER SIX: GENERAL DISCUSSION AND CONCLUSION

The application of metagenomics in the discovery of novel biocatalysts from natural materials has been a milestone achievement in biotechnology and microbiology. However, there are still technical hindrances that limit this technology. Different strategies have been developed to eliminate some of these technical hurdles in order to improve access to novel biocatalysts and important microbial compounds. For instance, Craig *et al.* (2010) reported the successful use of six different proteobacteria as heterologous hosts to overcome the limitation associated with the use of domesticated *E. coli* as the only heterologous expression host. While this approach shows the usefulness of multiple hosts for overcoming host expression related barriers, it also increases the labour burden required to complete screening. In addition, this approach usually requires the development of suitable shuttle vectors which express efficiently in all host systems (Martinez *et al.* 2004) and consequently, the costs and time required for this method increases drastically.



Recently, Colin *et al.* (2015) reported the use of FACS to improve the throughput and screening coverage of metagenomic clones. Since the probability of identifying a desired clone among thousands to millions of others is generally low, high-throughput screening (HTS) protocols such as FACS can greatly improve the chances of obtaining an active clone by allowing higher numbers of clones to be screened simultaneously. Despite this advantage of FACS screening, functional metagenomic approaches still face other limitations. To successfully identify a gene or protein candidate of interest, a series of sequential steps in the cloning and screening process must occur adequately and effectively. Transcription of the entire gene, translation of its mRNA, correct protein folding, and secretion of the active protein from the surrogate host must all be achieved before functional screening even begins.

This study is reporting the first *in vitro* functional metagenomic screening system. Here, uncloned environmental DNA was used to minimise the loss of DNA during cloning and transfection. The general overview of the work done in this study is represented in Figure 54.

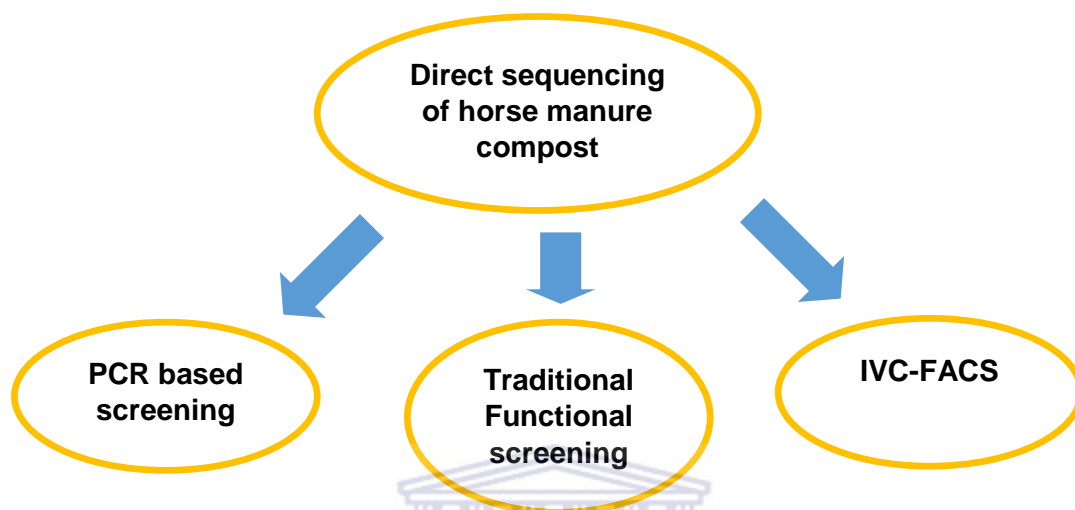


FIGURE 54: General overview of this study. High molecular weight metagenomic DNA was isolated and purified from composting horse manure. The DNA was first assessed for microbial diversity and GHs content before PCR screening, clone library screening and IVC-FACS.

In addition to challenges associated with heterologous expression, another shortcoming of functional screening approaches is that even large libraries only provide a small fraction of the environmental diversity and the low frequency of novel enzyme encoding genes among vast unrelated DNA sequences (Ferrer *et al.* 2009). Raes *et al.* (2007) reported that a pristine soil sample contains more than 10^4 different microbial species and over one million open reading frames (ORFs), many of which encode putative enzymes. The target genes encoding for novel enzymes represent a tiny fraction of these ORFs, in some cases less than 0.01% of the total nucleic acid sample extracted from environment total sources (Gabor *et al.* 2007). In Chapter 2 of this study, it was shown that only 0.28% of 829Mb environmental DNA derived from horse manure contained sequences with homology to GHs. When analysed for beta-xylosidase encoding sequences, only 0.0002% of the sequence dataset represented beta-xylosidase encoding sequences. While further sequencing of this sample is still required for complete coverage, it is clear that this fraction can be significantly

reduced during the cloning and transfection steps. For this reason, screening of uncloned metagenomic DNA increases the probability of identifying rare genes within diverse environmental DNA.

Screening of uncloned metagenomic DNA was facilitated by the use of cell-free transcription/translation systems in double emulsion droplets, followed by high throughput FACS sorting. IVC-FACS is routinely used in different laboratories where the linkage between a gene and the product it encodes has to be maintained within a single emulsion droplet (Stapleton & Swartz 2010). However, the selection of an uncloned metagenomic DNA fragment and the proteins encoded within this fragment is more challenging than the selection of a plasmid encoded gene and its product. As a result, we have reported in chapter 4 and 5 that the major difficulty with mIVC-FACS is high level of background DNA. The background DNA is of no consequence when plasmid DNA is being isolated since specific primers are normally used to isolate the desired gene. Indeed, the presence of background DNA is a major challenge that needs to be addressed if mIVC-FACS is to be applied routinely for metagenomic studies. In the presence of background DNA, the linkage between the gene of interest and its product is generally unclear.

In this study, we reasoned that even in the presence of background DNA, the activity observed during FACS screening and the gene that encodes for that activity is still within the dE droplet that was selected based on the observed activity. The high background sequence greatly affects the integrity of the screening system in general. However, results obtained suggested that we have indeed selected for beta-xylosidase encoding genes based on the activity we observed during FACS screening, as opposed to the random *in silico* identification of beta-xylosidases in amongst the high background sequence. For instance, we identified MbgIX in seven of nine analysed positive dE droplets which were selected through *E. coli* mIVC-FACS. While it can be argued that the distribution of this gene within droplets is dependent on its abundance within the metagenome sample, we have also

reported that the gene could not be identified within the 829Mb shotgun data derived from the same metagenome. In addition, it was reported in Chapter five that five different beta-xylosidase encoding genes were isolated and all of them could not be traced back to the shotgun data. Even more convincing was that the beta-xylosidases identified were completely different for the two different cell-free extracts used. Moreover, the majority of dE droplets contained the same gene. Based on the fact that they were selected from the same gate which has relatively the same fluorescence intensity, it further strengthens the arguments that that these genes were selected based on the expression and/or activity observed. Although a small number of events were analysed, these results suggest that the genes analysed were not abundant in the original sample and weremost likely selected based on the expression and activity observed.

One of the objectives of this study was to compare the mIVC-FACSScreening platform with traditional function-based screening, and also to compare *E. coli* based mIVC-FACS with *R. erythropolis* based mIVC-FACS system. Comparison based on hit ratesobtained from traditional *E. coli* based fosmid library screening shows that mIVC-FACS is more sensitive (Figure 55). The *E. coli* based mIVC-FACS hit rate was over 2.5 times higher than the clone library system. Although only a few hits from mIVC-FACS screening were confirmed through heterologous expression, this result suggests that a significant number of positive hits are left un-accessed through clone library screening.

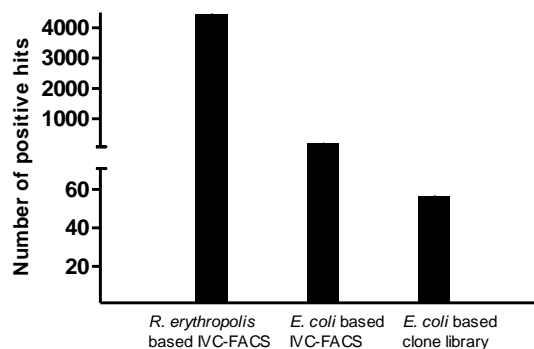


FIGURE 55: Comparison of hit rate. The hit rate obtained from mIVC-FACS systems were compared to traditional clone library screening method. The amount of DNA screened was normalised to 1Gb.

In addition, a very stringent gate was used to minimise the level of false positives, thus it is expected that many xylosidase hits obtained, albeit with lower activities, have yet to be analysed. These results also suggest that the *E. coli* expression system might be intrinsically sub-optimal for screening for hemicellulases from thermophilic environmental genomes compared to the *R. erythropolis* system that shows a significantly higher hit rate (30 fold higher). The simplicity of the *E. coli* system and the vast body of knowledge of its biochemical machinery makes it a good organism for application in functional metagenomic screening. However, once again, this study corroborates many others which have shown the limitations associated with using *E. coli* for metagenomic screening (Fakruddin *et al.* 2012; Schlegel *et al.* 2013)

This study also demonstrates that mIVC-FACS is more selective than traditional *E. coli* based library screening. Through the *R. erythropolis* system, five different genes were identified from the same gate. The fact that the majority of the analysed events contain the same genes is a strong indication of the ability of mIVC-FACS to separate gene products based on their effectiveness in hydrolysing substrates under the given conditions. This characteristic was also observed when the *E. coli* system was used where only *mbglX* was recovered from all analysed events derived from the same gate. This is one of the advantages of mIVC-FACS given the high hit rate nature of this screening system. Since the positive hits can be grouped in distinct gates, which is in part a proxy to their biochemical characteristics, it is easier to isolate only hits which show greatest or desired activity for further analysis. However, as with conventional activity screening, the expression conditions also dictate the nature of enzymes to be identified.

In addition to selection of enzymes based on their activity, mIVC-FACS also offers a platform to improve these enzymes through directed evolution. Metagenomic bio-prospecting is the first step in identifying an enzyme for a commercial process. Commercial enzymes are required to have high activity and stability under process conditions, desired substrate

specificity, and high selectivity. More often than not, naturally occurring enzymes do not fulfil the requirements of these harsh industrial conditions, and optimisation is necessary to obtain a suitable enzyme catalyst for production needs (Bundy & Swartz 2010). The high hit rate obtained through mIVC-FACS system can serve to generate a library of enzymes which can further be manipulated for desired characteristics through direct mutagenesis. Figure 56 demonstrate a proposed model for metagenomic mIVC-FACS screening and direct enzyme evolution.

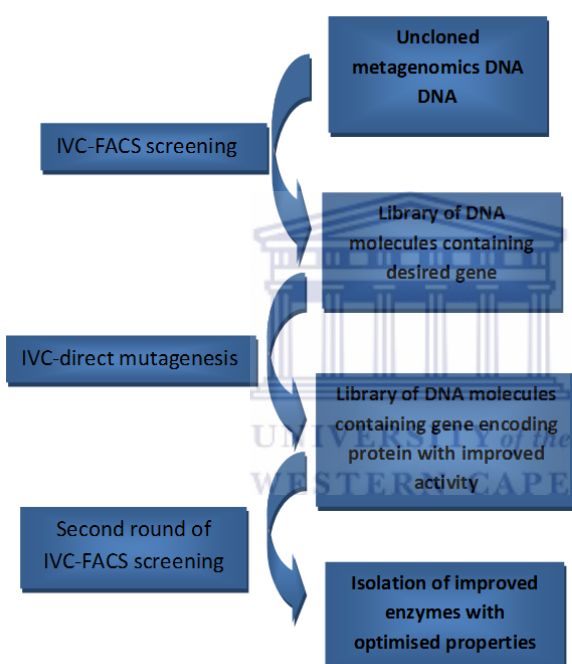
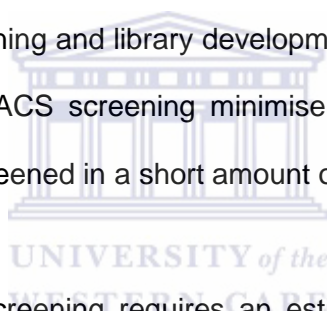


FIGURE 56: Flow chart illustrating the identification of highly active novel genes through metagenomic mIVC-FACS and directed evolution approaches. Newly identified enzymes from IVC-FACS screening can serve as an ideal starting point for the directed evolution of enzymes with improved properties. IVC-FACS hits can be subjected to direct enzyme evolution through direct mutagenesis, resulting in isolation of enzyme with improved properties.

One of the major challenges of metagenomic bio-prospecting is the continuous re-discovery of already known activities. Microbial communities from approximately 2192 different sites across the planet have been examined for their metagenomic content (Eugster & Schlesinger 2013). They include habitats such as terrestrial, marine and freshwater, non-

marine saline and alkaline lakes, acid mine drainage systems, wastewater treatment sludges, compost and marine sponge, termite and earthworms gut, rumen, human microbiota, etc. Within all these sites and environments, approximately 6100 clones containing new enzymes have been described to date (Ferrer *et al.* 2016), while significantly high number of positive clones contain already described genes. Singh & Macdonald (2010) reported that insufficient screening methods for rare enzymatic activities are one of the challenges facing the discovery of industrially applicable enzymes. In this study, we have demonstrated that mIVC-FACS is able to isolate genes in low abundance from mixed microbial consortia. For instance, genes isolated here were not identified in the shotgun data of this metagenome. The identification of these genes through mIVC-FACS demonstrates the astonishingly ultra-high sensitivity of this platform. While these genes could have been lost through cloning and library development, the elimination of these steps coupled with high throughput FACS screening minimises DNA loss while increasing the amount of DNA which can be screened in a short amount of time.



The workflow of clone library screening requires an estimated 320 hours to screen one million fosmid clones compared to 112 hours required for mIVC-FACS screening (Figure 57). The most time consuming process in fosmid library screening is colony picking and re-inoculation in screening media. Even with robotic systems, this process can still take at least 168 hours to pick one million colonies. On the contrary, FACS screening, DNA recovery and amplification can be done in at least 9 hours. When the commercial CFPS system is used instead of an in house CFPS system, the time required for mIVC-FACS screening can be reduced to at least 88 hours. In addition, the number of events to be screened does not have a major impact on time since preparation and screening of up to 10^{10} events can be done in a single reaction. Although sub-screening in this study was done *in vivo*, this cell-free system can be used for confirmation of isolated DNA molecules in much less time than required for PCR and cloning process. The MDA amplicons derived from a positive dE droplet can be used directly as template in an *in vitro* CFPS system followed by activity assay. Furthermore,

the screening with multiple CFPS extracts can be conducted consecutively and without the need of years' worth of shuttle vector development.

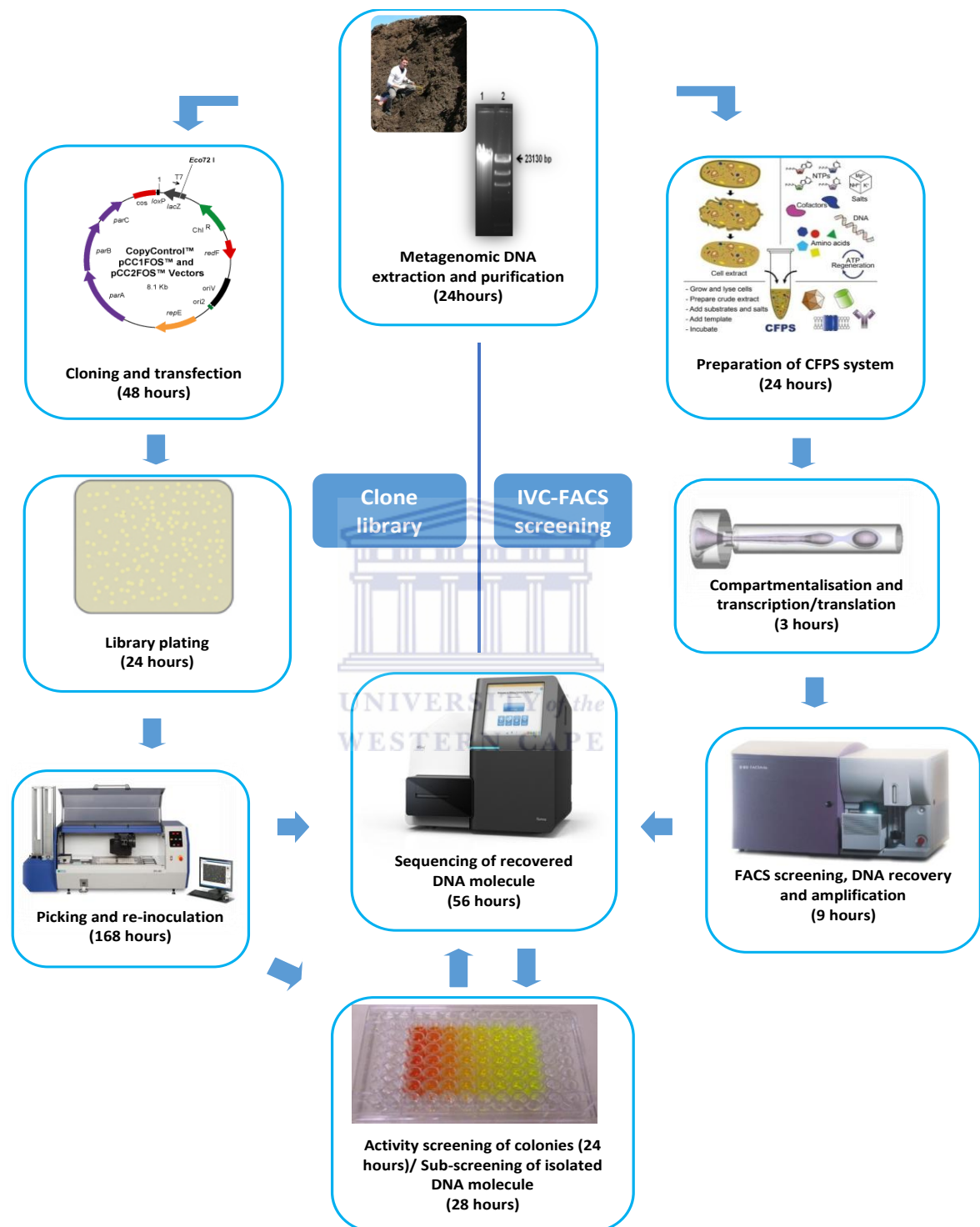


FIGURE 57: Comparison of mIVC-FACS and clone library screening systems. The screening was estimated for one million clones in the case of clone library screening or one million events for IVC-FACS.

