Metagenome sequencing and *in silico* gene discovery: From genetic potential to function

by

Dominique Elizabeth Anderson

A thesis submitted in fulfilment of the requirements for the degree of
DOCTOR OF PHILOSOPHY

Department of Biotechnology,
University of the Western Cape
Bellville

Supervisor: Professor D. A. Cowan
Co-supervisor: Professor M. I. Tuffin and Dr. M. P. Taylor

August 2012

Index

ACKNOWLEDGEMENTS

This thesis is dedicated to my mom and dad, who have supported me in all my endeavours and encouraged me to reach my goals with enthusiasm and vigour. I would like to thank Dr. Sam Easton for valuable comments and corrections, as well as for being the best friend anyone could ask for. I thank you for being there, through thick and thin, with all of life's challenges and joys. To my fiancé, F. v. d. Berg, thank you for understanding how important this PhD is to me, for dealing with late nights, weekends in the lab and good (and bad) days. To Dr. M. P. Taylor and Dr. M. I. Tuffin, thank you for reviewing this thesis, excellent supervision and allowing me freedom to explore different aspects in this project. In addition, thank you for your words of encouragement and motivation when things did not go according to plan. To all my family and friends, you have made me who I am and have contributed to my life in a great way. To students of IMBM, past and present, you have made every day memorable and to those of you that I have supervised, you have made me so proud.

To my principle supervisor, Professor D. A. Cowan, I thank you for giving me this opportunity and inviting me to join you in Antarctica. I will never forget the adventure. Thank you for having faith in me and my abilities as a scientist and I wish you all the best for your future endeavours.

Finally, I would like to thank the NRF and SANAP for funding, Professor N. Birkeland for your hospitality in Bergen and all involved in the K021 2011 Antarctic expedition for a once-in-a-lifetime experience.

ABSTRACT

Metagenome sequencing and *in silico* gene discovery:

From genetic potential to function.

Dominique E Anderson

PhD thesis, Department of Biotechnology,

University of the Western Cape

In a previous study, metagenomic DNA extracted from Antarctic Dry Valley soils was used to construct a large contig bacterial shotgun fosmid library (Anderson, 2008). In the current study, clones were selected based on a functional screen for putative lipolytic enzymes, which incorporated tributyrin in agar screening plates. Clones were subsequently subjected to next-generation sequencing and bioinformatic analysis, which allowed for further investigation of a portion of the Antarctic metagenome. Assembly and annotation of the genetic data encoded on three fosmid clones allowed for the identification of the genes responsible for tributyrin hydrolysis. Furthermore, hypotheses relating to survival and adaptation to abiotic conditions prevalent in the extreme Antarctic environment were developed (Chapter 3). A cold adapted esterase was subsequently characterised and showed substrate preference for *para*-nitrophenyl propionate. The optimum temperature and pH for the enzyme, *DEase*I was 25 ° C and 8.5, respectively. In addition, results indicated that *DEase*I was sensitive to thermal inactivation (Chapter 4). Furthermore, in fosmid clone LD13, one particular ORF annotated as a Water HYpersensitity response protein, became the focus of further study. When sub-cloned into a heterologous host, both ionic and osmotic stress tolerance was observed *in vivo*. The protein also exhibited a cryoprotective function *in vitro*, preventing cold denaturation of malate dehydrogenase during cycles of freeze-thaw (Chapter 5). This study demonstrates the value of combinatorial *in silico* and '-omic' based techniques for the discovery and functional characterisation of potentially novel genes from bacteria which inhabit Antarctic Dry Valley soils.

Keywords: Antarctica, metagenomics, next-generation sequencing, bioinformatic analysis cold adaptation, lipolytic enzymes, Water HYpersensitity response protein.

DECLARATION

I declare that *Metagenome sequencing and in silico gene discovery: From genetic potential to function* is my own 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.

Dominique Elizabeth Anderson 28 March 2012

List of Figures

Chapter 3. Metagenome sequencing and *in silico* gene discovery

Chapter 4. From genetic potential to function-Lipolytic genes

Chapter 5. From genetic potential to function-Water HYpersensitivity response gene

List of Tables

Chapter 2. General materials and methods

Chapter 3. Metagenome sequencing and *in silico* gene discovery

Chapter 4. From genetic potential to function-Lipolytic genes

Chapter 5. From genetic potential to function-Water HYpersensitivity response gene

List of Abbreviations

AFL	*Archeoglobus fulgidus* Lipase
AFPs	Antifreeze proteins
Ai	Aliphatic index
APS	Ammonium persulphate
ATP	Adenosine triphosphate
AU	Absorbance units
BOD	Biological oxygen demand
BR	Broad range
BSA	Bovine serum albumen
× g	Centrifugal force
CAPS	3-Cyclohexylamino-1-propanesulfonic acid
CAPs	Cold accumulatory proteins
cfu	Colony forming units
COG	Clusters of orthologous groups
CRT	Cyclic reversible terminator
CSPs	Cold shock proteins
C-terminus	Carboxy-terminus
DNA	Deoxyribonucleic acid
dNTP	Deoxyribonucleotides
EDTA	Ethylenediamine tetra-acetic acid
et al	et alia (and others)
EtBr	Ethidium bromide
EtOH	Ethanol
FPLC	Fast protein liquid chromatography
Gbp	Giga Basepairs
GO	Gene ontology
GRAVY	Grand Average of Hydropathy
GSP	General secretory pathway
HMW	High molecular weight