Lignocellulases are required for a wide range of applications, including the production of biofuels. While the production of biofuels from the renewable lignocellulosic biomass is gradually considered a promising way to replace fossil fuels, its bioconversion has been limited by the need for effective hydrolysis of lignin-carbohydrate complexes (LCC). This represents a major challenge in global efforts to utilize renewable resources in place of fossil fuels to meet the rising energy demands. Enzymatic hydrolysis is the most common process to degrade the cellulose and hemicellulose into fermentable sugars such as glucose and xylose, and multiple substrate enzymes are the most promising in this regards. This study has identified 38 beta-xylosidases encoding ORFs, ten of which have been confirmed through activity assay. None of the sequences identified in the shotgun sequence data were identified in the genes that were confirmed through function-based assay. Some of these enzymes show potential industrial application through their ability of hydrolyse multiple substrates. Figure 58 summarises all enzymes which have been reported in this study.

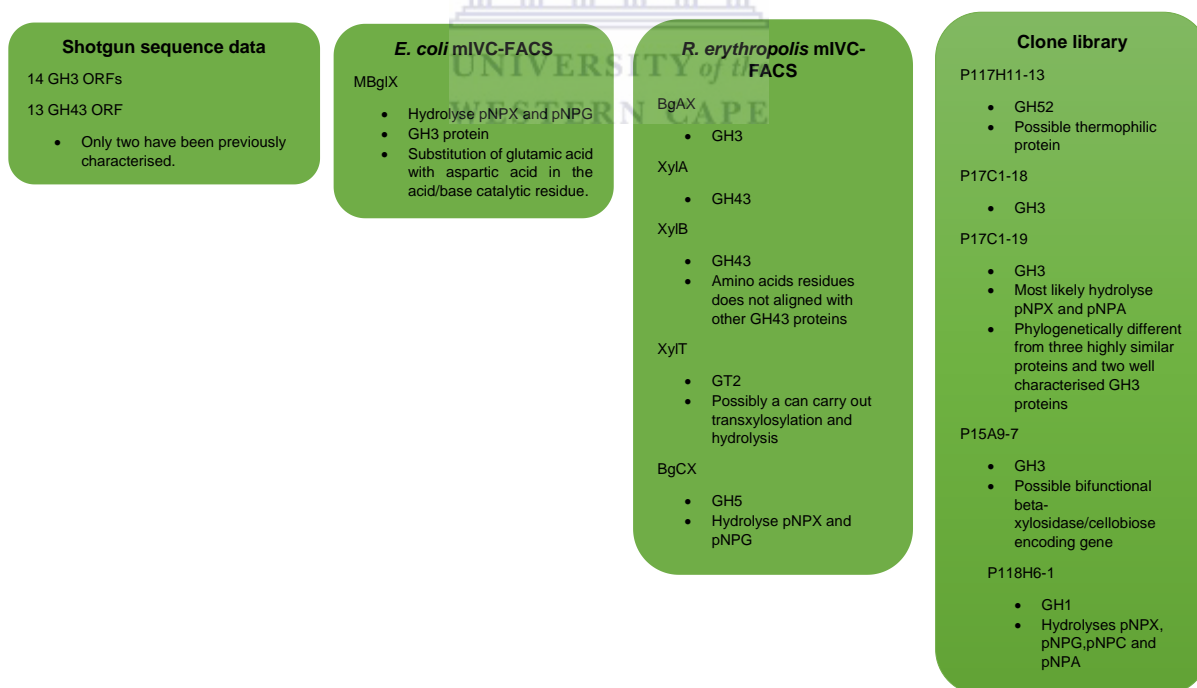


FIGURE 58: Beta-xylosidase encoding which were identified in this study.

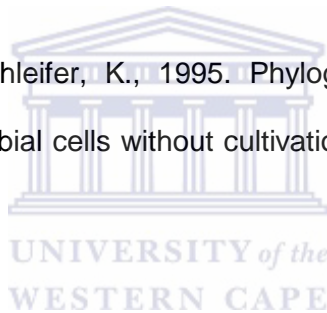
CONCLUSION

We have reported in this work a novel metagenomic approach to identify beta-xylosidase encoding genes within a complex, horse manure microbial metagenome. To increase the probability of identifying rare genes, we adopted a strategy based on cell-free protein synthesis, *in vitro* compartmentalisation and FACS sorting of uncloned DNA. While this platform still requires optimisation, we have demonstrated that this technique can be used to isolate genes encoding enzymes from mixed microbial genomes. mIVC-FACS is a promising technology with the potential to take metagenomic studies to the second generation of novel natural products bio-prospecting. The astonishing sensitivity and ultra-high throughput of this technology offers numerous advantages in metagenomic bio-prospecting.



REFERENCES

- Agresti, J.J. *et al.*, 2005. Selection of ribozymes that catalyse multiple-turnover Diels-Alder cycloadditions by using *in vitro* compartmentalisation. *Proceedings of the National Academy of Sciences of the United States of America*, 102(45), pp.16170–5.
- Åkerman, B. & Cole, K.D., 2002. Electrophoretic capture of circular DNA in gels. *Electrophoresis*, 23(16), pp.2549–2561.
- Allgaier, M. *et al.*, 2010. Targeted discovery of glycoside hydrolases from a switchgrass-adapted compost community. *PLoS ONE*, 5(1), pp.1-15
- Amann, R.L., Ludwig, W. & Schleifer, K., 1995. Phylogenetic identification and *In situ* detection of individual microbial cells without cultivation. *Microbiological reviews*, 59(1), pp.143–169.
- Andersen, D.C. & Krummen, L., 2002. Recombinant protein expression for therapeutic applications. *Current opinion in biotechnology*, 13(2), pp.117–23.
- Artimo, P. *et al.*, 2012. ExPASy: SIB bioinformatics resource portal. *Nucleic Acids Research*, 40(W1), pp.597–603.
- Avenaud, P. *et al.*, 2000. Expression and activity of the cytolethal distending toxin of *Helicobacter hepaticus*. *Infection and Immunity*, 68(3), pp.184–191.
- Aziz, R.K. *et al.*, 2015. Multidimensional metrics for estimating phage abundance, distribution, gene density, and sequence coverage in metagenomes. *Frontiers in Microbiology*, 6(381), pp.1-18



- Bailey, T.L. *et al.*, 2009. MEME Suite: Tools for motif discovery and searching. *Nucleic Acids Research*, 37(2), pp.202–208.
- Baneyx, F., 1999. Recombinant protein expression in *Escherichia coli*. *Current opinion in Chemical Engineering*, 10 (5), pp.411–421.
- Barker, I.J., Petersen, L. & Reilly, P.J., 2010. Mechanism of xylobiose hydrolysis by GH43 - xylosidase. *The journal of physical chemistry*, 114 (46), pp.15389–15393.
- Barone, R. *et al.*, 2014. Marine metagenomics, a valuable tool for enzymes and bioactive compounds discovery. *Frontiers in Marine Science*, 1(38), pp.1-11
- Bartossek, R. *et al.*, 2010. Homologues of nitrite reductases in ammonia-oxidizing archaea: diversity and genomic context. *Environmental Microbiology*, 12(4), pp.1075–1088.
- Bastien, G. *et al.*, 2013. Mining for hemicellulases in the fungus-growing termite *Pseudacanthotermes militaris* using functional metagenomics. *Biotechnology for biofuels*, 6(78), pp.1-15
- Bause, E. & Legler, G., 1974. Isolation and amino acid sequence of a hexadecapeptide from the active site of beta-glucosidase A3 from *Aspergillus wentii*. *Physiological chemistry*, 355(4), pp438-442.
- Bell, P.J.L. *et al.*, 2002. Prospecting for novel lipase genes using PCR. *Microbiology*, 148(8), pp.2283–2291.
- Beloqui, A., Domí, P. & Ferrer, M., 2008. Recent trends in industrial microbiology. *Current opinion in Microbiology*, 11(3), pp.240–248.

Bernath, K. *et al.*, 2004. *In vitro* compartmentalisation by double emulsions: sorting and gene enrichment by Fluorescence Activated Cell Sorting. *Analytical Biochemistry*, 325(1), pp.151–157.

Bernath, K., Magdassi, S. & Tawfik, D., 2009. *In vitro* compartmentalisation (IVC): A High-throughput screening technology using emulsion and FACS. *Discovery Medicine*, 4(20), pp.49-53.

Binga, E.K., Lasken, R.S. & Neufeld, J.D., 2008. Something from (almost) nothing: the impact of multiple displacement amplification on microbial ecology. *The ISME Journal*, 2(3), pp.233–241.

Blackwell, J.R. & Horgan, R., 1991. A novel strategy for production of a highly expressed recombinant protein in an active form. *FEBS Letters*, 295(1-3), pp.10–12.

Blake, R.D. & Delcourt, S.G., 1996. Thermodynamic effects of formamide on DNA stability. *Nucleic Acids Research*, 24(11), pp.2095–2103.

Bourdichon, F. *et al.*, 2012. Food fermentations: microorganisms with technological beneficial use. *International journal of food microbiology*, 154(3), pp.87–97.

Bourne, Y. & Henrissat, B., 2001. Glycoside hydrolases and glycosyltransferases: Families and functional modules. *Current Opinion in Structural Biology*, 11(5), pp.593–600.

Brüx, C. *et al.*, 2006. The Structure of an Inverting GH43 β -Xylosidase from *Geobacillus stearothermophilus* with its substrate reveals the role of the three catalytic residues. *Journal of Molecular Biology*, 359(1), pp.97–109.

- Bundy, B.C. & Swartz, J.R., 2011. Efficient disulfide bond formation in virus-like particles. *Journal of Biotechnology*, 154(4), pp.230–239.
- Bundy, B.C. & Swartz, J.R., 2010. Site-specific incorporation of p-propargyloxyphenylalanine in a cell-free environment for direct protein-protein click conjugation. *Bioconjugate Chemistry*, 21(2), pp.255–263.
- Carlson, E.D. *et al.*, 2011. Cell-free protein synthesis: applications come of age. *Biotechnology advances*, 30(5), pp.1185–1194.
- Charnock, S.J. & Davies, G.J., 1999. Structure of the nucleotide-diphospho-sugar transferase , SpsA from *Bacillus subtilis* in native and nucleotide-complexed forms. *Biochemistry*, 38(20), pp.6380–6385.
- Chauhan, P.S. *et al.*, 2012. Mannanases: microbial sources, production, properties and potential biotechnological applications. *Applied microbiology and biotechnology*, 93(5), pp.1817–1830.
- Chávez-Páez, M. *et al.*, 2012. Coalescence in double emulsions. *Langmuir*, 28(14), pp.5934–5939.
- Chen, G.T. & Inouye, M., 1994. Role of the AGA/AGG codons, the rarest codons in global gene expression in *Escherichia coli*. *Genes & Development*, 8(21), pp.2641–2652.
- Chir, J. *et al.*, 2002. Identification of the two essential groups in the family 3 beta-glucosidase from *Flavobacterium meningosepticum* by labelling and tandem mass spectrometric analysis. *Biochemistry*, 165(pt3), pp.857–863.

- Chor, B. *et al.*, 2009. Genomic DNA k-mer spectra: models and modalities. *Genome biology*, 10(10), R108.
- Christel, M. *et al.*, 2011. Characterization of a new β -glucosidase/ β -xylosidase from the gut microbiota of the termite (*Reticulitermes santonensis*). *FEMS Microbiology Letters*, 314(2), pp.147–157.
- Clarridge, J.E. & Zhang, Q., 2002. Genotypic diversity of clinical Actinomyces species: phenotype, source, and disease correlation among genospecies. *Journal of clinical microbiology*, 40(9), pp.3442–3448.
- Colin, P.-Y. *et al.*, 2015. Ultrahigh-throughput discovery of promiscuous enzymes by picodroplet functional metagenomics. *Nature communications*, 6(10008), p.100-118.
- Courtois, S. *et al.*, 2001. Quantification of bacterial subgroups in soil: comparison of DNA extracted directly from soil or from cells previously released by density gradient centrifugation. *Environmental Microbiology*, 3(7), pp.431–439.
- Coutinho, P.M. *et al.*, 2003. An Evolving Hierarchical Family Classification for Glycosyltransferases. *Journal of Molecular Biology*, 328(2), pp.307–317.
- Craig, J.W. *et al.*, 2010. Expanding small-molecule functional metagenomics through parallel screening of broad-host-range cosmid environmental DNA libraries in diverse proteobacteria. *Applied and Environmental Microbiology*, 76(5), pp.1633–1641.
- Currie, D.H. *et al.*, 2014. Profile of Secreted hydrolases associated proteins, and SlpA in *Thermoanaerobacterium saccharolyticum* during the degradation of hemicellulose. *Applied and Environmental Microbiology*, 80(16), pp.5001–5011.

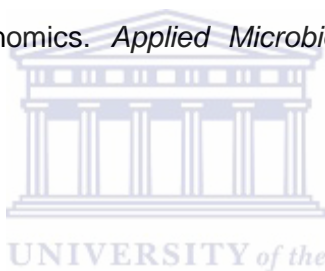
- Dabek, M. *et al.*, 2008. Distribution of beta-glucosidase and beta-glucuronidase activity and of beta-glucuronidase gene *gus* in human colonic bacteria. *FEMS microbiology ecology*, 66(3), pp.487–495.
- Daniel, R., 2005. The metagenomics of soil. *Nature Reviews Microbiology*, 3(6), pp.470–478.
- Danon, M. *et al.*, 2008. Molecular analysis of bacterial community succession during prolonged compost curing. *FEMS Microbiology Ecology*, 65(1), pp.133–144.
- Dedhia, N. *et al.*, 2000. Improvement in recombinant protein production in ppGpp-deficient *Escherichia coli*. *Biotechnology and Bioengineering*, 53(4), pp.379–386.
- Delmont, T.O. *et al.*, 2011. Metagenomic comparison of direct and indirect soil DNA extraction approaches. *Journal of Microbiological Methods*, 86(3), pp.397–400.
- Devasahayam, M., 2007. Factors affecting the expression of recombinant glycoproteins. *The Indian journal of medical research*, 126(1), pp.22–27.
- Diaz-Torres, M.L. *et al.*, 2006. Determining the antibiotic resistance potential of the indigenous oral microbiota of humans using a metagenomic approach. *FEMS microbiology letters*, 258(2), pp.257–262.
- Dodd, D. *et al.*, 2010. Functional diversity of four glycoside hydrolase family 3 enzymes from the rumen bacterium *Prevotella bryantii* B14. *Journal of Bacteriology*, 192(9), pp.2335–2345.
- Dohm, J.C. *et al.*, 2008. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Research*, 36(16), pp.1-10.

Doi, N. *et al.*, 2004. *In vitro* selection of restriction endonucleases by *in vitro* compartmentalisation. *Nucleic acids research*, 32(12), pp.1-8.

Dougherty, M.J. *et al.*, 2012. Glycoside hydrolases from a targeted compost metagenome , activity-screening and functional characterization. *BMC Biotechnology*, 12(38), pp1-9.

Eckburg, P.B. *et al.*, 2005. Diversity of the human intestinal microbial flora. *Science*, 308(5728), pp.1635–1638.

Ekkers, D.M. *et al.*, 2012. The great screen anomaly- A new frontier in product discovery through functional metagenomics. *Applied Microbiology and Biotechnology*, 93(3), pp.1005–1020.



Eugster, M. & Schlesinger, T., 2013. Osmar: OpenScreenMap and R. *The R Journal*, pp.53-63.

Fahnert, B., Lilie, H. & Neubauer, P., 2004. Inclusion bodies: formation and utilisation. *Advances in biochemical engineering*, 89, pp.93–142.

Fangman, W.L., 1978. Separation of very large DNA molecules by gel electrophoresis. *Nucleic acids research*, 5(3), pp.653–665.

Felczykowska, A., Bloch, S.K. & Nejman-faleńczyk, B., 2012. Metagenomic approach in the investigation of new bioactive compounds in the marine environment. *Acta biochimica polonica*, 59(4), pp.501–505.

- Ferrer, M. *et al.*, 2016. Estimating the success of enzyme bio-prospecting through metagenomics: Current status and future trends. *Microbial Biotechnology*, 9(1), pp.22–34.
- Ferrer, M. *et al.*, 2012. Functional metagenomics unveils a multifunctional glycosyl hydrolase from the family 43 catalysing the breakdown of plant polymers in the calf rumen. *PLoS one*, 7(6), p.e38134.
- Ferrer, M. *et al.*, 2009. Interplay of metagenomics and *in vitro* compartmentalisation. *Microbial biotechnology*, 2(1), pp.31–39.
- Ficheux, M., Bonakdar, L. & Bibette, J., 1998. Some stability criteria for double emulsions. *Langmuir*, 14(11), pp.2702–2706.
- Florence, A.. & Whitehill, D., 1981. Transfer phenomena across the oil phase in water-oil-water multiple emulsions evaluated by coulter counter. *Journal of colloid interface*, 101(2) , pp.587-591.
- Gabor, E., Liebeton, K. & Lorenz, P., 2007. Updating the metagenomics toolbox. *Biotechnology Journal*, 2(2), pp.201–206.
- Gabor, E.M., Alkema, W.B.L. & Janssen, D.B., 2004. Quantifying the accessibility of the metagenome by random expression cloning techniques. *Environmental Microbiology*, 6(9), pp.879–886.
- Gadd, G.M., 2010. Metals, minerals and microbes: geomicrobiology and bioremediation. *Microbiology (Reading, England)*, 156(Pt 3), pp.609–643.
- Garner, M. & Chrambach, A., 1992. Resolution of circular, nicked circular and linear DNA, 4.4 kb in length, by electrophoresis in polyacrylamide solutions. *Electrophoresis*, 13(3),

pp.176–178.

Gershenson, A. & Gierasch, L.M., 2011. Protein folding in the cell: challenges and progress.

Current opinion in structural biology, 21(1), pp.32–41.

Ghadessy, F.J. & Holliger, P., 2004. A novel emulsion mixture for *in vitro* compartmentalisation of transcription and translation in the rabbit reticulocyte system.

Protein Engineering, Design and Selection, 17(3), pp.201–204.

Ghrayeb, J., Kimura, H. & Takahara, M., 1984. Secretion cloning vectors in *Escherichia coli*.

, 3(10), pp.2437–2442.

Glavina Del Rio, T. *et al.*, 2010. Complete genome sequence of *Chitinophaga pinensis* type strain (UQM 2034). *Standards in genomic sciences*, 2(1), pp.87–95.



Glick, B.R. & Whitney, K., 1987. Factors affecting the expression of foreign proteins in

Escherichia coli. , pp.277–282.

Gomez del Pulgar, E. & Saadeddin, A., 2014. The cellulolytic system of *Thermobifida fusca*.

Microbiology, 40(3), pp.236–247.

Gonzalez, J.M. *et al.*, 2012. Amplification by PCR artificially reduces the proportion of the rare biosphere in microbial communities. *PlosOne*, 7(1), e29973.

Goujon, T. *et al.*, 2003. AtBXL1, a novel higher plant (*Arabidopsis thaliana*) putative beta-xylosidase gene, is involved in secondary cell wall metabolism and plant development.

Plant Journal, 33(4), pp.677–690.

Grabnitz, F. *et al.*, 1991. Structure of the beta-glucosidase gene *bgIA* of *Clostridium thermocellum*. *Reactions*, 200(2), pp.301–309.

Granström, T.B., Izumori, K. & Leisola, M., 2007. A rare sugar xylitol. Part II: biotechnological production and future applications of xylitol. *Applied Microbiology and Biotechnology*, 74(2), pp.273–276.

Green, S.J. *et al.*, 2004. Similarity of bacterial communities in sawdust- and straw-amended cow manure composts. *FEMS Microbiology Letters*, 233(1), pp.115–123.

Griffiths, A.D. & Tawfik, D.S., 2003. Directed evolution of an extremely fast phosphotriesterase by *in vitro* compartmentalisation. *The EMBO journal*, 22(1), pp.24–35.

Hall, N., 2007. Advanced sequencing technologies and their wider impact in microbiology. *The Journal of experimental biology*, 210(Pt 9), pp.1518–25.

Handelsman, J., 2004. Metagenomics: Application of Genomics to Uncultured Microorganisms. *Microbiology and Molecular Biology Reviews*, 68(4), pp.669–685.

Handelsman, J. *et al.*, 1998. Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. *Chemistry & biology*, 5(10), pp.245–249.

Hartl, F.U. *et al.*, 2011. Molecular chaperones in protein folding and proteostasis. *Nature review*, 475(7337), pp324-332.

Heath, C. *et al.*, 2009. Identification of a novel alkaliphilic esterase active at low temperatures by screening a metagenomic library from antarctic desert soil. *Applied*

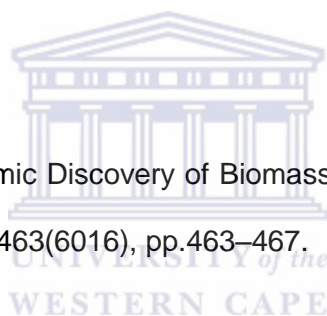
and Environmental Microbiology, 75(13), pp.4657–4659.

Henrissat, B. *et al.*, 1995. Conserved catalytic machinery and the prediction of a common fold for several families of glycosyl hydrolases. *Proceedings of the National Academy of Sciences of the United States of America*, 92(15), pp.7090–7094.

Henrissat, B., Vegetales, M. & Grenoble, F.-, 1991. A classification of glycosyl hydrolases based sequence similarities amino acid. *Biochemical journal*, 280(pt2), pp.309–316.

Herzenberg, L. a. *et al.*, 2002. The history and future of the Fluorescence Activated Cell Sorter and flow cytometry: A view from Stanford. *Clinical Chemistry*, 48(10), pp.1819–1827.

Hess, M. *et al.*, 2011. Metagenomic Discovery of Biomass-Degrading Genes and Genomes from Cow Rumen. *Science*, 463(6016), pp.463–467.



Hindson, B.J. *et al.*, 2011. High-throughput droplet digital PCR system for absolute quantitation of DNA copy number. *Analytical Chemistry*, 83(22), pp.8604–8610.

Hiraga, S.I., Sugiyama, T. & Itoh, T., 1994. Comparative analysis of the replicon regions of eleven ColE2-related plasmids. *Journal of Bacteriology*, 176(23), pp.7233–7243.

Hjort, K. *et al.*, 2010. Chitinase genes revealed and compared in bacterial isolates, DNA extracts and a metagenomic library from a phytopathogen-suppressive soil. *FEMS Microbiology Ecology*, 71(2), pp.197–207.

Hoffman, S., 1985. Fusions of secreted proteins to alkaline phosphatase: An approach for studying protein secretion. *PNAS*, 82(15), pp.5107–5111.

- Hosaka, T. *et al.*, 2009. Antibacterial discovery in actinomycetes strains with mutations in RNA polymerase or ribosomal protein S12. *Nature Biotechnology*, 27(5), pp.462–464.
- Howard, R.L. *et al.*, 2003. Lignocellulose biotechnology: issues of bioconversion and enzyme production. *African journal of biotechnology*, 2(12), pp.602–619.
- Hrmova, M. *et al.*, 1998. Substrate binding and catalytic mechanism of a barley β -D-glucosidase/(1,4)- β -D-glucan exohydrolase. *Journal of Biological Chemistry*, 273(18), pp.11134–11143.
- Hu, J., Arantes, V. & Saddler, J.N., 2011. The enhancement of enzymatic hydrolysis of lignocellulosic substrates by the addition of accessory enzymes such as xylanase: is it an additive or synergistic effect? *Biotechnology for biofuels*, 4(1), pp.36-42.
- Huang, Z. *et al.*, 2014. GH52 xylosidase from *Geobacillus stearothermophilus*: Characterization and introduction of xylanase activity by site-directed mutagenesis of Tyr509. *Journal of Industrial Microbiology and Biotechnology*, 41(1), pp.65–74.
- Hudson, R. *et al.*, 1993. Purification and properties of a β -1,4-xylanase from a cellulolytic extreme thermophile expressed in *Escherichia coli*. *International Journal of Biochemistry*, 25(4), pp.609–617.
- Ilmberger, N. *et al.*, 2014. A comparative metagenome survey of the fecal microbiota of a breast and a plant-fed asian elephant reveals an unexpectedly high diversity of glycoside hydrolase family enzymes. *PLoS ONE*, 9(9), pp.101-107.
- Isaksson, L. *et al.*, 2012. Expression screening of membrane proteins with cell-free protein synthesis. *Protein expression and purification*, 82(1), pp.218–25.

Jami, E. & Mizrahi, I., 2012. Composition and similarity of bovine rumen microbiota across individual animals. *PLoS ONE*, 7(3), pp.1–8.

Jogler, C. *et al.*, 2009. Toward cloning of the magnetotactic metagenome: Identification of magnetosome island gene clusters in uncultivated magnetotactic bacteria from different aquatic sediments. *Applied and Environmental Microbiology*, 75(12), pp.3972–3979.

Jones, G.W., Doyle, S. & Fitzpatrick, D.A., 2014. The evolutionary history of the genes involved in the biosynthesis of the antioxidant ergothioneine. *Gene*, 549(1), pp.161–170.

Jordan, D.B. & Wagschal, K., 2010. Properties and applications of microbial beta-D-xylosidases featuring the catalytically efficient enzyme from *Selenomonas ruminantium*. *Applied microbiology and biotechnology*, 86(6), pp.1647–1658.

Kaeberlein, T., Lewis, K. & Epstein, S.S., 2002. Isolating “Uncultivable” Microorganisms in pure culture in a simulated natural environment. *Science*, 296(5570) pp.1127–1130.

Kang, C.-H. *et al.*, 2011. A novel family VII esterase with industrial potential from compost metagenomic library. *Microbial Cell Factories*, 10(1), pp.41-48.

Kanokratana, P. *et al.*, 2011. Insights into the phylogeny and metabolic potential of a primary tropical peat swamp forest microbial community by metagenomic analysis. *Microbial Ecology*, 61(3), pp.518–528.

Keegan, K.P. *et al.*, 2012. A platform-independent method for detecting errors in metagenomic sequencing data: DRISSE. *PLoS Computational Biology*, 8(6), pp.1-11

- Keenleyside, W.J., Clarke, A.J. & Whitfield, C., 2001. Identification of residues involved in catalytic activity of the inverting glycosyl transferase WbbE from *Salmonella enterica* Serovar Borreze. *Journal of Bacteriology*, 183(1), pp.77–85.
- Kennedy, J., Marchesi, J.R. & Dobson, A.D.W., 2007. Metagenomic approaches to exploit the biotechnological potential of the microbial consortia of marine sponges. *Applied microbiology and biotechnology*, 75(1), pp.11–20.
- Kigawa, T. *et al.*, 2004. Preparation of *Escherichia coli* cell extract for highly productive cell-free protein expression. *Journal of structural and functional genomics*, 5(1-2), pp.63–8.
- Kim, D.M. *et al.*, 1996. A highly efficient cell-free protein synthesis system from *Escherichia coli*. *European journal of biochemistry / FEBS*, 239(3), pp.881–886.
- Kim, H.-C. & Kim, D.-M., 2009. Methods for energizing cell-free protein synthesis. *Journal of bioscience and bioengineering*, 108(1), pp.1–4.
- Kim, T.-W. *et al.*, 2006. Simple procedures for the construction of a robust and cost-effective cell-free protein synthesis system. *Journal of biotechnology*, 126(4), pp.554–561.
- Kim UJ, Shizuya H, Dejong PJ, Birren B, S.M., 1992. Stable propagation of cosmid sized human DNA inserts in an F-factor based vector. *Nucleic Acids Res*, 20(2), pp.1883-1885.
- Kimura, I., Yoshioka, N. & Tajima, S., 1999. Purification and characterization of a beta-glucosidase with beta-xylosidase activity from *Aspergillus sojae*. *Journal of bioscience and bioengineering*, 87(4), pp.538–541.

- Kojima, T. *et al.*, 2005. PCR amplification from single DNA molecules on magnetic beads in emulsion: Application for high-throughput screening of transcription factor targets. *Nucleic Acids Research*, 33(17), pp.1–9.
- Komar, A., Lesnik, T. & Reiss, C., 1999. Synonymous codon substitution affects ribosome traffic and protein folding during *in vitro* translation. *FEBS letters*, 462(3), pp.387–391.
- Kousar, S., Mustafa, G. & Jamil, A., 2013. Microbial Xylosidases: Production and biochemical characterization. *Pakistan Journal of Life and Social Sciences*, 11(2), pp.85–95.
- Kurakake, M. *et al.*, 2005. Characteristics of transxylosylation by β -xylosidase from *Aspergillus awamori* K4. *Japan Biochimica et Biophysica Acta*, 1726(3), pp.272–279.
- Lama, L. *et al.*, 2004. Purification and characterization of thermostable xylanase and beta-xylosidase by the thermophilic bacterium *Bacillus thermantarcticus*. *Res Microbiol*, 155(4), pp.283–289.
- Lange, J. & Solutions, S.G., 2007. Lignocellulose conversion: an introduction to chemistry ., pp.39–48.
- Larsson, a., 2014. AliView: a fast and lightweight alignment viewer and editor for large data sets. *Bioinformatics*, 30(22), pp.531–542
- Leis, B., Angelov, A. & Liebl, W., 2013. Screening and expression of genes from metagenomes. *Advances in applied microbiology*, 83(13), pp.1-68 .
- Lemmel, S., Datta, R. & Frankiewicz, J., 1986. Fermentation of xylan by *Clostridium acetobutylicum*. *Enzyme and Microbial Technology*, 8(4), pp.217–221.

- Li, M. *et al.*, 2010. A comparison of primer sets for detecting 16S rRNA and hydrazine oxidoreductase genes of anaerobic ammonium-oxidizing bacteria in marine sediments. *Applied Microbiology and Biotechnology*, 86(2), pp.781–790.
- Liles, M.R. *et al.*, 2008. Recovery, purification, and cloning of high-molecular-weight DNA from soil microorganisms. *Applied and environmental microbiology*, 74(10), pp.3302–3305.
- Liu, D.V., Zawada, J.F. & Swartz, J.R., 2005. Streamlining *Escherichia coli* S30 extract preparation for economical cell-free protein synthesis. *Biotechnology progress*, 21(2), pp.460–465.
- Liu, L. *et al.*, 2012. Comparison of next-generation sequencing systems. *Journal of biomedicine & biotechnology*, 12 (251), pp.1-11.
- Lykidis, A. *et al.*, 2007. Genome sequence and analysis of the soil cellulolytic actinomycete *Thermobifida fusca* YX. , 189(6), pp.2477–2486.
- Margulies, M. *et al.*, 2005. Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, 437(7057), pp.376–80.
- Maroti, G. *et al.*, 2009. Discovery of [NiFe] hydrogenase genes in metagenomic DNA: Cloning and heterologous expression in *Thiocapsa roseopersicina*. *Applied and Environmental Microbiology*, 75(18), pp.5821–5830.
- Martinez, A. *et al.*, 2004. Genetically modified bacterial strains and novel bacterial artificial chromosome shuttle vectors for constructing environmental libraries and detecting heterologous natural products in multiple expression hosts. *Applied and environmental microbiology*, 70(4), pp.2452–2463.

- Mastrobattista, E. *et al.*, 2005. High-throughput screening of enzyme libraries: *in vitro* evolution of a beta-galactosidase by fluorescence-activated sorting of double emulsions. *Chemistry & biology*, 12(12), pp.1291–300.
- Matsuzawa, T., Kaneko, S. & Yaoi, K., 2015. Screening, identification, and characterization of a GH43 family beta-xylosidase/alpha-arabinofuranosidase from a compost microbial metagenome. *Applied Microbiology and Biotechnology*, 99(21), pp.8943–8954.
- Meyer, F., Paarmann, D., D'Souza, M., Olson, R., *et al.*, 2008. The metagenomics RAST server – a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics*, 9(1), pp.386.
- Mhuantong, W. *et al.*, 2015. Comparative analysis of sugarcane bagasse metagenome reveals unique and conserved biomass-degrading enzymes among lignocellulolytic microbial communities. *Biotechnology for Biofuels*, 8(1), pp.1-17
- Miller, O.J. *et al.*, 2006. Directed evolution by *in vitro* compartmentalisation. *Nature methods*, 3(7), pp.561–570.
- Moore, L.D., Kocun, F.J. & Umbreit, W.W., 1966. Cell-free protein synthesis: effects of age and state of ribosomal aggregation. *Science*, 154(754), pp.1350–1353.
- Morimoto, S. & Fujii, T., 2009. A new approach to retrieve full lengths of functional genes from soil by PCR-DGGE and metagenome walking. *Applied Microbiology and Biotechnology*, 83(2), pp.389–396.
- Morozova, O. & Marra, M., 2008. Applications of next-generation sequencing technologies in functional genomics. *Genomics*, 92(5), pp.255–264.

Myronovskyi, M. *et al.*, 2011. Beta-glucuronidase as a sensitive and versatile reporter in actinomycetes. *Applied and environmental microbiology*, 77(15), pp.5370–83.