HPLC	High-performance liquid chromatography
HSL	Hormone sensitive lipase
IPTG	Isopropyl-b-D-thiogalactopyranoside
kDa	Kilo Dalton
kV	Kilovolts
LB	Luria Bertani
LBA	Luria Bertani agar
Mbp	Mega basepairs
µF	Micro Farad
MCS	Multiple cloning site
MPE	Metallophosphoesterase
NGS	Next generation sequencing
N-termius	Amino-terminus
Ω	Ohm
OD	Optical density
ORF	Open reading frame
p-NP	para-nitrophenyl
ppGpp	guanosine tetraphosphate
RT	Reverse transcriptase
SDS	Sodium dodecyl sulphate
sec	Second
SNP’s	Single nucleotide polymorphisms
TAE	Tris acetic acid
TCA	Trichloracetic acid
TEMED	N,N,N',N'-Tetramethylethylenediamine
Tris	Tris-hydroxymethyl-aminomethane
U	Units
UHQ	Ultra high quality
UTR	Untranslated region
WGS	Whole genome sequencing

Chapter 1: General Introduction

1.1 Antarctica

The Antarctic continent harbours unique and diverse terrestrial and aquatic habitats and, while most of this remote continent is covered by an expansive ice sheet, 0.3 % of the land mass is ice-free (Balks and Campbell, 2001). These so-called Dry Valleys consist of exposed soils, glaciers, streams and lakes (both freshwater and saline) and permanently ice-covered lakes and none of these environments are homogenous (Cowan *et al*., 2004). Climatic conditions in this environment vary greatly and directly impact the diversity of life forms that inhabit the various niches (Wynn-Williams, 1996; Hogg *et al*., 2006). The atmosphere contains very low levels of water vapour due to the cold temperatures, which range from – 40 °C in winter to 0 °C in summer. A lack of precipitation in the Dry Valleys ultimately transforms them into cold deserts (Balks and Campbell, 2001). Dry Valley soils accumulate salts which, in combination with low buffering capacity, cause high salinity and fluctuations in soil pH (Balks and Campbell, 2001). Other abiotic conditions which strongly influence the physical, biological, ecological and chemical properties of the soil and are limiting factors for microbial populations that inhabit the Valley soils include low organic accumulation of carbon and nitrogen, low-humidity, excessive solar radiation and strong kabatic winds (Balks and Campbell, 2001; Aislabie *et al*., 2006). The remains of immature crabeater seal carcasses influence the ecosystem dynamics in the Dry Valley soils by contributing a pool of organic nutrients to microbial communities directly beneath, or in close proximity to them. In addition, and perhaps more importantly, these carcasses also provide protection to soil microbes from desiccating winds and ultraviolet exposure (Barwick and Balham, 1967; Dort, 1982; Tiao *et al*., 2011). Studying the unique biota of this environment may provide clues as to how these organisms have adapted to survive extreme abiotic conditions.

1.2 Cold adaptation

The ability of microorganisms to adapt to natural stress factors has made them Earth's most successful colonisers. Approximately 80 % of our planets biosphere is permanently cold, making low temperature the most wide-spread natural stress condition (Hébraud and Poitier, 1999; Rodrigues and Tiedje, 2008; Russell, 1990). Low temperature environments have been successfully colonised by a number of organisms, making psychrophiles the most abundant extremophiles in terms of distribution, biomass and diversity (Piette *et al*., 2011). In addition, microbes inhabiting these niche environments generally also encounter more than one stress factor, such as desiccation, high or low pH, high osmotic pressure and low nutrient availability (Morgan-Kiss *et al*., 2006; Tehei and Zaccai, 2005). In cold environments, the physical properties of water changes and, coupled with cold-stress modifies all physico-chemical parameters of a living cell. It influences cell integrity, solute diffusion rates, membrane fluidity, enzyme kinetics and macromolecule interactions and therefore, the capacity of an organism to compete in that environment (Rodrigues and Tiedje, 2008; Gualerzi *et al*., 2003). The ability of an organism to survive and grow in cold conditions is dependent on a number of adaptive strategies and the modification of pre-existing biosynthetic pathways, in order to maintain vital cellular functions at cold temperatures (Rodrigues and Tiedje, 2008)

In cold environments, membrane fluidity is essential for the transport of substrates and nutrients (Deming, 2002). Alteration of lipid content in cellular membranes is a common strategy employed by organisms (Ray *et al*., 1998) and the rate at which these changes occur is of great importance in habitats where thermal fluctuations occur (Hébraud and Poitier, 1999). In general, a decrease in temperature is accompanied by an increased ratio of polyunsaturated fatty acids which reduces the phospholipid melting point and the rigidity of membrane structures (Nichols *et al*., 1993; Ulusa and Tezcan, 2001). Further alterations in

membrane composition include changes in the size and charge of lipid head groups which affects glycerophospholipid packing, changes in the isomerisation of straight chain fatty acids to cyclic and / or branched isomer and conversion of trans- unsaturated fatty acids to cis-isomers (Chintalapati *et al*., 2004). Furthermore, the post-biosynthetic transformation of saturated fatty acids into unsaturated derivatives by acyl lipid desaturase, is a common mechanism employed by Cyanobacteria for the maintenance of membrane fluidity under cold conditions (Chintalapati *et al*., 2004). Other possible modulators of membrane fluidity have been proposed, particularly carotenoid pigment molecules and sensory proteins that are associated with cell membranes. The sensory proteins which span the length of the membrane generally act as phospo-relay systems for the sensing of temperature changes while carotenoid pigments may buffer membrane fluidity and maintain homeoviscosity during temperature fluctuations (Chattopadhyay, 2006; Ray *et al*., 1998; Rodrigues and Tiedjie, 2008). To combat the low diffusion rates which occur at low temperatures, transport systems for nutrients, substrates and compatible solutes are required and as such, ABC-type membrane transporters are found to be upregulated in proteomic studies (Bakermans *et al*., 2007; Cacace *et al*., 2010).

In cold conditions, basic functions such as transcription, translation and ribosome assembly are hindered by the formation of stable secondary structures in nucleic acids (Gualerzi *et al*., 2003). The cold shock response is induced when an organism is subjected to sub-optimal growth temperatures and involves a number of cold inducible proteins which are preferentially expressed at low temperatures and can be further characterised into Cold shock proteins (CSPs) or Cold acclimation proteins (CAPs) (Cacace *et al*., 2010; Bakermans *et al*., 2007). The gene products may either be directly or indirectly involved in protein transcription and translation (Horn *et al*., 2007). CspA and its homologs are the major proteins involved in this response (Ray *et al*., 1998). Under cold stress the mRNA encoding CspA is stabilised and

its expression is favoured. CspA can up-regulate its own transcription as well as that of other CSPs, by binding to the 5' UTR of Csp mRNA's (Horn *et al*., 2007). This enhances the half-life of RNA and reduces the degradation of mRNA by RNase by decreasing secondary structure formation in transcribed mRNAs. Other proteins associated with transcription and post-transcriptional events include NusA and polynucleotide phosphorylase (Ray *et al*., 1998; Rodrigues and Tiedje, 2008).

To combat the increased negative supercoiling of DNA which occurs at lower temperatures, genes encoding nucleoid associating proteins which are involved in maintaining functional topology of DNA are required for cell survival at low temps (Rodrigues and Tiedje, 2008; Ray *et al*., 1998). These include gyrase A, HU-beta, and H-NS. To cope with transcript stabilisation and degradation, DEAD-box helicase enzymes accumulate in the cell and assemble into degradosomes, with RNase E. Another essential role for the DeaD proteins is in 50 S ribosomal subunit assembly (Rodrigues and Tiedje, 2008; Gualerzi *et al*., 2003). Additional proteins involved in ribosome biogenesis and function include; ribosome binding factors, initiation factors, proteins of the cold shock family, and RNA chaperones. Correct folding of proteins requires the presence of chaperones such as trigger factor and DnaK (Gualerzi *et al*., 2003).

Considering that temperature is one of the most important environmental factors governing biochemical reactions, enzymes need to be suitably adapted in order to perform their catalytic activity (D'Amico *et al*., 2002). According to the Arrhenius equation, this reduction in enzyme reaction rate is correlated to temperature decrease in an exponential fashion (D'Amico *et al*., 2002). Psychrophilic enzymes display increased protein plasticity coupled to weak thermostabilty and higher specific activity at lower temperatures. The structural flexibility of cold-adapted enzymes allows for efficient substrate interaction and reduces the

amount of energy required for generation of intermediate products of catalysis which, in turn, increases substrate turnover (Rodrigues and Tiedje, 2008). A combination of several features contributes to the flexibility of these cold-adapted enzymes and includes; an increase in solvent-exposed hydrophobic side chains, a decrease in hydrophobic residues in the enzyme core, decreased amounts of aromatic-aromatic interactions, increase in glycine (increases thermal agitation) and lysine, decrease in proline and arginine (which has higher hydrogen bond potential), decreased arginine: lysine ratios as well as decreased isoleucine content when compared to mesophilic and thermophilic homologs. In addition, proper solvation at lower temperatures is ensured by an overall increase in the occurrence of charged residues and decrease in salt bridges (Rodrigues and Tiedje, 2008; D'Amico *et al*., 2002; Gerday *et al*., 2000; Cavicchioli *et al*., 2002; Nichols *et al*., 1999; Ray *et al*., 1998; Russell, 2000). Different strategies of structural adaptation may be adopted by different enzyme families and may be unique to each enzyme (Gerday *et al*., 2000; Gianese *et al*., 2001). For example, in a comparison of structures of 21 psychrophilic enzymes belonging to different families, significant substitution of proline residues was only observed for the α-amylase family (Gianese *et al*., 2001). Considering that other ecological and physico-chemical parameters are involved with protein structure and modification, it is important that all the characteristics of an environment be taken into account when assessing adaptive strategies utilised by microorganisms (D'Amico *et al*., 2002).