Nair, H.P., Vincent, H. & Bhat, S.G., 2014. Evaluation of five in situ lysis protocols for PCR amenable metagenomic DNA from mangrove soils. *Biotechnology Reports*, 4(14), pp.134–138.

Nakashima, N. & Tamura, T., 2004. A novel system for expressing recombinant proteins over a wide temperature range from 4 to 35 degrees C. *Biotechnology and bioengineering*, 86(2), pp.136–148.

Nanmori, T., Watanabe, T., Shinke, R., Kohno, a, *et al.*, 1990. Purification and properties of thermostable xylanase and beta-xylosidase produced by a newly isolated *Bacillus stearothermophilus* strain. *Journal of bacteriology*, 172(12), pp.6669–6672.

Nanmori, T., Watanabe, T., Shinke, R., Kohno, A., *et al.*, 1990. Purification and properties of thermostable xylanase and beta-xylosidase produced by a newly isolated thermostable xylanase Strain. *Journal Of Bacteriology*, 172(12), pp.6669–6672.

Napolitano, D.R. *et al.*, 2008. Identification of Mycobacterium tuberculosis ornithine carboamyltransferase in urine as a possible molecular marker of active *Pulmonary Tuberculosis*. *Clinical abd vaccine immunology*, 15(4), pp.638–643.

Oulas, A. *et al.*, 2015. Metagenomics : Tools and insights for analyzing Next-Generation Sequencing data derived from biodiversity studies. *Bioinformatics and biology insight*, 15(9), pp.75–88.

Papadopoulos, J.S. & Agarwala, R., 2007. COBALT: Constraint-based alignment tool for

- multiple protein sequences. *Bioinformatics*, 23(9), pp.1073–1079.
- Pelletier, E. *et al.*, 2008. Candidatus Cloacamonas Acidaminovorans: Genome sequence reconstruction provides a first glimpse of a new bacterial division. *Journal of Bacteriology*, 190(7), pp.2572–2579.
- Perez, J. *et al.*, 2002. Biodegradation and biological treatments of cellulose , hemicellulose and lignin : an overview. *International microbiology*, 5(2), pp.53–63.
- Piel, J., 2002. A polyketide synthase-peptide synthetase gene cluster from an uncultured bacterial symbiont of Paederus beetles. *PNAS*, 99(22), pp.14002–14007.
- Pierce, J. & Gutteridge, S., 1985. Large-scale preparation of ribulosebiphosphate carboxylase from a recombinant system in *Escherichia coli* characterized by extreme plasmid instability. *Applied environmental microbiology*, 49(5), pp.1094–1100.
- Polizeli, M.L.T.M. *et al.*, 2005. Xylanases from fungi: properties and industrial applications. *Applied Microbiology and Biotechnology*, 67(5), pp.577–591.
- Polz, M.F. & Cavanaugh, C.M., 1998. Bias in template-to-product ratios in multitemplate PCR. *Applied and Environmental Microbiology*, 64(10), pp.3724–3730.
- Postma, D., 2012. Modification of the wood based hemicelluloses for use in the pulp and paper. M.sc.thesis, *University of Stellenbosch*. <http://hdl.handle.net/10019.1/20174>
- Purohit, M.K. & Singh, S.P., 2009. Assessment of various methods for extraction of metagenomic DNA from saline habitats of coastal Gujarat (India) to explore molecular diversity. *Letters in applied microbiology*, 49(3), pp.338–44.

Py, B. *et al.*, 1991. Cellulase EGZ of *Erwinia chrysanthemi*: structural organization and importance of His98 and Glu133 residues for catalysis. *Protein engineering, design & selection : PEDS*, 4(3), pp.325–333.

Raes, J., Foerstner, K.U. & Bork, P., 2007. Get the most out of your metagenome : computational analysis of environmental sequence data. *Current opinion in Microbiology*, 10(15), pp.490-498.

Reddy, P., Peterkofsky, a & McKenney, K., 1985. Translational efficiency of the *Escherichia coli* adenylate cyclase gene: mutating the UUG initiation codon to GUG or AUG results in increased gene expression. *Proceedings of the National Academy of Sciences of the United States of America*, 82(17), pp.5656–5660.

Rensburg, P. Van & Pretorius, I.S., 2000. Enzymes in wine making : Harnessing natural catalysts for efficient biotransformations - A Review. *South african journal of enology and viticulture*, 21(2000), pp.52-73.

Rhoads, A. & Au, K.F., 2015. PacBio sequencing and its applications. *Genomics, Proteomics & Bioinformatics*, 13(5), pp.278–289.

Rodriguez-R, L.M. & Konstantinidis, K.T., 2014. Nonpareil: A redundancy-based approach to assess the level of coverage in metagenomic datasets. *Bioinformatics*, 30(5), pp.629–635.

Rondon, M.R. *et al.*, 2000. Cloning the soil metagenome : a strategy for accessing the genetic and functional diversity of uncultured microorganisms. *Applied environmental microbiology*, 66(6), pp.2541–2547.

Rosano, G.L. & Ceccarelli, E. a., 2014. Recombinant protein expression in *Escherichia coli*:

Advances and challenges. *Frontiers in Microbiology*, 5(4), pp.1–17.

Saha, B.C., 2000. Alpha - L -Arabinofuranosidases : Biochemistry , molecular biology and application in biotechnology. *Biotechnology advances*, 18(5), pp.403–423.

Salah, A. *et al.*, 2012. Tapping uncultured microorganisms through metagenomics for drug discovery. *African journal of Biotechnology*, 11(92), pp.15823–15834.

Sanger, F. & Nicklen, S., 1977. DNA sequencing with chain-terminating. *PNAS*, 74(12), pp.5463–5467.

Sastalla, I. *et al.*, 2009. Codon-optimized fluorescent proteins designed for expression in low-GC gram-positive bacteria. *Applied and Environmental Microbiology*, 75(7), pp.2099–2110.

Saxena, I.M. *et al.*, 1995. Multidomain architecture of beta-glycosyltransferases : Implications for mechanism of action. *Journal of bacteriology*, 177(6), pp.1419–1424.

Schlegel, S. *et al.*, 2013. Optimizing heterologous protein production in the periplasm of *E . coli* by regulating gene expression levels. *Microbial Cell Factories*, 12(24), pp.1-12.

Schmeisser, C., Steele, H. & Streit, W.R., 2007. Metagenomics , biotechnology with non-culturable microbes. *Applied microbiology and Biotechnology*, 75(5), pp.955–962.

Schmidt, O., Drake, H.L. & Horn, M. a, 2010. Hitherto unknown [Fe-Fe]-hydrogenase gene diversity in anaerobes and anoxic enrichments from a moderately acidic fen. *Applied and environmental microbiology*, 76(6), pp.2027–2031.

- Scholz, M.B., Lo, C.-C. & Chain, P.S., 2012. Next generation sequencing and bioinformatic bottlenecks: the current state of metagenomic data analysis. *Current Opinion in Biotechnology*, 23(1), pp.9–15.
- Sepp, A. & Choo, Y., 2005. Cell-free selection of zinc finger DNA-binding proteins using *in vitro* compartmentalisation. *Journal of molecular biology*, 354(2), pp.212–219.
- Shallom, D. & Yuval, S., 2003. Microbial hemicellulases. *Current opinion in microbiology*, 6(3) pp.219–228.
- Shao, W. *et al.*, 2011. Characterization of a Novel Xylosidase , XylC, from *Thermoanaerobacterium saccharolyticum* JW/SL-YS485. *Applied environmental microbiology*, 77(3), pp.719–726.
- Shao, W. & Wiegel, J., 1992. Purification and Characterization of a Thermostable 13-Xylosidase from *Thermoanaerobacter ethanolicus*. *Journal of Bacteriology*, 174(18), pp.5848–5853.
- Sharma, A., Adhikari, S. & Satyanarayana, T., 2006. Alkali-thermostable and cellulase-free xylanase production by an extreme thermophile *Geobacillus thermoleovorans*. *World Journal of Microbiology and Biotechnology*, 23(4), pp.483–490.
- Sharma, P. *et al.*, 2008. From bacterial genomics to metagenomics: concept, tools and recent advances. *Indian Journal of Microbiology*, 48(2), pp.173–194.
- Sharma, P.K., Capalash, N. & Kaur, J., 2007. An improved method for single step purification of metagenomic DNA. *Molecular Biotechnology*, 36(1), pp.61–63.

- Sharma, S.S., Blattner, F.R. & Harcum, S.W., 2007. Recombinant protein production in an *Escherichia coli* reduced genome strain. *Metabolic engineering*, 9(2), pp.133–141.
- Shi, H. *et al.*, 2013. Biochemical properties of a novel thermostable and highly xylose-tolerant β -xylosidase / α -arabinosidase from *Thermotoga thermarum*. *Biotechnology for Biofuels*, 6(1), pp.1-10
- Shimizu, Y., Kanamori, T. & Ueda, T., 2005. Protein synthesis by pure translation systems. *Methods*, 36(3), pp.299–304.
- Shin, J. & Noireaux, V., 2010. Efficient cell-free expression with the endogenous *E. coli* RNA polymerase and sigma factor 70. *Journal of biological engineering*, 4(8), pp.1-9.
- Shrestha, P., Holland, T.M. & Bundy, B.C., 2012. Streamlined extract preparation for *Escherichia coli*-based cell-free protein synthesis by sonication or bead vortex mixing. *Biotechniques*, 53(3), pp.163–74.
- Silhavy, T.J., Kahne, D. & Walker, S., 2010. The bacterial cell envelope. *Cold Spring Harbor perspectives in biology*, 2(5), a000414.
- Simon, C. *et al.*, 2009. Rapid identification of genes encoding DNA polymerases by function-based screening of metagenomic libraries derived from glacial ice. *Applied and environmental microbiology*, 75(9), pp.2964–2968.
- Simon, C. & Daniel, R., 2011. Metagenomic Analyses: Past and Future Trends. *Applied and Environmental Microbiology*, 77(4), pp.1153–1161.
- Singh, B.K. & Macdonald, C. a, 2010. Drug discovery from uncultivable microorganisms.

Drug Discovery Today, 15(17-18), pp.792–799.

Singh, S.P., Sagar, K. & Konwar, B.K., 2013. Strategy in metagenomic DNA isolation and computational studies of humic acid. *Current research in microbiology and Biotechnology*, 1(1), pp.9–11.

Stapleton, J. a & Swartz, J.R., 2010. Development of an *in vitro* compartmentalisation screen for high-throughput directed evolution of [FeFe] hydrogenases. *PloSOne*, 5(12), e15275.

Stoletzki, N. & Eyre-walker, A., 2007. Synonymous codon usage in *Escherichia coli*: Selection for translational accuracy. *Molecular biology and evolution*, 24(2), pp.374–381.

Stolt, P. & Stoker, N.G., 1996. Functional definition of regions necessary for replication and incompatibility in the *Mycobacterium fortuitum* plasmid pAL5000. *Microbiology*, 142(10), pp.2795–2802.

Subramaniyan, S. & Prema, P., 2002. Biotechnology of microbial xylanases: enzymology, molecular biology, and application. *Critical reviews in biotechnology*, 22(1), pp.33–64.

Sudek, S. *et al.*, 2007. Identification of the putative bryostatin polyketide synthase gene cluster from “*Candidatus endobugula sertula*”, the uncultivated microbial symbiont of the marine bryozoan *Bugula neritina*. *Journal of Natural Products*, 70(1), pp.67–74.

Suzuki, H. & Okazaki, F., 2013. Genome mining and motif modifications of glycoside hydrolase family 1 members encoded by *Geobacillus kaustophilus* HTA426 provide thermostable 6-phospho- β -glycosidase and β -fucosidase. *Applied microbiology and biotechnology*, 97(7), pp.2929–2938.

- Tanaka, T. *et al.*, 2014. A hidden pitfall in the preparation of agar media undermines microorganism cultivability. *American Society for Microbiology*, 24, pp.7659–7666.
- Tanaka, Y., 2004. *Catellibacterium nectarophilum* genes which requires a diffusible compound from a strain related to the genus *Sphingomonas* for vigorous growth. *International Journal of Systematic and Evolutionary Microbiology*, 54(3), pp.955–959.
- Tasse, L. *et al.*, 2010. Functional metagenomics to mine the human gut microbiome for dietary fiber catabolic enzymes. *Genome Research*, 20(11), pp.1605–1612.
- Tawfil, D. & Graffiths, A., 1998. Man-made cell-like compartments for molecular evolution. *Nature Biotechnology*, 16, pp.652–656.
- Terrasán, C.R. *et al.*, 2013. Xylanase and xylosidase from *Penicillium janczewskii*: Production, physico-chemical properties, and application of the crude extract to pulp biobleaching. *Bioresources*, 8(1), pp.1292–1305.
- Thongaram, T. *et al.*, 2012. Metagenomic analysis of novel lignocellulose-degrading enzymes from higher termite guts inhabiting microbes. *Journal of microbiology and biotechnology*, 22(4), pp.462–469.
- Treadway, S. *et al.*, 1999. Isolation and characterization of indene bioconversion genes from *Rhodococcus strain* 124. *Applied microbiology and biotechnology*, 51(6), pp.786–793.
- Troeschel, S.C. *et al.*, 2010. Novel tools for the functional expression of metagenomic DNA. *Methods in molecular biology*, 668(10), pp.117–139.
- Uchiyama, T. *et al.*, 2005. Substrate-induced gene-expression screening of environmental

metagenome libraries for isolation of catabolic genes. *Nature Biotechnology*, 23(1), pp.88–93.

Uchiyama, T. & Miyazaki, K., 2009. Functional metagenomics for enzyme discovery: challenges to efficient screening. *Current opinion in biotechnology*, 20(6), pp.616–622.

Uchiyama, T. & Miyazaki, K., 2010. Product-Induced Gene Expression , a product-responsive reporter assay used to screen metagenomic libraries for enzymes encoding genes. *Applied and environmental microbiology*, 76(21), pp.7029–7035.

Uchiyama, T. & Watanabe, K., 2007. The SIGEX Scheme: High throughput screening of environmental metagenomes for the isolation of novel catabolic genes. *Biotechnology and genetic engineering review*, 24, pp.107–116.

Uroz, S. *et al.*, 2013. Functional assays and metagenomic analyses reveals differences between the microbial communities inhabiting the soil horizons of a norway spruce plantation. *PLoS ONE*, 8(2), e55929.

Varaljay, V. a *et al.*, 2010. Deep sequencing of a dimethylsulfoniopropionate-degrading gene (*dmdA*) by using PCR primer pairs designed on the basis of marine metagenomic data. *Applied and environmental microbiology*, 76(2), pp.609–617.

Villegas, A. & Kropinski, A.M., 2008. An analysis of initiation codon utilization in the domain bacteria - Concerns about the quality of bacterial genome annotation. *Microbiology*, 154(9), pp.2559–2561.

Voget, S. *et al.*, 2003. Prospecting for novel biocatalysts in a soil metagenome. , 69(10), pp.6235–6242.

- Wagschal, K. *et al.*, 2008. Cloning, expression and characterization of a glycoside hydrolase family 39 xylosidase from *Bacillus halodurans* C-125. *Applied biochemistry and biotechnology*, 146(1-3), pp.69–78.
- Wallecha, A. & Mishra, S., 2003. Purification and characterization of two beta-glucosidases from a thermo-tolerant yeast *Pichia etchellsii*. *Proteins and Proteomics*, 1649(1), pp.74–84.
- Walton, S.L. *et al.*, 2010. Production of lactic acid from hemicellulose extracts by *Bacillus coagulans* MXL-9. *Journal of Industrial Microbiology and Biotechnology*, 37(8), pp.823–830.
- Warnecke, F. *et al.*, 2007. Metagenomic and functional analysis of hindgut microbiota of a wood-feeding higher termite. *Nature*, 450(7169), pp.560–565.
- Williamson, L.L. *et al.*, 2005. Intracellular screen to identify metagenomic clones that induce or inhibit a quorum-sensing biosensor. 71(10), pp.6335–6344.
- Wong, H.. & Chang, S., 1986. Identification of a positive retroregulator that stabilizes mRNAs in bacteria. *PNAS*, 83(4), pp.3233–3237.
- Wu, N. *et al.*, 2011. A double-emulsion microfluidic platform for *in vitro* green fluorescent protein expression. *Journal of Micromechanics and Microengineering*, 21(11), p054032.
- Xiong, J. *et al.*, 2007. Molecular cloning of a bifunctional b -xylosidase / a - L - arabinosidase from alfalfa roots : heterologous expression in *Medicago truncatula* and substrate specificity of the purified enzyme. , 58(11), pp.2799–2810.