Other mechanisms of adaptation employed by some microorganisms during cold stress include slower overall growth rates, reduction or inhibition of cell division and long life cycles (Peck *et al*., 2005; Ulusa and Tezcan, 2001). The formation of dormant cell types which continue to respire and utilise substrates, is also a possible survival strategy employed by bacteria under adverse conditions (Chattopadhyay, 2006). The production of antifreeze

proteins (AFPs) in Antarctic fish species such as *Trematomus bernacchi* has been well documented (De Vries *et al*., 1970). Antifreeze glycoproteins bind to water molecules and lower the freezing temperature, thereby preventing the formation of ice-crystals (Ulusa and Tezcan, 2001). Uptake or production of cryoprotectants such as glycine betaine in bacteria is thought to prevent protein aggregation that is induced by cold stress (Chattopadhyay, 2006). Genes for the synthesis and degradation of polyesters and polyamides are evident in the genome of *Colwiella psychrerythaea* and may serve as intracellular carbon and nitrogen reserves (Methé *et al*., 2005). Prolonged exposure to extreme cold conditions may restrict the uptake of these molecules and the reserves can therefore ensure a constant supply of carbon and nitrogen. Similarly, a proteomic study of the psychrotolerant microbial pathogen *Listeria monocytogenes*, grown at 37 °C and 4 °C revealed ten proteins which were present only in the 4 °C 2DE maps. These included not only heat shock proteins DnaK and GrpE, but also proteins involved in metabolism such as mutase and aldolase as well as a number of hypothetical proteins (Cacace *et al*., 2010). Cold shock also affects the process of cell division as evident in a study by Duplantis *et al* (2010). In this study, mesophilic bacterial pathogens were rendered temperature sensitive by the substitution of several genes with homologs from psychrophilic microorganisms and one of the genes under investigation was the cell division protein, FtsZ. In three of the most cold-tolerant *Shewanella* species, an operon encoding the subunits for a Na^+/H^+ antiporter, were observed. It is hypothesised that this pump may be involved in both cold and salt tolerance in this microorganism (Karpinets *et al*., 2009). It appears that a lack of common features indicates that diverse strategies are employed for adaptation to the cold and these mechanisms appear to be constrained by species specific cellular structure and organisation (Piette *et al*., 2011). This topic is discussed extensively (as it relates to this study) in Chapter 3, section 3.3.3

1.3 Metagenomics

Currently, it is estimated that the total number of prokaryotic cells on our planet comprises over 10^8 separate genospecies (Singh *et al*., 2009; Amann *et al*., 1995). In soils, the presence of more than 10^9 bacteria per gram supports an approximate biomass of 3000 kg per hectare (Ranjard and Richaume, 2001). Population diversity and heterogeneity in soils is integral to ecosystem function with soils acting as reservoirs for metabolic and phylogenetic biodiversity (Hunter-Cevera, 1998). However, an estimated 99 % of microbes have remained recalcitrant to culturing mainly due to strict physicochemical requirements, and interdependence with other organisms (Lorenz and Schleper, 2002; Amann *et al*., 1995) In order to study organisms that cannot be maintained in pure culture, a metagenomic approach can be used. Metagenomics is a DNA-based, culture independent approach and focusses on the entire genetic complement of microbes in a habitat or niche (Cowan *et al*., 2004; Schmeisser *et al*., 2007). The two most common strategies for the screening of metagenomic libraries include homology based screening, which requires sequence data in order to target genes, and activity-based screening, whereby clones are functionally selected. There are pro's and con's to both strategies, but both have the potential for isolating genes and/ or gene products of interest (Lorenz and Schleper, 2002; Daniel, 2005; Ferrer *et al*., 2005). Initially, most metagenomic research endeavours were driven by bioprospecting of unknown and improved gene products with the aim of exploitation for biotechnological and biomedical purposes (Cowan *et al*., 2004), and the diversity of characterised, metagenome-derived enzymes have been extensively reviewed by Steele *et al* (2009) and Tuffin *et al* (2009). Recent trends in this rapidly developing research area are aimed at investigating microbial communities, and sequence-based screening methods which employ high-throughput next generation sequencing technologies are a powerful tool for the analysis of large genomic datasets, allowing for the identification of gene function and putative assignment of organismal roles

in communities (Simon and Daniel, 2011). Novel and improved methodologies have driven demand for the development of new software and bioinformatics tools and have found applications in comparative genomics, metatranscriptomics, metaproteomics and metabolomics (Chistoserdova, 2010). In combination, the '-omic' based approaches provide powerful insight into the genetic potential of the microbial world and as these technologies develop and improve, researchers can begin to define complex environmentally dependant pathways of growth, survival and adaptation.