- Yonezawa, M. *et al.*, 2003. DNA display for *in vitro* selection of diverse peptide libraries. *Nucleic acids research*, 31(19), p.e118.
- Yuan, K. *et al.*, 2015. Exchange of active site residues alters substrate specificity in extremely thermostable beta-glycosidase from *Thermococcus kodakarensis* KOD1. *Enzyme and Microbial Technology*, 77, pp.14–20.
- Yun, J. & Ryu, S., 2005. Screening for novel enzymes from metagenome and SIGEX, as a way to improve it. *Microbial Cell Factories*, 4(1), pp.1-15
- Zhang, Y. *et al.*, 2014. Impact of electrolytes on double emulsion systems (W/O/W) stabilized by an amphiphilic block copolymer. *Colloids and Surfaces B: Biointerfaces*, 122(10), pp.368–374.
- Zhou, J. *et al.*, 2012. Beta-xylosidase activity of a GH3 glucosidase/xylosidase from yak rumen metagenome promotes the enzymatic degradation of hemicellulosic xylans. *Letters in Applied Microbiology*, 54(2), pp.79–87.
- Zhou, J., Bruns, M. a & Tiedje, J.M., 1996. DNA recovery from soils of diverse composition. *Applied and Environmental Microbiology*, 62(2), pp.316–322.
- Zimbardi, A.L.R.L. *et al.*, 2013. Optimization of β -glucosidase , β -xylosidase and xylanase production by *Colletotrichum graminicola* under solid-state fermentation and application in raw sugarcane trash saccharification. *International journal of molecular science*, 14(9), pp.2875–2902.
- Zylicz-Stachula, A. *et al.*, 2014. Modified “one amino acid-one codon” engineering of high GC content TaqII-coding gene from thermophilic *Thermus aquaticus* results in radical expression increase. *Microbial cell factories*, 13(7), pp.3-19 .

APPENDICES

Appendix A: Tables of ORFs identified from clone library

Table 17: Blastx results of genes contained in P17C1

ORF	Highest identity	Identity (%)	EC number
P17C1.1			
P17C1-1	hypothetical protein [<i>Chondromycescrocatus</i>]	40	WP_050428680.1
P17C1-2	dimethylhistidine N-methyltransferase [<i>Vulgatibacter incomptus</i>]	55	WP_050727294.1
P17C1-3	class II glutamine amidotransferase [<i>Vulgatibacter incomptus</i>]	60	WP_050727295.1
P17C1-4	ergothioneine biosynthesis protein EgtB [<i>Vulgatibacter incomptus</i>]	44	WP_050727296.1
P17C1-5	hypothetical protein [<i>Sorangiumcellulosum</i>]	43	WP_012234306.1
P17C1-6	hypothetical protein [<i>Sandaracinusamylolyticus</i>]	68	WP_053236857.1
P17C1-7	hypothetical protein AKJ09_05333 [<i>Labilithrixluteola</i>]	43	AKU98669.1
P17C1-8	diguanylate cyclase response regulator [<i>Desulfuromonas</i> sp. TF]	45	WP_035056363.1
P17C1-9	hypothetical protein AYL33_000670 [<i>CandidatusBathymarchaeota archaeon B63</i>]	42	KYH42790.1
P17C1-10	glycosyltransferase [<i>Sorangiumcellulosum</i>]	57	KYF71941.1
P19C1-11	hypothetical protein [<i>Sorangiumcellulosum</i>]	52	WP_061614813.1
P17C1-12	ornithine carbamoyltransferase [<i>Myxococcales bacterium SG8_38</i>]	57	KPK12947.1
P17C1-13	acetylornithine aminotransferase [<i>Sorangium cellulosum</i>]	63	WP_012232639.1
P17C1.2			
P17C1-14	hypothetical protein AWW69_09755 [<i>Bacillus cereus</i>]	96	KWW51416.1
P17C1-15	alpha-glucuronidase [<i>Melioribacterroseus</i>]	64	WP_014857156.1
P17C1-16	LacI family transcriptional regulator [<i>Bacteroides</i> sp. SM23_62_1]	60	KPK85919.1
P17C1-17	sodium solute transporter superfamily protein [<i>Haliscomenobacter hydrossis</i>]	76	WP_013766271.1
P17C1-18	glucan 1,4-alpha-glucosidase [<i>Teredinibacter</i> sp. 1162T.S.0a.05]	63	WP_052705472.1
P17C1-19	glycosyl hydrolase [<i>Nafulsella turpanensis</i>]	69	WP_017730211.1
P17C1-20	copper transporter [<i>Sediminibacter</i> sp. Hel_I_10]	81	WP_026754896.1

Table 18: Blastx results of genes contained in P15A9

ORF	Highest identity	Identity (%)	EC number
P15A9.1			
P15A9-1	Isochorismatase EC 3321 CDS [<i>Bradyrhizobium</i> sp.]	42	CUU22097.1
P15A9-2	hypothetical protein [<i>Gimesiamaris</i>]	38	WP_002647960.1
P15A9-3	hypothetical protein AMJ85_05725 [Candidate division BRC1 bacterium SM23_51]	57%	KPL10432.1
P15A9-4	hypothetical protein [<i>Gimesiamaris</i>]	44	WP_002643619.1
P15A9-5	hypothetical protein [<i>Gimesiamaris</i>]	44	WP_002643619.1
P15A9-6	hypothetical protein AW736_07565 [<i>Opitutaceae</i> bacterium TSB47]	73%	OAM90494.1
P15A9-7	glycoside hydrolase [<i>CandidatusKoribacterversatilis</i>]	41%	CDA85370.1
P15A9-8	transposase [<i>Methylothermobacter</i>]	38%	WP_012510657.1
P15A9-9	hypothetical protein [<i>Parachlamydiaacanthamoebae</i>]	40%	WP_006340994.1
P15A9.2			
P15A9-10	glycosyl hydrolase [<i>Nafulsella turpanensis</i>]	69	WP_017730211.1
P15A9-11	sodium solute transporter superfamily protein [<i>Haliscomenobacter hydrossis</i>]	76	WP_013766271.1
P15A9-12	Lacl family transcriptional regulator [<i>Bacteroides</i> sp. SM23_62_1]	60	KPK85919.1
P15A9-13	alpha-glucuronidase [<i>Melioribacterroseus</i>]	64	WP_014857156.1
P15A9-14	hypothetical protein AWW69_09755 [<i>Bacillus cereus</i>]	96	KWW51416.1
P15A9-15	hypothetical protein AWW69_09755 [<i>Bacillus cereus</i>]	96	KWW51416.1
P15A9-16	Glutamine--tRNA ligase [<i>Cesiribacter andamanensis</i> AMV16]	79	EMR03274.1
P15A9-17	hypothetical protein [<i>Hassalliabysssoidea</i>]	51	WP_039746484.1
P15A9-18	hypothetical protein [<i>Hassalliabysssoidea</i>]	53	KIF34477.1

Table 19: Blastx results of genes contained in P118H6

ORF	Highest identity	Identity (%)	Accession
P118H6.1			
P118H6-1	beta-glycosidase [<i>Thermobifida fusca</i>]	99	WP_011292055.1
P118H6-2	dipeptide ABC transporter ATP-binding protein [<i>Thermobifida fusca</i>]	100	WP_061783668
P118H6-3	dipeptide ABC transporter ATP-binding protein [<i>Thermobifida fusca</i>]	100	WP_061783668
P118H6-4	ABC transporter permease [<i>Thermobifida fusca</i>]	99	WP_011292058.1
P118H6-5	ABC transporter permease [<i>Thermobifida fusca</i>]	99	WP_011292058.1
P118H6-6	ABC transporter substrate-binding protein [<i>Thermobifida fusca</i>]	99	WP_061783669
P118H6-7	23S rRNAmethyltransferase [<i>Thermobifida fusca</i>]	100	WP_041428040
P118H6-8	DNA polymerase subunit beta [<i>Thermobifida fusca</i>]	100	WP_041428042.1
P118H6.2			
P118H6-9	thioredoxin reductase [<i>Streptomyces griseorubens</i>]	67	KEG43590.1
P118H6-10	acyl-CoA dehydrogenase [<i>Actinospicaacidiphila</i>]	99	WP_033277204.1
P118H6-11	hypothetical protein [<i>Actinospicaacidiphila</i>]	97	WP_033277206.1
P118H6-12	MBL fold metallo-hydrolase [<i>Streptomyces griseorubens</i>]	97	WP_037638335.1
P118H6-13	xylose repressor [<i>Streptomyces gancidicus</i>]	98	WP_006135179.1

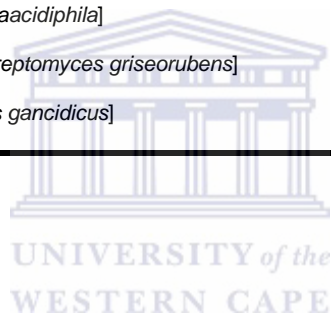


Table 20: Blastx results of genes contained in P117F11

ORF	Highest identity	Identity (%)	Accession
P117F11.1			
P117F11-1	glucose-fructose oxidoreductase [<i>Phaeodactylibacter xiamenensis</i>]	63	WP_044222474.1
P117F11-2	hypothetical protein [<i>Devosia</i> sp. A16]	63	WP_055046703.1
P117F11-3	sugar ABC transporter permease [<i>Hahellagangh wensis</i>]	75	WP_020410100.1
P117F11-4	multidrug ABC transporter ATP-binding protein [<i>Caulobacter</i> sp. AP07]	74	WP_007668705.1
P117F11-5	hypothetical protein [<i>Sandaracinus amyolyticus</i>]	56	WP_053230726.1
P117F11-6	hypothetical protein [<i>Nodosilineanodulosa</i>]	44	WP_017296853.1
P117F11-7	hypothetical protein [<i>Salinibacter ruber</i>]	34	WP_051010711.1
P117F11-8	two-component sensor histidine kinase [<i>Cystobacter fuscus</i>]	54	WP_002629048.1
P117F11-9	hypothetical protein UT69_C0013G0054 [<i>Candidatus Yanofskybacteria</i> bacterium GW2011_GWE1_40_10]	48	KKR36832.1
P117F11-10	hypothetical protein [<i>Sorangium cellulosum</i>]	47	WP_020741934.1
P117F11-11	hypothetical protein BE21_22790 [<i>Sorangium cellulosum</i>]	36	KYG08578.1
P117F11-12	hypothetical protein [<i>Rhizobacter</i> sp. Root1221]	46	WP_056655007.1
P117F11-13	beta-xylosidase [<i>Paenibacillus mucilaginosus</i>]	63	WP_014369630.1
P117F11.2			
P117F11-14	hypothetical protein [<i>Paenibacillus ginsengihumi</i>]	53	WP_019534959.1
P117F11-15	peptide ABC transporter ATP-binding protein [<i>Paenibacillus terrigena</i>]	70	WP_018756692.1
P117F11-16	peptide ABC transporter ATP-binding protein [<i>Paenibacillus terrigena</i>]	69	WP_018756693.1
P117F11-17	diguanylate cyclase [<i>Paenibacillus wuensis</i>]	71	ANE47356.1
P117F11-18	peptide ABC transporter permease [<i>Paenibacillus wuensis</i>]	67	ANE47355.1
P117F11-19	peptide-binding protein [<i>Cohnella</i> sp. VKM B-2846]	54	WP_041064887.1
P117F11-20	hypothetical protein [<i>Paenibacillus</i> sp. VKM B-2647]	40	WP_041046243.1
P117F11-21	precorrin-3B C(17)-methyltransferase [<i>Paenibacillus</i> sp. UNC451MF]	69	WP_028549263.1
P117F11-22	hypothetical protein SY83_14940 [<i>Paenibacillus wuensis</i>]	45	ANE47350.1

Appendix B: Full nucleotide and predicted amino acid sequences (in standard one letter code) for each of the genes targeted in this study. The stop codon is marked by an asterisk.

ORF P17C1-18

1 ATGATGCAAACCAAATACCGTGTAGCACGAACAGGCTATGCCGTCCTCTTTCTCCTGATT
1 M M Q T K Y R V A R T G Y A V L F L L I

61 TCATTTTGGCGCCGGCGCAGACCGAAATCTTCCGGGATCCGAATCAACCACTCGATGCA
21 S F C A G A Q T E I F R D P N Q P L D A

121 AGGATCAACGATCTTGTGCGACGATTGACTCTGGAGGAAAAGGCCCTCCAGATGATGCAT
41 R I N D L V A R L T L E E K A L Q M M H

181 CGCAGTCCGGCGATTCCCCGATTGAATATTCCCGCATACAATTGGTGAACGAGGCACTT
61 R S P A I P R L N I P A Y N W W N E A L

241 CATGGTGTGGTTCGTTCCGGCGTTGCTACGATCTTCCCCAGGCAATCGGTCTCGGCGCT
81 H G V G R S G V A T I F P Q A I G L G A

301 ACGTTTGACGAAGACCTCATATTTAGAGTCGCCACCGCGATCTCGGATGAGGCCCGCGCC
101 T F D E D L I F R V A T A I S D E A R A

361 AACTACAACGTGTCCACCAAAAAAGGATATCATCTTCAATACAGTGGACTGACGTTCTGG
121 N Y N V S T K K G Y H L Q Y S G L T F W

421 ACACCAAACATCAACATCTTTCGCGATCCTCGGTGGGGCAGAGGGCAGGAGACTTATGGA
141 T P N I N I F R D P R W G R G Q E T Y G

481 GAAGATCCATACCTGACGTCGCGACTGGGAATGGCCTTCGTGCGGGGGCTCCAGGGTGAT
161 E D P Y L T S R L G M A F V R G L Q G D

541 CACCCAGGATATCTCAAGACGGCAGCGTGCGCCAAACACTTTGCAGTCCATAGCGGTCCC
181 H P G Y L K T A A C A K H F A V H S G P

601 GAACGGTTGCGCCATGAGTTCGACGCAACGTCTTCGAAAAAGGATCTTTATGAACTTAC
201 E R L R H E F D A T S S K K D L Y E T Y

661 TTACCCGCATTCAAGGCACTTGTGAAACGGGAGTTGAAGCAGTGATGTGCGCTTACAAC
221 L P A F K A L V E T G V E A V M C A Y N

721 AGCACAAACGGTGAACCTGTTGTTCAAACAGCTATCTCCTTCAGGATGTGTTGCGTGGT
241 S T N G E P C C S N S Y L L Q D V L R G

781 GAATGGCAGTTCAAGGGTCACATTGTGAGCGACTGTTGGGCTTTGGTCGATTTGTATTCC
261 E W Q F K G H I V S D C W A L V D L Y S

841 GATAATGGCCATAAGACCGTACCCACCAAAGAAGAGGCGGTTGCGCTGGCTGTCAAGCGT
281 D N G H K T V P T K E E A V A L A V K R

901 GGTGTTAACCTGAATTGTGGGGATGAGTTTCCCGCGCTCGTGGATGCTGTCAAGAAAGGG
301 G V N L N C G D E F P A L V D A V K K G

961 CTTATCACCGAGAAGGAGATCGACGACGCATTGAAAGTGTGCTTAGAACAAGATTCAAA
321 L I T E K E I D D A L K V L L R T R F K

1021 CTTGGTATGTTTACCCGGTTGATGATAATCCCTGGGCTTCCCTTGATGCAAAAGTCATC
341 L G M F D P V D D N P W A S L D A K V I

1081 GACAGCGACAAGCATCGCCGCTCGCTCGGGAAGCGGCCCTGAAGTCGATCGTCATGCTG
361 D S D K H R R L A R E A A L K S I V M L

1141 AAGAACGATGGTGTCTTGCCCTTGAGGAACGACCTGAAGAAGTATTTTGTACCGGCCCC
381 K N D G V L P L R N D L K K Y F V T G P

1201 AATGCTGAAACTGTCTACGCGCTGATCGGGAACACTACTACGGCGTGAATCCGAATATGGTA
401 N A E T V Y A L I G N Y Y G V N P N M V

1261 ACTTACCTTGAGGGAATTGCTGCTGCCATAGAACCCGGAAGTCAACTGCATTACAAGCCT
421 T Y L E G I A A A I E P G S Q L H Y K P

1321 GGTATCATGATCGATCGGGCGAATGTGAATCCGATTGACTGGACTACGGGTGACGCGAGG
441 G I M I D R A N V N P I D W T T G D A R

1381 TCCAGTGATGCGACTATAGTCGTGTTGGGTCTGACGGGTGCATTGGAAGGCCAAGAAGGT
461 S S D A T I V V L G L T G A L E G E E G

1441 GAATCTATCGCGTCCCCGCATTACGGAGACCGTCTGGATTACAATCTGCCCCAGGTTTCAG
481 E S I A S P H Y G D R L D Y N L P Q V Q

1501 ATCGATTTCTCAAGAAGCTGCGAGAGGGTCATAACAGACCGATCATCGCCGTTATCACA
501 I D F L K K L R E G H N R P I I A V I T

1561 GCGGAAGCCCGATGAACCTCTCGGAAGTGCATGAGCTCGCGGATGCTGTGTTGCTCGTG
521 G G S P M N L S E V H E L A D A V L L V

1621 TGGTATGCCGGTGAGGAGGCTGGTAACGCCCTCGCTGATATCGTTTTTGGAAAGGTGGCT
541 W Y A G E E A G N A L A D I V F G K V A

1681 CCTTCGGGCCGACTACCCATCACGTTTCCGAAATCCCTTGATCAACTGCCTCCGTATGAG
561 P S G R L P I T F P K S L D Q L P P Y E

1741 GATTATAGCATGAAGGGCCGACCTACCGCTACATGGAAGCGGAACCGATGTACCCGTTT
581 D Y S M K G R T Y R Y M E A E P M Y P F

1801 GGCTATGGCCTTAGCTACGCCAAATTCGCTTACGGCGACATAAACTTTCAGCAACAACG
601 G Y G L S Y A K F A Y G D I K L S A T T

1861 ATCAAAAAAGGTCAATCCATTGACGTGGATGTAACCGTCAGCAATCCTGGTCAGCTCGAA
621 I K K G Q S I D V D V T V S N P G Q L E

1921 GCTGAAGATGTCGTGCAATTGTATCTCACTGATTTGGCTTCGAAAGAGGAACAGGTACCG
641 A E D V V Q L Y L T D L A S K E E Q V P

1981 TTGTTTTCGTTGAAGGGCATTAAACGCGTCAACGTGGCGCCGGGACAATCTCAGGTTGTC
661 L F S L K G I K R V N V A P G Q S Q V V

2041 CGGTTTACGATCACCCCTGACATGATGGCGATTGTTAATAACAACGGGGGAATCGGTCGTG
681 R F T I T P D M M A I V N T T G E S V V

2101 GAGCCGGGTGATTTTTCGCGTTTCAGTCGGTGGCGCTGTGCCGATCAAACGAAGCATTGAC
701 E P G D F R V S V G G A V P I K R S I D

2161 CTCGGTGTGAGTGCGCCAGCCCAGGCGACGTTTCAGGGTTCGGTAA
721 L G V S A P A Q A T F R V R *

ORF P17C1-19

1 ATGCAAAATGTTCTTTCCCGCATAAGGAAATCGCTACCCTGCTGCCTGTTGTTCCTTTTT
1 M Q N V L S R I R K S L P C C L L F L F

61 ATGATCGCGTGTTCGGTTCACCAGGATCGGGTTCATTTCCTTTCAGGATAACAACGCTG
21 M I A C S V H Q D R V A F P F Q D T T L

121 GATATCGACACTCGGGTGGAGGACCTTGTTCGCGGGCTCACGCTGGAAGAAAAGATCGGT
41 D I D T R V E D L V A R L T L E E K I G

181 CAGATGATGCACAACGCGCCGGCTATTGAACGCCTGGGTATTCCTGCATACAATTGGTGG
61 Q M M H N A P A I E R L G I P A Y N W W

241 TCGGAAGCATTACACGGCGTTGCTCGCGCCGGACTGGCTACAGTTTATCCGCAGGCCATC
81 S E A L H G V A R A G L A T V Y P Q A I

301 GGCCTTGCCGCAACGTGGGATGAAGACCTCATGTTTCAAGTCGCTACCTCTATTTCCGAT
101 G L A A T W D E D L M F Q V A T S I S D

361 GAGGCACGCGCCAAGCATCATGACTTTGCCCGTAATGGAAAACGATTTATTTACCAGGGA
121 E A R A K H H D F A R N G K R F I Y Q G

421 CTGACTTTCTTTTCTCCCAACATCAACATTTTTCGGGATCCGCGATGGGGCCGCGGGCAA
141 L T F F S P N I N I F R D P R W G R G Q

481 GAAACCTATGGCGAAGATCCGTACCTCTCGGGGAAGAATGGGGGTGCAGTTTGTTCGCGGA
161 E T Y G E D P Y L S G R M G V Q F V R G

541 ATGCAGGGCGACGATCCGGAATACTTCAAGACGATCGCAACGGTCAAACACTTTGCGGTG
181 M Q G D D P E Y F K T I A T V K H F A V

601 CACAGTGGTCCGGAACCTGAGCGCCACAGCTTCGACGCTAAGACCAGCCGGCGTGATCTT
201 H S G P E P E R H S F D A K T S R R D L

661 CTCGATATGTATGTACCACAGTTTGAATGGGTATTCGGGAAGGCAAGGCTTATTCATTG
221 L D M Y V P Q F E M G I R E G K A Y S L

721 ATGTGTGCATACAACCGTTACAACGGGGAGGCGTGTGTGGCAGCGATAAACTCCTCAAC
241 M C A Y N R Y N G E A C C G S D K L L N

781 CAGATGCTAAGGGAAGAATGGGAATTTGAAGGATATGTTGTTTCCGATTGTTGGGCGGTA
261 Q M L R E E W E F E G Y V V S D C W A V

841 TCGGACATCTACCAGTTCACAAGTTGGTTCGATACCCCTGAAGAAGCGGCAGCCCTCGCC
281 S D I Y Q F H K L V D T P E E A A A L A

901 GTGAAGTCGGGTACCGAGTTGGAATGTGGGGAGACTTACCAGACGTTGACCAAAGCCGTT
301 V K S G T E L E C G E T Y Q T L T K A V

961 GAAAAAGGCCTTATTACTGAGAAAAGAAATTGACGTCGCGGTAAAAAACTTTTTAAGGCC
321 E K G L I T E K E I D V A V K K L F K A

1021 CGGTTACAGGCTGGGGTTATTCGACCCGCCGCAAAGGTTAAGTATGCCAGCATCCCGTAC
341 R F R L G L F D P P A K V K Y A S I P Y

1081 GATGTTGTGGACAGTGAGCCGCATAGGGCGCTGGCGCTTGATGCAGCGCGAAAATCAATC
361 D V V D S E P H R A L A L D A A R K S I

1141 GTACTCCTGAAGAATGACCCGTTTCAGCGGTAACCCTGTTTTGCCATTGAAGGGCGATTTA
381 V L L K N D P F S G N P V L P L K G D L

1201 AAAAGGATTGCCGTCATCGGCCCAATGCTGACCAGTGGCTGATGTTATTAGGCAACTAC
401 K R I A V I G P N A D Q W L M L L G N Y

1261 AACGGCGTGCCTTCCGACCCCGTGACTCCACTTGAAGGTATACGTAAAAAGTTTCCCGAC
421 N G V P S D P V T P L E G I R K K F P D

1321 GCAGAGGTACTGTACGCTCAGGGGTGCGAGCTCGCGAAGGGAATGCCGATGTTCAATGTG
441 A E V L Y A Q G C E L A K G M P M F N V

1381 ATACCTGCTACGGCGCTAAGCCACGATGACGAACCTGGCTTGCAAGTTGAATTCCTTTAC
461 I P A T A L S H D D E P G L Q V E F F H

1441 GGTGCCGATTCAATTCAGAACACGTATTTGTGCGAGAGACATGAAACCCTGGATGCGAAT
481 G A G F N S E H V F V E R H E T L D A N

1501 TGGCGAGACAAAGCTCCTCGCGACGACATGGACAACGATGGATTGAGTGTGCGCTGGACC
501 W R D K A P R D D M D N D G F S V R W T

1561 GGAGATCTGTGCGCCAGACGTTACCGCGGAATATCAGCTCGGCGTGATTACCACTTGCAAC
521 G D L S P D V T A E Y Q L G V I T T C N

1621 ACAGAACTGTACTTGAATGATTCTCTGATTGCGGAAACAGTTTATCATACGTTTCGACGAA
541 T E L Y L N D S L I A E T V Y H T F D E

1681 TACGGAGATCCGAGACTGGTGAAGTCAAGTCCCATTTCGACTGGAGGCTGGAAAGAAGTAC
561 Y G D P R L V K S S P I R L E A G K K Y

1741 AGGATCAGGGTTGAAGCAATCGAGTCGTATGCGGATGCCCAAGTCCAAGTGTCTGGGCA
581 R I R V E A I E S Y A D A Q V Q L V W A

1801 AGACCGCAGCCGAATCTAAAGGAGGAAGCCATACGTGTTGCCCGTGATGCCGATGTTGTG
601 R P Q P N L K E E A I R V A R D A D V V

1861 CTTATGTTTATGGGGCTGACTGCGAGGATGGAAGGAGAGGAAATGGACATCGCCATTGAG
621 L M F M G L T A R M E G E E M D I A I E

1921 GGATTCCGCGGAGGCGATCGGACCCGGGTGGACTTGCCGCAGACTCAACAGGATTTGATC
641 G F R G G D R T R V D L P Q T Q Q D L I

1981 CGCAGTATACAAGCACTTGGCAAACCCGTGGTGCTCGTGTGCTCAATGGGAGTGCCTC
661 R S I Q A L G K P V V L V L L N G S A L

2041 GCTGTTAATTGGGCCGACAAGAATGTGCCGGCCATACTCGAGGCATGGTATCCGGGACAG
681 A V N W A D K N V P A I L E A W Y P G Q

2101 GCAGCGGGTGACGCCATCGCCGATGTGCTTGC GGCGACTACAACCCTGCAGGCAGATTG
701 A A G D A I A D V L A G D Y N P A G R L

2161 CCTGTTACGTTTTATCGTTCCGAGAAAAGACCTTCCCCCGTTTGAGGAGTACCAGGTCAGC
721 P V T F Y R S E K D L P P F E E Y Q V S

2221 ACGCAGACGTACCGCTATTTCCCGGTGAAGTGTGTACCCGTTTGGATACGGTTTGTGAGT
741 T Q T Y R Y F P G E V L Y P F G Y G L S

2281 TATACCACGTTTCAGCTACAGCGACCTGGAATTGGAAGAGAATCAACAGGCTGGAGACAGC
761 Y T T F S Y S D L E L E E N Q Q A G D S

2341 GTGCGCGTGTCCGGTACGTCTAAAGAATACTGGTTCTCTCGACGGAGATGAAGTAGTTTCAG
781 V R V S V R L K N T G S L D G D E V V Q

2401 GTCTATGTGTCCCACTTGAATGCTCCTGTAAGAGTTCGGATCCGGTTCGTTGCAGGCTTTC
801 V Y V S H L N A P V R V P I R S L Q A F

2461 GATAGAATTCACCTGAAGGCTGGAGAGGAGAGGGCGGTCACTTTCGTTTTGCCGCCGGCT
821 D R I H L K A G E E R A V T F V L P P A

2521 GCATTCTCTTTGATTGATGCCAACGATGAGCGTGTCGTGCTCCCGGGCGAGTTTGAGATT
841 A F S L I D A N D E R V V L P G E F E I

2581 TCCGTTGGAGGAGGGCAGCCGAAAGCCAGGGGGTCAAATGCGATTTCAAGGCGAATGCGA
861 S V G G G Q P K A R G S N A I S R R M R

2641 CTTATGGAGTAG
881 L M E *

ORF P15A9-7

1 ATGACAAGAACCGGTTCACTTCTGATTGGCGCTCTTGCCCTGCTCACGCTCGGGCGCACCG
1 M T R T R S L L I G A L A L L T L G A P

61 CATGTCCCGGCGGAGCCGCTGCAGAAGCCTTTCGTGACGCCCTCGCTACCAGTCGAGGAG
21 H V P A R A A A E A F R D A S L P V E E

121 CGGGTCAACGATCTGGTGGCGCGGCTGACGCTGGACGAGAAGATCGATCTGCTCGCCGGC
41 R V N D L V A R L T L D E K I D L L A G

181 AATCGTCGCGTCCGGCGCCGTGCCGCGACTGGGTCTGGAACCGCTCCTGCTCGCCGACGGC
61 N R R V G A V P R L G L E P L L L A D G

241 CCGCTCGGCATCCGCAAGCGCGACGCGGCCACTGCGTATCCCGCTTCGATCAGCCTCGCC
81 P L G I R K R D A A T A Y P A S I S L A

301 GCGAGTTGGGATGTGACCTCGCCACGGGTTTCGGTACATCGATCGGCCGCGACAGCCG
101 A S W D V D L A H G F G T S I G R D S R

361 GCGCGTGGTATCCACTTCTCCTCGCACCCGGCACCAACATCATCCGCGTCCGCGACAAC
121 A R G I H F L L A P G T N I I R V P H N

421 GGGCGGAATTTTCGAGTACTACTCCGAGGACCCGTTTCTGTCGTCGGCGATGGTCTCCAC
141 G R N F E Y Y S E D P F L S S A M V S H

481 GTCGTGCGCGGCGTGCAGGGTCAGGGTGTGGTTCGCGACGGTGAAGCACTTTGCCGCCAAC
161 V V R G V Q G Q G V V A T V K H F A A N

541 AATCAGGAGACGAACCGGCAAAGCATCGATGTCGTCGTCGACGAACGTACGCTCCGTGAA
181 N Q E T N R Q S I D V V V D E R T L R E

601 ATCTACCTGCCGGCCTTTTCGCGCCGCGGTTTCAGGAAGGCGGCCGGTTCGGTGTGATGGCG
201 I Y L P A F R A A V Q E G G A G A V M A

661 GCCTACAACAAGCTCAACGGCATGCACGCCCTCGGAGCACGGCTGGCTCCTCCGCGAGGTA
221 A Y N K L N G M H A S E H G W L L R E V

721 TTGAAAGAGCAATGGGAGTTTCAGGGATTCGTGATGTCGACTGGCGCGCCACCCAGAGC
241 L K E Q W E F Q G F V M S D W R A T Q S

781 ACGCTCGGCGGCTCGAGGGCGGACTCGATCTCGAGATGCCGAACCCGGTGCACCTCAAT
261 T L G A L E G G L D L E M P N P V H L N

841 GCCGGCACGATCAAACCGCTGCTGGATTCCGGCCGAATCACGACGGAATTGATCGACGAA
281 A G T I K P L L D S G R I T T E L I D E

901 AAATCCGGCGTCTCTTCCGGGTCATCATCGCCCATGGCTTCTCGACCGTCCCCAGAAG

301 K L R R L F R V I I A H G F L D R P Q K
 961 GACGCGACGATCCCCCTCGATGATCCGAAGAGCGACGCGGTTGCACGCGACATCGCCATG
 321 D A T I P L D D P K S D A V A R D I A M
 1021 CGGGGCACGGTTCTGCTCAAGAACGAAGGCGACCTCCTGCCACTCGAGCGCTCGAAGCTG
 341 R G T V L L K N E G D L L P L E R S K L
 1081 AAGTCCGTCGTCGTGCTCGGTCCCAATGCCGACCGGATGCCGGTCGCCGGAGGCAGCTCC
 361 K S V V V L G P N A D R M P V A G G S S
 1141 GAGGTTTCGTCGTTCCGCCATGTCTCCCTGCTCGAGGGCATCCGCGCGGTCGTGCGTCCG
 381 E V R P F R H V S L L E G I R A V V G P
 1201 GACACGGTGGTGCATGCGATCACCGTCGACGCCAGCGTGAATTCGACCGGCTGTTTCCG
 401 D T V V H A I T V D A Q R E F D R L F P
 1261 GAGTCGCACTACGACGGGCGCTCGACATTGAATTTTCGTCGCCGATTCCTCGACGATGACC
 421 E S H Y D G P L D I E F R P D S S T M T
 1321 GTGCTCGGTTCGTGCCACGGAGGATACGGTGGACATCGACTGGAGCGGACGCCGGCCTGCC
 441 V L G R A T E D T V D I D W S G R R P A
 1381 GAGGGTGTTCGCCACCGACTACTACCACGGGAGTGGAAACGGACGGATCACCGCGCGG
 461 E G V P D P T Y Y H A E W N G R I T A R
 1441 CAGACGGGCAACCATCTGTTTCTCGTGCCTACCCACGGCGCATGACAGTGGAGATCGAC
 481 Q T G N H L F L V R T H G G M T V E I D
 1501 GGCGAAAACTCTTCTTCGCGAACGACAATCCCACCGATGCCGTGCTCTGGGGCGAGACC
 501 G E K L F F A N D N P T D A V L W G E T
 1561 TGGCTGGAAGCGGGACGTTTCGTACGCCCTCAAGCTCCGCTATATGTGCCGGCGCGAGCAT
 521 W L E A G R S Y A L K L R Y M C R R E H
 1621 TCGTACATCAAGCTCGCCTGGGGGCGGCTCCCGCCCCGTTGCGCCGGAGCAGGCCGAG
 541 S Y I K L A W G P A P A P L A P E Q A E
 1681 CTCGTCCGCTCCGCGGACGCGGTTCGTGGTGTGCGGCGGTTTCGATGTCGCCACGGAGGCC
 561 L V R S A D A V V V S A G F D V A T E A
 1741 GAGGGACGCGACCGCCCGTATGCGATTCTTGGCACGCAGCGCGAATTGTTGCGGGCCGTG
 581 E G R D R P Y A I P G T Q R E L L R A V
 1801 ACCGATCTGAACCGCAACACGGTTCGTGCTCAACAGCGGCGGAGTGTGCAAACCGCC
 601 T D L N R N T V V V L N S G G S V E T A
 1861 GACTGGATCGAGCAGGTGCCCGCACTGATGCACGCTGGTACCCGGGCAATCAGGAGGC
 621 D W I E Q V P A L M H A W Y P G Q S G G
 1921 ACCGCGCTCGCCGCTTGTGTTTCGGGGATGCCAACCCCTCCGGCAAACCTCCGTTTCCG
 641 T A L A A L L F G D A N P S G K L P F T
 1981 TGGAAACGGCGATGGGAGGATCATCCGGCGTACGGAACTTTCCGGAAAGTGGAGAACGT
 661 W E R R W E D H P A Y G N F P G S G E R
 2041 GTCACCTACGCCGAAGGCGTTTTTCGTGCGTTACCGTTTCTTCGATGCGCGACAGGTGGAG
 681 V T Y A E G V F V G Y R F F D A R Q V E
 2101 CCGCTGTTCCCTTCGGCCACGGCCTGAGTTACACGACCTTTTCGTTATGACGATCTTCG
 701 P L F P F G H G L S Y T T F R Y D D L R