Objectives

In order to investigate the capacity of *in silico* bioprospecting of established large- insert metagenomic libraries, several clones were selected for next-generation sequencing and analysis. Using this approach, it is hypothesised that data mining of cloned DNA fragments may provide fundamental information relating to microbial survival and adaptation to extreme abiotic conditions (Chapter 3). Sequencing data also provides a platform to investigate the functionality of genes encoding for industrially relevant metabolic enzymes (Chapter 4). Furthermore, the wealth of data also allows for prediction of novel genes, which may lead to the discovery of new functions (Chapter 5). This study aims to demonstrate the value of combining computational data mining with experimental strategies to add new knowledge to the field of metagenomics and microbial adaptation.

Aims (Chapter 3)

1. Assemble and annotate full fosmid sequences using both manual and automated approaches
2. Describe genetic potential contained within the sequences
3. Discuss key genes possibly linked to adaptation to environmental conditions experienced in Antarctic Dry Valley soils

Aims (Chapter 4)

1. Identify possible lipolytic genes from fully assembled fosmid clones
2. Clone candidate genes, verify their respective enzyme activities and confirm the accuracy of bioinformatic predictions
3. Overexpress and kinetically characterise lipolytic enzymes
4. Perform comparative genomic study to identify possible sequence and structure modifications conferring cold-adaptation

Aims (Chapter 5)

1. Perform bioinformatic analysis on 13ORF6 (dWHy1)
2. Sub-clone, express and purify dWHy1 for *in vivo* and *in vitro* assays in order to experimentally validate the putative role of dWHy1 in desiccation tolerance.

Chapter 2: General materials and methods.

2.1 Materials

2.1.1 Chemical reagents

All chemical reagents used in this study were of analytical grade and obtained from a range of suppliers.

2.1.2 Antibiotics

Supplementation with the appropriate filter sterilised antibiotic was performed aseptically after the autoclaved media was cooled to ~45 °C. Final concentrations of antibiotics were: (unless otherwise stated) chloramphenicol (Cam), 34 µg/ml; carbenicillin (Carb), 50 µg/ml; ampicillin (Amp), 150 µg/ml; and kanamycin (Kan), 30 µg/ml. In the case where fosmid containing cells were cultured, the final concentration of chloramphenicol was 12.5 µg/ml.

2.1.3 Enzymes

The various enzymes used for both DNA manipulations as well as enzymatic assay, along with the supplier are shown in table 2.1.1.

Table 2.1.1 Enzymes used in this study

Enzyme	Function	Supplier
Restriction endonucleases	Restriction enzyme digestion	Fermentas
Alkaline phosphatase	Dephosphorylation	Fermentas
Klenow DNA polymerase	Blunt end generation	Fermentas
T4 DNA Ligase	Ligation	Fermentas
High fidelity polymerase	PCR	Various
DreamTaq™ DNA Polymerase	PCR	Fermentas
Malate dehydrogenase	Enzyme assays	Sigma

2.1.4 Strains, vectors and primers

A list of strains, vectors and primers utilised in this study are shown in table 2.1.2

2.2 General microbiological techniques

2.2.1 Media

Luria-Bertani (LB) broth consisted of 1 % [w/v] tryptone, 0.5 % [w/v] yeast extract and 1 % [w/v] NaCl. Luria-Bertani (LB) agar consisted of LB broth prepared as above with the addition of 1.3 % [w/v] bacteriological agar. Tributyrin agar consisted of 1 % [w/v] tryptone, 0.5 % [w/v] yeast extract, 1 % [w/v] NaCl, 1.3 % [w/v] bacteriological agar, 1 % [v/v] tributyrin, 1 % [w/v] gum arabic. SOB broth consisted of 2 % [w/v] tryptone, 0.5 % [w/v] yeast extract, 0.05 % [w/v] NaCl, 0.02 % [w/v] KCl. All components were mixed together with distilled water and the pH was adjusted to 7.0 using 1 M NaOH. Media was autoclaved at 121 °C for 20 minutes. Sodium chloride agar was prepared from LB agar with the addition of 3.5 % [w/v] NaCl. Mannitol agar was prepared from LB agar with a reduced NaCl concentration (0.5 % [w/v]) and 21 % [w/v] D-Mannitol. SOC was prepared from SOB with the addition of filter sterilised 0.5 % [w/v] 2 M $MgCl_2$ and 2 % [w/v] 1 M glucose.

2.2.2 Growth of *E. coli* strains

Bacterial strains were grown in broth or on solid media supplemented with the appropriate antibiotic. The native *E. coli* strains; BL21 and GeneHog were grown on media with no antibiotic. Strains were inoculated using aseptic technique. Unless otherwise stated, cultures were incubated at 37 °C. If strains were grown in broth, incubation was accompanied by agitation at 150 to 225 rpm.