2161 ATCGATTCTGCGACCGATGAGGACGCGTTGATCGTCTCCTTCGACATCACCAACACCGGC
 721 I D S A T D E D A L I V S F D I T N T G

 2221 AAACGTGCCGGCGCGGAGATCGCGCAAGTGTACATCGCCCCACCCGCTTCGGCTGAACCG
 741 K R A G A E I A Q V Y I A P P A S A E P

 2281 CGCCCGCCACGCGAGCTCAAGGGTTTCGCCCCGGTTAAAACTCGCACCCGGGGAAACACAG
 761 R P P R E L K G F A R L K L A P G E T Q

 2341 CGCGCACAGGTGCGGATCGCCCCGTGCCGATCTCGCGTGGTTTGACGCCGAAACGCGTGCG
 781 R A Q V R I A R A D L A W F D A E T R A

 2401 TGGCGCACCGACGCGGGCACGTACCGCATTCACGTTGGCGCCTCGTCCC GCGATCTTCGT
 801 W R T D A G T Y R I H V G A S S R D L R

 2461 CTTGAAGGCCGCTCTCCGTTCCGACACCGGTGAGTCACCCGCTGCAATGA
 821 L E G R L S V P T P V S H P L Q *

ORF P118H8-1

1 ATGCCCACGGGATCAGGCGACGTCAACCAGCGCGCCTGGACTTCTACGACCGTCTCGTC
 1 M P T G S G D V N Q R G L D F Y D R L V

 61 GATGCGCTGCTCGAAGCGGGGATCACCCCGGTGCCACCCCTTACCCTGGGACCTGCCG
 21 D A L L E A G I T P V P T L Y H W D L P

 121 CAGGCCCTGGAGGATGCCGGCGGCTGGCGTTCCCGCAGCACCGCGGAGCACTTCGCCGCC
 41 Q A L E D A G G W R S R S T A E H F A A

 181 TACACCCGGGCGGTCGTTGCCCGTCTCGGCGACCGGATCCGCCACTGGATCACGCTCAAC
 61 Y T R A V V A R L G D R I R H W I T L N

 241 GAGCCGTTCTGCAGCGCTTTCCTGGGCTATGCCGTGGGCGGCGACGCCCCGGGGCCCG
 81 E P F C S A F L G Y A V G R H A P G A R

 301 GAAGGCACGCCTGCGCTCGCCGCCGCGCACACCCTGCTCCTCGCCCACGGCCTGGCAGTG
 101 E G T P A L A A A H H L L L A H G L A V

 361 CAGGAACTGCGCGCTGCCGATGCGGGCGAGGTCCGCATCACCCCTCAACCCCGACCGGATC
 121 Q E L R A A D A G E V G I T L N P D R I

 421 CTGCCCCCTCCGACAGTGCGGCCGACCGCGCGGCTGCCGACCGCGCGAAACCCCTGCAC
 141 L P A S D S A A D R A A A D R A E T L H

 481 AACCGGGTCTGGTTCGACCCGCTGTTTCGCCGGCCGCTACCCCGGAACGAAGCCGACGTG
 161 N R V W F D P L F A G R Y P A N E A D V

 541 TGGGGAGAGCTCGCGGACGGCTCGTACCGGCGGACGGCGACCTGGAGATCATCGGCCAG
 181 W G E L A D G S Y R R D G D L E I I G Q

 601 CCGCTGGACTTCCCTCGGCATCAACTACTACCGGCCGCTCAAAGTCGCCGACGGCCCGCTG
 201 P L D F L G I N Y Y R P L K V A D G P L

 661 ACCGAAGCCGACCCGGCACGCCGACCGCCGTAGACATCCGCGCCCACCAGCAGCGTTTC
 221 T E A D P A R R T A V D I R A H Q Q R F

 721 GATGGAGTGCGGCACACAGCCATGGACTGGCCGGTTCGTCGCCGAATCCTTCACGGACCTG
 241 D G V R H T A M D W P V V P E S F T D L

 781 CTGCTGGACCTCACGGAGCGCTACCCGAACCTTCCGCCCATCTACATCACAGAGAACGGC
 261 L L D L T E R Y P N L P P I Y I T E N G

841 TCAGCCGAGCACGACGTCGTCTCCCCTGACGGACGGGTGCACGACACGGACCGCATCGCC
281 S A E H D V V S P D G R V H D T D R I A

901 TATCTCAACGACCACCTGCATGCCCTGGCCGCGGCGATCCGCGCCGGGGTGGACGTGCGC
301 Y L N D H L H A L A A A I R A G V D V R

961 GGCTACTTCGTCTGGTCGCTGCTCGACAACCTTCGAGTGGGCGTTCGGATACGAGCGGCGT
321 G Y F V W S L L D N F E W A F G Y E R R

1021 TTCGGCATCGTCCGGGTGGACTACGACACCTTGAACGGCTTCCGAAGGACAGCTACTTC
341 F G I V R V D Y D T L E R L P K D S Y F

1081 TGGTATCAGCGGCTCATCGAGCACCACCGCGCCCGCCACCGCGGCTGA
361 W Y Q R L I E H H R A R H R G *

ORF P17F11-1

1 ATGTTTCATGAACAATCCAATCGTGTATAATGCCCATCATTACCAATGGGCGCTTACGCC
1 M F M N N P I V Y N A H H S P M G A Y A

61 TCATTTACTACTGGGTTACCATGGGGCGAAGGGTGGACTTGGCCTGGAGTTGGACAAACCA
21 S F T L G Y H G A K G G L G L E L D K P

121 GCAGACCAACAAGTATATATATCGGCTTTGAAGCTGCCGAAGGTGGTTACTATGAATCTTTA
41 A D Q Q V Y I G F E A A E G G Y Y E S L

181 CCATTTTTTCGGGGCAGGATTTGATGAAAAGTCGGAGATTCGACGTAGAGAAGGAAGATGAC
61 P F F G A G F D E S R R F D V E K E D D

241 GGAAGCGGAAGAAGAAACGAATTATTCCGTATGCCAGTCAGGACGTTAGCAGAGCATTTC
81 G K R K K K R I I P Y A S Q D V S R A F

301 AATTTAAGTACGGATACTTGGAGCACGAAGGAGTTGTCGTTTACGATTTATTCTCCTGCA
101 N L S T D T W S T K E L S F T I Y S P A

361 CATTTCGATTCCAGACCCCGCTCATGCGTCAGACCATGAACTGATGGATGCGTTAGTTCCT
121 H S I P D P A H A S D H E L M D A L V P

421 GCTGTATTTGCAGAGCTTACGCTGGATAACCGCAATGGTAAGCAAGCTCGTTGCGGGTTT
141 A V F A E L T L D N R N G K Q A R C G F

481 ATTAGCTACAAGGGAAATGCAAGTGAGCCGTATAGCCATATGAGAAAATGGGATCACAAT
161 I S Y K G N A S E P Y S H M R K W D H N

541 GCAGATGGCAAGCTTCGCGGTATCGGTGTTGGGCGTTCAACAGCAATCGTGTGATGGAT
181 A D G K L R G I G V G R S T A I V S M D

601 GAAGATGTAATGACCGCACAAAGGCTTCAATTTGGACGATATATTAAGTGAACGTTTTCAA
201 E D V M T A Q G F N L D D I L S E R F Q

661 GAGAATTGGACATTTGCGCTTGGTGCAACGGGTGCATTACTCGTCGATGTACCTGCTGGT
221 E N W T F A L G A T G A L L V D V P A G

721 CAATGTAAAACATATCGGTTTCGCGATTTGTTTTTATCGCGGTGGGCTGGTGACGACAGGA
241 Q C K T Y R F A I C F Y R G G L V T T G

781 ATCGATGCAGCCTATTATTACACACGTTATTTTAGCAAAATGAGGAAGTCGCCGAATAC
261 I D A A Y Y Y T R Y F S K I E E V A E Y

841 GCACTCACCCACTTCGACAGACTGACTACGTTATGTAAAAATAACAATCGTCTGCTAGAT
 281 A L T H F D R L T T L C K N N N R L L D

 901 CAATCTGCATTGTCTGATGATCAAAAAGTTTATGCTCGCCACGCTATACATAGTTATTAC
 301 Q S A L S D D Q K F M L A H A I H S Y Y

 961 GGAATACAGAGCTCTTAATTGCCGATGGGAAGCCGTTGTGGATTGTGAATGAAGGCGAG
 321 G N T E L L I A D G K P L W I V N E G E

 1021 TATCGGATGATGAATACACTGGACTTAACGGTGGATCAGTTGTTCTTTGAACTGGAGATG
 341 Y R M M N T L D L T V D Q L F F E L E M

 1081 AATCCTTGACGGTTAAAAATGTGCTTGATCAGTTCACATCAAGGTACAGTTATATAGAT
 361 N P W T V K N V L D Q F T S R Y S Y I D

 1141 ACGTTACAGTCAAATGATGGGACGGAATATGAAGGCGGCATTAGCTTTACACATGATATG
 381 T L Q S N D G T E Y E G G I S F T H D M

 1201 GGAGTCATGAATCAATTTTCCCGACCTCAGCATTCGTGTTATGAGAAAATATGGAATTCGC
 401 G V M N Q F S R P Q H S C Y E K Y G I R

 1261 GATTGCTTCTCTCATATGACACACGAGCAATTGGTGAACCTGGGTGAGCTGCGCAACCCTT
 421 D C F S H M T H E Q L V N W V S C A T V

 1321 TACGTTGCATATACGAAAGACGTCGAATGGTTACATGCCAATGCGAGCATATTGCAACAA
 441 Y V A Y T K D V E W L H A N A S I L Q Q

 1381 TGCTTCCAGAGCATGGTCCGTCGCGATCATCCAGATCCGACCAAGCGTAACGGCATGATG
 461 C F Q S M V R R D H P D P T K R N G M M

 1441 TCATTTGATAGCAGTCGTACGATGGGTGGCGCAGAAATTACGACCTACGACAGCCTGGAT
 481 S F D S S R T M G G A E I T T Y D S L D

 1501 GTGTCTTTAGGGCAAGCCGTAACAATATATATATCGCAGGAAAATGCTGGGCGTCTTAT
 501 V S L G Q A R N N I Y I A G K C W A S Y

 1561 GTTGGCTTAGCCAACATATTCACGCAGTTAGGTTGGACAGAGCTGGCGATAGAAGCGGAA
 521 V G L A N I F T Q L G W T E L A I E A E

 1621 CAACAAGCGAAGCGAACAGCTGCTACGATTACAGCTCACTTACAGCCGGGTGGATACATT
 541 Q Q A K R T A A T I T A H L Q P G G Y I

 1681 CCAGCGGTGATCGGTGAGAACAATGACTCCCGAATTATTCCGGCCATTGAAGGTTTGATA
 561 P A V I G E N N D S R I I P A I E G L I

 1741 TTCCCTTATTATACTGGCTGTAAGGAAGCTCTTGCCGAAGATGGTCCATACGGTGAATAT
 581 F P Y Y T G C K E A L A E D G P Y G E Y

 1801 ATTCAAGCATTAAAGAAAACATCTGGAGACCATCCTGGTTCCGGGGACGTGTTTGTTCGAT
 601 I Q A L K K H L E T I L V P G T C L F D

 1861 GATGGCGGCTGGAAGCTGTCCCTCGACAAGCCATAATTCGTGGTTAAGCAAAAATATACTTA
 621 D G G W K L S S T S H N S W L S K I Y L

 1921 TGTCAGTTTATCGCTAGGCGAATTCTCGGAATGAAGTGGGATGAGCAAGATGCGATTGCT
 641 C Q F I A R R I L G M K W D E Q D A I A

 1981 GATCAAGCGCATGTGAGTTGGTTGAAACATTCAGAACTATCTTATTGGTGTGGAGCGAT
 661 D Q A H V S W L K H S E L S Y W C W S D

 2041 CAGATCGTATCTGGCAATATTATTGGAAGTAAATACTACCCACGAGGCGTAACGGCCATC
 681 Q I V S G N I I G S K Y Y P R G V T A I

2101 TTATGGCTGCAAGAACAACCACAAGCTTAA
701 L W L Q E Q P Q A *

mbglX

1 ATGCTGCCCTTCGAGATTGCCGTTCTGGGAGGCCGAGCCGGGCAGCGTCATGTGCGCGTAC
1 M L P F E I A V R E A E P G S V M C A Y
61 AACAAAGCTCACGCTTGATAAATATCCCGCTGATATATTTGCTTGCCAGCATTATCACACG
21 N K L T L D K Y P A D I F A C Q H Y H T
121 TTGACAGAAATTCTCAAGAATGAGTGGGGGTTCAAAGGGCAGGTTTCAGACTGATTGGCAA
41 L T E I L K N E W G F K G Q V Q T D W Q
181 GCCATACATTCTACCGCTGACGCGATCAATGCCGGTGTGATGAGGAAGAGGATTGGCAG
61 A I H S T A D A I N A G V D E E E D W Q
241 GCGGCTACTTTTTTCTTCCCGCTAATGTGAAGCCGCTTCTGAATGACGGAACCATCTCC
81 A A T F F L P A N V K P L L N D G T I S
301 ATATCCCGTCTGGACGATATGGTACGGCGCAAGCTGCGAACCATGATCAAGGTCGGTGTCT
101 I S R L D D M V R R K L R T M I K V G V
361 ATGGACAATCCGCCAGTCGATAACAAAGTCGCCGACTTCAAACCAAAAATCGATTTTGAT
121 M D N P P V D N K V A D F K P K I D F D
421 AGCGGTGCAGCCGTCGCGCAACGGGTGGCTGAAGAATCGATCGTTCTGCTTAAAAATCGG
141 S G A A V A Q R V A E E S I V L L K N R
481 GAACCACAAGCGCCACTCAGTGC CGCACCAAAATGCACACCTCCTGCCATTGAACGCGGTG
161 E P Q A P L S A A P N A H L L P L N A V
541 AATCTGAGAAAGATCGCGGTTATTGGCGCCCATGCCGATGACGCAGTCTTGTCCGGGGGT
181 N L R K I A V I G A H A D D A V L S G G
601 GGTCTGGAAGCACCATACACCCGGTTGGGGGCAGTTACGGCACGTGTGGGAAAGTGAAA
201 G S G S T I H P V G G S Y G T C G K V K
661 CTACATAGAGATGGCAGCTGCGGCTGGTGGGCTATTCCGTGGACGCGGGTACGAACGTGC
221 L H R D G S C G W W A I P W T R V R T S
721 ATCCTTCAGGCAATCAAGGATATTGTGCCGGACGCTGACGTAAATTATGGCGGAAACAGC
241 I L Q A I K D I V P D A D V N Y G G N S
781 GACCGTGATCAGCCGTTTTCGTCCCTACACGGCACAAGAAATAGATGATGCCGTCAATCTG
261 D R D Q P F R P Y T A Q E I D D A V N L
841 GCGAGCACATCGGACGTGGCAATTGTCGTTGTCGCTCAGCCCTCCGGTGAGGACGTGACC
281 A S T S D V A I V V V A Q P S G E D V T
901 TCGCTCTCGCTCAGTTTGGAGCGTTGTTACGATGACAAAGACAATCCAGCAATCAAGAC
301 S L S L S L S V V H D D K D N P S N Q D
961 GAATTGGTAACGAGGGTAGCCGCTGTCAATAAAAAACACGATCGTCATTATCGAAAGTGGAA
321 E L V T R V A A V N K N T I V I I E S G
1021 AATCCTATCCTGATGCCTTGGATCGATAATGTGGCAGCTGTACTGGAAACCTGGTATCCT
341 N P I L M P W I D N V A A V L E T W Y P