Table 2.1.2 Strains, vectors and primers used in this study.

	Characteristics	Source
Bacterial strains	Genotype	
E. coli		
EPI-300	F- mcrA D(*mrr-hsd*RMS-*mcr*BC) f80d*lac*ZDM15 D*lac*X74 *rec*A1 *end*A1 *ara*D139 D(*ara, leu*)7697 *gal*U *gal*K l- *rps*L *nup*G *trf*A *ton*A *dhfr*	Epicentre Biotechnology (USA)
GeneHog	F- *mcr*A Δ(*mrr-hsd*RMS-*mcr*BC) φ80*lac*ZΔM15 Δ*lac*X74 *rec*A1 *ara*D139 Δ(*ara-leu*)7697 *gal*U *gal*K *rps*L (StrR) *end*A1 *nup*G *fhu*A::IS2	Invitrogen (USA)
Rosetta(DE3)pLysS	F^- *ompT hsdS*$_B$(r_B^-, m_B^-) *gal dcm* (DE3) pLysSRARE (Cam^R)	Novagen (USA)
BL21(DE3)	F-, *omp*T, *hsd*S_B (r_B–, m_B–), *dcm*, *gal*, λ(DE3)	Invitrogen (USA)
Plasmids/vectors		
pCCFos1	$Chloramphenicol^R$ 12.5 µg/ml	Novagen (USA)
pUC19	$Ampicillin^R$ 100 µg/ml	Novagen (USA)
pET21a	$Chloramphenicol^R$ 34 µg/ml	Novagen (USA)
pGemT-easy	None	Fermentas (SA)
pRareMod7	$Kanamycin^R$ 50 µg/ml	This study
pET28a	Kanamycin ^{R} 50 µg/ml	Novagen (USA)
pCOLD	$Ampicillin^R$ 100 µg/ml	Takara (Japan)
Primers *		
Primer walking		
LD4C7R	CCATACAACAAACTGGTCAACA	This study
LD4C7F	GCCATTTAGACAAGTTCATCAC	This study
LD4N187ds	GATAGCGACTACAACTGGCAAGCCG	This study
LD4N212ds	CCACGGGACTGGCAAAAGATGCATTGC	This study
LD7C5F	GCCGAGCCCGAATTCATCG	This study
LD7N14us	CGGCAGCAATCATTCCCAGC	This study
LD7N118ds	GAGCGGGTCTCATCAGGTCAG	This study
LD7C3F	CGATCAAGGACGGTTCAAGCGTG	This study
Transposon mutagenesis		
MUKAN-1 FP-1	CTGGTCCACCTACAACAAAGG	Epicentre Biotechnology
MUKAN-1 RP-1	AGAGATTTTGAGACAGGATCCG	Epicentre Biotechnology
General sequencing		
T7 forward	TAATACGACTCACTATAGGG	
T7 reverse	GCTAGTTATTGCTCAGCGG	
pCC1 reverse	CTCGTATGTTGTGTGGAATTGTGAGC	Epicentre Biotechnology
Sub-cloning		
		This study
CL1-R21	ACCTCGAGGAATGCCTCGCGCAGCGAC	This study
CL1-F21	TGCATATGGGATCAATGCCGCTACGGACACG	

CL2-F21	ACCTCGAGCAGGACAGGCTCAGGCTGG	This study
CL2-R21	TGCATATGCCTGTCCTACCAATCCCATCCGTG	This study
DEA4-R-N	TGGCGCATATGTGGTTTTTTATTATTATTAC	This study
DEA4-F-X	ACCTCGAGTTATTCTGCTTGTTGTAAAG	This study
DA5-F21	ACCTCGAGCAAACACTTCTTCAGCATGGC	This study
DEA5-R-N	TGCATATGCATATTGATGGCAGCAGTG	This study
LEA-F21	CGTGAATTCATGAGCTATTTAGCAACTATA	This study
LEA-R21	CTACTCGAGCTCGCGAATATAGTCG	This study
W17R	CTGGAATTCTTACTCGCGAATATAGTCGC	This study
W17F	GACGCTAGCATGAGCTATTTAGCAACTATAAA	This study

*All primer sequences are given in the 5' to 3' orientation.

2.3 General molecular biology techniques

2.3.1 Plasmid DNA extraction- Alkaline lysis

Selected colonies were picked from agar plates and inoculated into 10 ml LB broth supplemented with the appropriate antibiotic. Cultures were incubated overnight at 37°C with agitation. After incubation, cells were collected in by centrifugation at 10000 × g for 5 minutes at 4 °C. The supernatant was discarded and excess media removed. Cells were re-suspended in 1 ml ice cold GET buffer (50 mM glucose, 10 mM EDTA, 25 mM Tris-HCl) and 24 µl of RNaseA (10 mg/ml) was added. After incubation at room temperature for 5 minutes, 1 ml of lysis solution (0.2 M NaOH, 1% [w/v] SDS) was added and the tubes were gently inverted to mix. Following the addition of 1 ml of 3 M KOAc (pH 5.5), the tubes were inverted again to mix and cells were incubated on ice for 5 minutes. The tubes were inverted again and incubated on ice for a further 10 minutes. After centrifugation at 10 000 × g at 4 °C, the plasmid DNA was precipitated by the addition of 0.7 volumes of isopropanol to the recovered supernatant in a sterile tube followed by incubation for either 3 hours or overnight at room temperature. Tubes were centrifuged to collect the plasmid DNA and the resultant pellets were washed twice in 70 % [v/v] ethanol in order to remove residual salt. After air drying, the DNA pellets were resuspended in UHQ Millipore water and aliquots were

analysed by agarose gel electrophoresis (section 2.3.5). The method was scaled up depending on the amount of plasmid DNA required and adjustments in solution volumes were made accordingly.

2.3.2 Fosmid DNA extraction

Selected clones were inoculated into 5 ml LB broth supplemented with chloramphenicol and incubated overnight. One millilitre of the culture was inoculated into a tube containing 9 ml LB broth and 10 µl of filter sterilised 1 % [w/v] arabinose and grown for 5 hours at 37 °C with agitation. Tubes were centrifuged at 4000 × *g* for 30 minutes at 4 °C. Fosmid DNA was obtained using the alkaline lysis protocol for plasmid DNA extraction described in section 2.3.1.

2.3.3 Sequencing

Fosmids were extracted from 6 selected clones (section 2.3.2). DNA was quantified by fluorimetry (section 2.3.6). The fosmids were pooled together and sent for Solexa sequencing at the University of the Western Cape. The fosmids were also sent with the T7–promoter primer and pCCfos1 primer to the University of Stellenbosch sequencing facility for end-sequencing using the ABI PRISM 377 automated DNA sequencer. In the case where the orientation and sequence of sub-cloned genes needed to be verified, plasmids were extracted using alkaline lysis and sent to the same sequencing facility along with the T7-promoter and terminator primers.

2.3.4 Restriction enzyme digestion

Restriction enzyme digestions were performed in sterile eppendorf tubes in small reaction volumes (10–50 µl). The reactions contained the appropriate volume of 1 × or 2 × buffer (supplied by the manufacturer for the specific enzyme) and 5-10 U of enzyme per µg of plasmid or genomic DNA. Reactions were incubated for either a 2 hour period, or overnight

at 37 °C. The digestion products were analysed by gel electrophoresis on 0.7 % or 1 % [w/v] agarose gels (section 2.3.5).

2.3.5 Agarose gel electrophoresis

Zero-point seven percent or one percent [w/v] agarose was dissolved in 1 × TAE buffer (0.2 % [w/v] Tris base, 0.5 % [v/v] glacial acetic acid, 1 % [v/v] 5 M EDTA [pH 8]). Cast gels were electrophoresed at 30 to 100 V in 1 × TAE buffer. To allow visualisation of the DNA on a UV transilluminator, the gels were supplemented with 0.5 µg/ml ethidium bromide. Samples were mixed with standard loading dye (60 % [v/v] glycerol, 0.25 % [w/v] Orange G) and loaded into the wells of the cast gels. DNA was sized according to its migration in the gel as compared to that of DNA molecular markers used (Lambda DNA restricted with HindIII; Lambda DNA restricted with PstI; 1 kb DNA marker and 100 bp DNA marker).

2.3.6 DNA quantification

Quantification was performed using the Nanodrop ND-1000. The instrument was blanked using 2 µl of the UHQ Millipore water as used for DNA resuspension, or elution. A volume of 1 µl of resuspended or eluted DNA was added to the cleaned reading platform and the DNA concentration recorded. For more accurate DNA quantification, the concentration was measured by fluorometery using the Quanti-iT ™ ds DNA BR assay kit and the Qubit™ system (Invitrogen, Oregon, USA) according to the manufacturers' specifications.

2.3.7 DNA purification

The NucleoSpin® Gel and PCR Clean-up kit (Macherey-Nagel, Germany) was used to purify DNA from solution or agarose, according the manufacturers specifications.

2.3.8 Preparation of competent *E. coli* cells

a. Electrocompetent *E. coli* cells

All glassware was thoroughly washed with 2 % [v/v] SDS, followed by an additional wash with 70 % ethanol, rinsed and autoclaved prior to use A single colony of an *E. coli* strain,

streaked from a glycerol stock onto LB agar and grown overnight, was used to inoculate 10 ml SOB. This starter culture was grown overnight at 37 °C with shaking (225 rpm). One litre of SOB media was inoculated with the starter culture and grown at 37 °C until an OD_{600} of 0.5-0.7 was reached. The cells were kept on ice and 250 ml aliquots were made in chilled Corning pointed bottles.