1081 GGTGAAAACGGCGGACCGGCTATTGCGAATATCTTGTTCGGCAAGATCAACCCTTCAGGA
361 G E N G G P A I A N I L F G K I N P S G

1141 AAAGTCCGATTACCTTTCCGAAAATGGAAATTGACACCCCGACAGGTGGTGGAGCCTGG
381 K L P I T F P K M E I D T P T G G G A W

1201 TCTGAAAATCCGGTTTATTCCGAAAACTTGAGGTCGGATACCGCTGGTACGATGCAAAA
401 S E N P V Y S E K L E V G Y R W Y D A K

1261 AGCGTAACGCCATCGTTTGAATTCCGGCTTTGGTCTGTCTTACACCAGTTTTTTCGTATTCC
421 S V T P S F E F G F G L S Y T S F S Y S

1321 GAATTGCGGGTGGATACGGACGGAAACGGATGGCACGAAAACAGTGTCTTTTTCAGTTGAA
441 E L R V D T D G T D G T K T V S F S V E

1381 AATACCGGTATGGTTTCCGGGAAAGAGGTGCCGCAAGTCTATATTCAGTTTCCTTCAGCA
461 N T G M V S G K E V P Q V Y I Q F P S A

1441 GCGGGGACCCGCTAAACGGCTTGTTCGGTTGGGAGAAAAGTTGATCTGAAGCCTGGCGAA
481 A G D P P K R L V G W E K V D L K P G E

1501 AAGAAAAAGTGA
501 K K K *



1 ATGAAGAAGATTCTTTTCCGGAGCAGCAATACTTTTCAGTATTTATAGCCAATGCCAGCAA
1 M K K I L F G A A I L S V F I A N A Q Q

61 AAAACCTATGCTAACCCCGTCAATGTAGACTATGGTTACACCCCTATTCCTAATTTTCGCA
21 K T Y A N P V N V D Y G Y T P I P N F A

121 ACACAGGGAAAGCACAGAGCCACTGCAGATCCGGTAATTGTAACTTTCAAAGGAAAATAC
41 T Q G K H R A T A D P V I V T F K G K Y

181 TTTATGTTTTCTACAAACCAATGGGGCTATTGGTGGAGTGACGACATGCTGAACTGGAAA
61 F M F S T N Q W G Y W W S D D M L N W K

241 TTTGTTTCCCGTAAATTCCTTCTTCCACAACATAAGGTATATGATGAGTTGTGTGCACCC
81 F V S R K F L L P Q H K V Y D E L C A P

301 GCTGTCTTTGTAATGAAAGATGCCATGTATGTTATTGGTTCTACTCATAATCCTGATTTCC
101 A V F V M K D A M Y V I G S T H N P D F

361 CCTATCTGAAAAGTACAGATCCAACCAAAGACAATTGGGAAAATTGCAGTAAAAGAATTT
121 P I W K S T D P T K D N W E I A V K E F

421 AAAGTAGGTGCATGGGATCCTGCCTTCCATTATGATGAAGATACAGACAAACTTTATCTA
141 K V G A W D P A F H Y D E D T D K L Y L

481 TATTGGGGTTCCAGTAACGCCTATCCTATTCTGGGAACGGAGATTAATACCAAGACCTTA
161 Y W G S S N A Y P I L G T E I N T K T L

541 CAATCCGAAGGTTATGTAAAACCTCTCTTAGGATTAGAACCTTCAGAACACGGCTGGGAA
181 Q S E G Y V K P L L G L E P S E H G W E

601 AGATTCGGGGAATATAATGACAATACCTTTTTGCCGCCTTTTATAGAAGGTGCATGGATG
201 R F G E Y N D N T F L P P F I E G A W M

661 ACCAAGCATAATGGAAAATATTACCTGCAATATGGTGCTCCGGGAACAGAATTCAGTGGC

221 T K H N G K Y Y L Q Y G A P G T E F S G
 721 TATGGGGATGGTGTATGTAAGCGATAAACCATTAGAGGGTTTCACCTACCAAAGCCAT
 241 Y G D G V Y V S D K P L E G F T Y Q S H
 781 AATCCTTTTTCTTACAAAACCTGGCGGATTTGCCAGAGGAGCCGGACACGGAGCTACATTT
 261 N P F S Y K P G G F A R G A G H G A T F
 841 GAAGATAATTACAAAACTGGTGGCATATTTCTACCATAGTTATATCAACCAAAAAATAAC
 281 E D N Y K N W W H I S T I V I S T K N N
 901 TTTGAAAGGAGAATGGGTATCTGGCCTGCCGATTCGATAAAGATGATGTTATGTACACT
 301 F E R R M G I W P A G F D K D D V M Y T
 961 AATACGGCTTATGGTGACTACCCTACTTACCTTCCACAATATGCACAGGGAAAAGATTTTC
 321 N T A Y G D Y P T Y L P Q Y A Q G K D F
 1021 AGTAAAGGTCTTTTTGCCGGATGGATGCTGCTCAATTATCAGAAACCCGTTTCAGGCTTCC
 341 S K G L F A G W M L L N Y Q K P V Q A S
 1081 TCTACATTAGGTGGATTTTCAGCCAAATCTCGCAGTAGATGAAGATATTTAAAACCTATTGG
 361 S T L G G F Q P N L A V D E D I K T Y W
 1141 AGTGCAAAAACCGGAAATGCCGGAGAATGGTATCAGACAGATTTAGGTGACATTTCTACA
 381 S A K T G N A G E W Y Q T D L G D I S T
 1201 GTCAACGCCATACAGATCAATTATGCTGATCAGGATGCCGAGTTTTTAGGTAAAACGCTG
 401 V N A I Q I N Y A D Q D A E F L G K T L
 1261 AACAAAGATGCATCAGTATAAAAATTTATGCTTCTAATGACGGAAAATCCTGGAAAACAATT
 421 N K M H Q Y K I Y A S N D G K S W K T I
 1321 GTAGACAAAAGTAAAAACCAAAAAGATGTACCACACGATTATATCGAACTGGAAACTCCG
 441 V D K S K N Q K D V P H D Y I E L E T P
 1381 GTGAAAGCACGTTTCTGAAAATGGAAAATTTGAAAATGCCTACCGGAAAAGTTTGCTTTA
 461 V K A R F L K M E N L K M P T G K F A L
 1441 AGCGGATTTTCGTGTATTTGGTAAAGGTACCGGAGCAAACCATCGGCAGTAGAAAACTTT
 481 S G F R V F G K G T G A K P S A V E N F
 1501 GTTGCACTCCGGGCAGAGCCAAGAAAAAATGCTGACAGAAGAAGCGTATGGTTTAAATGG
 501 V A L R A E P R K N A D R R S V W F K W
 1561 AAACAGAATGATCTCGCAGATGGTTATGTTATCTATTTTGGCAAATCTCCGGACAAATTA
 521 K Q N D L A D G Y V I Y F G K S P D K L
 1621 TACGGAAGCATTATGGTGTATGGTAAAGAATGAATACTACTTTTACAGGAGCCGATAAAAAGT
 541 Y G S I M V Y G K N E Y Y F T G A D K S
 1681 GATGCTTATTATTTCCAGATAGAAGCTTTTAATGCTAATGGTATTTCCGGAACGGACATCA
 561 D A Y Y F Q I E A F N A N G I S E R T S
 1741 GTAATGAAATCAGAATAA
 581 V M K S E *

xyIT

1 GTGAGTATTGTTTCCGAGGGAGAAAAATATAAAGGTTTCAGTATGTGTGCTGACTTATAAT
1 V S I V S E G E N I K V S V C V L T Y N

61 CAGGTTGGTTATATCAGCGAGTGCTTGTGAGCTTGATTAATCAGGTCACGGATTTCAAG
21 Q V G Y I S E C L L S L I N Q V T D F K

121 TTTGAGGTGATCGTCCGGTGTGACTGTTTCGACAGATGGCACCAGCGACGTAGTACGTAGT
41 F E V I V G D D C S T D G T S D V V R S

181 TTGGCGGAGCAATACCCCGATATTGTCAAGCCGCTTATACATGAAAAGAATGTTGGTATC
61 L A E Q Y P D I V K P L I H E K N V G I

241 ACGGCAAATTACCTTGAAGTGCACGCTTTGGCATGCGGCAAGTACATTGCTCATCTAGAT
81 T A N Y L E V H A L A C G K Y I A H L D

301 GGTGATGCTTACGCTCTTCCAGGGAAACTACAGGCTCAAAGTGACTTTCTGGACGAGCAC
101 G D A Y A L P G K L Q A Q S D F L D E H

361 CCGAACTATAATATCGTCTGGCACAGGATGCTCGTCGAGAATCCTGGCACCAAGGTTGTT
121 P N Y N I V W H R M L V E N P G T K V V

421 GTAGAAGACCTCATTGACTTATCTCGCGTGAGTAACTCGTTCACCAGAAAGGACATTTTT
141 V E D L I D L S R V S N S F T R K D I F

481 CAATATATAACGATTGGAATGAACAGTTCTAAAAATGTATCGGGCCACGGCAAGGGACATT
161 Q Y I T I G M N S S K M Y R A T A R D I

541 GAACTGCCAGATTTTCCTTTGCTGGATTACTTTGCAAACGTTGAGCAGGTTGGACAAGGC
181 E L P D F P L L D Y F A N V E Q V G Q G

601 TATGCAGGGTTTCGTCAGCGCCAAGCCTTTGGGTGTTTATCGTACAGGTATCGGTATTGCC
201 Y A G F V S A K P L G V Y R T G I G I A

661 TCCTCGGGAAACAAGACAAAAGTTTTGCTGGTGAAGTCGTTTGATTACTTTTTGGAAAAG
221 S S G N K T K V L L V K S F D Y F L E K

721 TACAAAGGTAGTGCCGCAGATATATCGGTGTCTATCAGCTTCTTGTTTATTGCCGCACTG
241 Y K G S A A D I S V S I S F L F I A A L

781 AAAAAAGGCGTTTTGAAGACTGTAACTGTTTTTCCCAATATTCGTTTCGCGCTTTTAGA
261 K N R R F E D C K L F F P I F V R A F R

841 CTGACGACTATACCCAAAGTATGGCGTGCCAGTAAGATGCTTTCAATGCTCAGAATCCCC
281 L T T I P K V W R A S K M L S M L R I P

901 GATGCGGTCAAGAGTAAATGA
301 D A V K S K *

BgCX

1 ATGTTACAGGAAGGTTATATGCTGAAAACAGCGGATTTTGCAGGTCCACAGTATCAGATA
1 M L Q E G Y M L K T A D F A G P Q Y Q I

61 AAGAATAAAAATAGCAGAACTCGCCGGTGAGGATGGTATGCAGGCATTTTATAAAAAGTAT
21 K N K I A E L A G E D G M Q A F Y K K Y

121 CTGGAAAACGGAATTACTAAAAAAGATATAGATGCGCTAAAAATCCTGGGGATTTAATTCT

41 L E N G I T K K D I D A L K S W G F N S
 181 GTAAGATTACCAATGCATTATAACCTTTATACCTTACCTATAGAGAAAAGAGGAGGTA
 61 V R L P M H Y N L Y T L P I E K E E V K
 241 GGAAAGGATACCTGGCTGGAAGAAGGTTTCCGTATGACGGATAATCTTCTGAAATGGTGT
 81 G K D T W L E E G F R M T D N L L K W C
 301 GCAGAAAATAAGATGTACTTGTCTGGATATGCATGCTCTGCCTGGAGGACAGGGGAAT
 101 A E N K M Y L F L D M H A L P G G Q G N
 361 GATGTTAATATTTCCGATAATGATAAATCAAACCGTCTTTGTGGGAGAGTGAAGAAAAT
 121 D V N I S D N D K S K P S L W E S E E N
 421 CAAAGAAAATCTGTTGCACTATGGAAAAAACTTGCCGAACGTTACAAAGATAGTCCCTGG
 141 Q R K S V A L W K K L A E R Y K D S P W
 481 ATTGGCGGGTACGATATTATCAATGAACCTAATTATGGATTTACAGGAAAGAACCTCAAT
 161 I G G Y D I I N E P N Y G F T G K N L N
 541 GGTTGCGATGAAGAATCTAATGCACCATTAAGGAAGTTCATGGTTGATGTTACAAAAGCA
 181 G C D E E S N A P L R K F M V D V T K A
 601 ATTCGCGAAGTAGATCAGAAGCACTTGATTATAAATTGAAGGTAAGTCTGCTGGGGAAATAAC
 201 I R E V D Q K H L I I I E G N C W G N N
 661 TATAAAGGAATATTTCCACTATGGGATAATAATCTGGTGCTAAGTTTCCATAAAACTG
 221 Y K G I F P L W D N N L V L S F H K Y W
 721 AATAAGAATGACCAAACTCTATTAAGCAGATGTTGGAGTATCGTAATCAGTACAATGT
 241 N K N D Q N S I K Q M L E Y R N Q Y N V
 781 CCTATCTGGTTAGGAGAAAAGTGGCGAGAATTCTAATGTATGGTTTACGGAAGCCATAAGC
 261 P I W L G E S G E N S N V W F T E A I S
 841 CTTATGGAAAATAATAATATTGGATGGGCTTTCTGGCCTATGAAAAAGATTGACAATATT
 281 L M E N N N I G W A F W P M K K I D N I
 901 GCCGGAGTAGCCAATGTTAAGATAACACCGGAATATGAAAAGCTACTAAATTACTGGAAA
 301 A G V A N V K I T P E Y E K L L N Y W K
 961 AATGGAGGAGAAAAGCCTTCTAAGGAATTTGCTTATAAAAACAATGATGCAGATAGCTGAC
 321 N G G E K P S K E F A Y K T M M Q I A D
 1021 AATTACAAATTCGAAAATACAGAAGTAAAAAGAGATGTGATCGATGCCATGTTCCGTCAG
 341 N Y K F E N T E V K R D V I D A M F R Q
 1081 ATAAAAAGCAATGAAGTATTACCATATACCAGTCATACAATTCCGGGAAGAATATTTGCT
 361 I K S N E V L P Y T S H T I P G R I F A
 1141 ACGGAATACGATTTGGGAAGGATTGGCGCTGCTTACTATGACAAAAGATGCAATTAACAT
 381 T E Y D L G R I G A A Y Y D K D A I N Y
 1201 CGTATAGATACCGGTGAACAGGTCAACTGGAACCTCCGGAGATAAGATGAGAAACGATGGT
 401 R I D T G E Q V N W N S G D K M R N D G
 1261 GTGGATATTTACAGCAATAAAGATAAAAATTTCCAATGGTTATTATGTAGGAAAAATTGAA
 421 V D I Y S N K D K I S N G Y Y V G K I E
 1321 GATGGTGAGTGGCTAAACTTCACTTTGAAAAGTGTAATAATCAGGAAAATATACTCTGGAA
 441 D G E W L N F T L K S V K S G K Y T L E

1381 ATTCGCTATGCAAATGCAAACAGTGCAGGGCAGTTATCCGTAACGAATAGTAAAGGACAG
461 I R Y A N A N S A G Q L S V T N S K G Q

1441 CAAATAGTGCAAACAGAGCTTCCTTCTACAGGCGGGGATCAGATATGGAAAACGATAACC
481 Q I V Q T E L P S T G G D Q I W K T I T

1501 GTAAAAAATGTAAACATAGCCAAAGGAACAGATAAAAATAAAGCTTCAGTTTGATAAAGGC
501 V K N V N I A K G T D K I K L Q F D K G

1561 AGTTTCAATCTGAATTATATAGAATTTAAATAA
521 S F N L N Y I E F K *

xyIB

1 ATGAAAAGCCTAAAGATAAAAAAAGAAGGAATTCCTTTTGAAGAAAACAGATTACAGTTTT
1 M K S L K I K K E G I L L K K T D Y S F

61 GAAAGTGAAGGTGTTTTAAATCCTGCTGTAATAAAGGATAATAAAAATAATACATTTATTT
21 E S E G V L N P A V I K D N K I I H L F

121 TACCGTGCTGTAGCTAAAGGAAATTTTTCAAGCATCGGGTACTGCCAATTGTCCGATCCT
41 Y R A V A K G N F S S I G Y C Q L S D P

181 CTTCATATAGAAAATCGAAAAGATGTTCTGTTTTGGTACCTGAATATGATTATGAAAAA
61 L H I E N R K D V P V L V P E Y D Y E K

241 CAAGGTATGGAAGATCCGCGCATTGTAAAGATTGATAATTTGTTTTACATTACCTATAACC
81 Q G M E D P R I V K I D N L F Y I T Y T

301 GCTTACGATGGAATTAATGCTCTTGGAGCCTTGGCTACTTCTCCGGATTTGAAAACATGG
101 A Y D G I N A L G A L A T S P D L K T W

361 AAAAAATAGGAATTATTGTTCCCCAGATTACTTTTGAAGAATTTAAACTCTTCTCAGAA
121 K K I G I I V P Q I T F E E F K L F S E

421 GACAAAAGTAATCTTAACGAAAAGTATATTCGTTATAATCAATTCAGATCAGTCATCAG
141 D K S N L N E K Y I R Y N Q F Q I S H Q

481 ACTGATACGAGTAGCGTTTATTTATGGAGCAAAAACCTTGTTTTTTTTTCCCAGAAGAATT
161 T D T S S V Y L W S K N L V F F P R R I

541 AATGGAACCTCTATTTTCTCCACAGAATAAGACCGGATATTCAGATGGTAACAGGTATT
181 N G N L Y F L H R I R P D I Q M V T G I

601 AAAAATCTAAAAGCATTGACTCTGGATTTTTTGAAAAGATTACTTTCTTCATTTCAAAGAC
201 K N L K A L T L D F W K D Y F L H F K D

661 CACATTGTA CTCTCCAATATACGCCCATGAAAATAAGCTATATAGGAAGCGGCTGCCCT
221 H I V L S P I Y A H E I S Y I G S G C P

721 CCTATAGAAACTCCTGAAGGCTGGCTTATTATTTACCACGGTGTATATGATTCCATAGAA
241 P I E T P E G W L I I Y H G V Y D S I E

781 GGCTATGTATATACAGCATGTGCTGCCTTGCTAGATTTAGAAAATCCAAAGAAAGAAATT
261 G Y V Y T A C A A L L D L E N P K K E I

841 GCAAGACTTCCCTACCCCTCTTTCAACCGGAAAAAATGGGAACTAAAGGGAGAAGTC
281 A R L P Y P L F Q P E K K W E L K G E V

901 AACAAATGTTTGCTTTCCAACAGGAGCTCTGGTAGAAAATGATATCCTCTATATCTACTAC
301 N N V C F P T G A L V E N D I L Y I Y Y

961 GGAGCAGCAGACCAGCGCATTGCAGTTGTCTCTGTAAACTATCAGAGCTCATAAAAGAA
321 G A A D Q R I A V V S V K L S E L I K E

1021 CTCCTGCAATATTCTTCATTATAA
341 L L Q Y S S L *