Bottles were centrifuged at 5000 × *g* for 25 minutes at 4 °C, the supernatant was poured off and the pellet was gently resuspended in 200 ml ice-cold demineralised water (dH_2O) [Millipore] before another round of centrifugation at 5000 × *g* for 25 minutes at 4 °C. Once the supernatant was removed, the cells were resuspended in 100 ml ice-cold dH_2O and centrifuged at 5000 × *g* for 25 minutes at 4 °C. Bottles were placed on ice, the supernatant was removed and the cell pellet was resuspended in 20 ml ice-cold 10 % [v/v] glycerol and centrifuged at 5000 × *g* for 25 minutes at 4 °C. After the supernatant was removed, each cell pellet was very gently resuspended in 1 ml 15 % [v/v] glycerol and 2 % [w/v] sorbitol. The cell suspension was kept on ice, aliquoted into 1.5 ml eppendorf tubes and stored at -80 °C. One microliter of pUC19 vector DNA (100 ng/µl) was used to test the electro-competency of the cells.

b. Chemically competent *E. coli* cells

All glassware was thoroughly washed with 2 % [v/v] SDS, followed by an additional wash with 70 % ethanol, rinsed and autoclaved prior to use A single colony of an *E. coli* strain, streaked from a glycerol stock onto LB agar and grown overnight, was used to inoculate 10 ml LB broth. This starter culture was grown overnight at 37 °C with shaking (225 rpm). One litre of LB broth was inoculated with the starter culture and grown at 37 °C until an OD_{600} of 0.3- 0.6 was reached. The cells were kept on ice and 250 ml aliquots were made in chilled Corning pointed bottles.

Bottles were centrifuged at 5000 × *g* for 25 minutes at 4 °C, the supernatant was poured off and the pellet was gently resuspended in 200 ml ice-cold demineralised water (dH_2O) [Millipore] before another round of centrifugation at 5000 × *g* for 25 minutes at 4 °C. Once the supernatant was removed, the cells were resuspended in 100 ml ice-cold $MgCl_2$ and kept on ice for 5 minutes before another round of centrifugation at 5000 × *g* for 25 minutes at 4 °C. Bottles were placed on ice, the supernatant was removed and the cell pellet was resuspended in 100 ml ice-cold $CaCl_2$, held on ice for a further 20 minutes and centrifuged at 5000 × *g* for 25 minutes at 4 °C. After the supernatant was removed, each cell pellet was very gently resuspended in 3 ml 85 % [v/v] $CaCl_2$ and 15 % [v/v] glycerol. The cell suspension was kept on ice, aliquoted into 0.6 ml eppendorf tubes and stored at -80 °C. Two microliters of pUC19 vector DNA (100 ng/µl) was used to test the electro-competency of the cells.

2.3.9 Transformation of *E. coli* cells

a. Electroporation

Aliquots of 50 µl of electrocompetent cells were thawed on ice. DNA was added directly to cells and incubated on ice for 5 minutes. The mixture was pipetted into pre-cooled electroporation cuvettes (Bio-Rad Laboratories, CA, USA). Electroporation was performed using the following conditions; 1.8 kV, 25 µF, 200 Ω. Nine hundred and fifty microliters of SOC was immediately added to the cuvette and once mixed, transferred to sterile tubes. The mixture was incubated for a maximum of 3 hours at 37 °C with agitation and aliquots were plated on LB-Agar plates supplemented with the appropriate antibiotic and grown overnight at 37 °C.

b. Heat shock

Plasmid DNA was added directly to 50 µl of competent cells, incubated on ice for 5 minutes and heat-shocked at 45 °C for 30 seconds. Cells were incubated on ice for a further 5 minutes, 250 µl of SOC was added and the cells were incubated for a maximum of 3 hours at

37 °C. The transformation mix was plated on media supplemented with appropriate antibiotic and incubated overnight at 37 °C.

2.3.10 Polymerase Chain reaction using gene specific primes

PCR reactions (20-50µl) typically contained ~ 10 ng of template DNA, 1 x PCR buffer, 0.2 mM of each dNTP, 0.5 pmol of each primer (Table 2.1) and 0.25- 0.5 U of high fidelity *Taq* DNA polymerase. For control purposes, forward primer, reverse primer and a negative control (a reaction mixture containing all reagents except template) was routinely included. Following PCR, an aliquot of each reaction mixture was analysed using gel electrophoresis as described.

2.3.11 Ligation

The PCR products were purified using the Nucleospin ® gel and PCR purification kit (section) and digested with the restriction endonucleases that were included in the primer design (section) One microgram of vector DNA was digested with the same restriction enzymes and the gene and vector were ligated, using a ratio of 1:2 vector to insert. The following reagents were used for the ligation reaction which was performed overnight at 18°C, 1 U of T4 DNA ligase in appropriate ligation buffer. The amount of insert required was calculated by the following equation assuming that 50 ng of vector DNA was used in the ligation reaction.

$$\text{ng insert required} = \frac{\text{ng vector} \times \text{kb insert}}{\text{kb vector}} \times 2$$

Three microliters of the ligation mixture was dialysed on a 0.02 nm nitrocellulose filter (Millipore). One and a half microliters of this mixture was electroporated into competent GeneHog *E. coli* cells. These cells were grown on LB-Agar supplemented with the appropriate antibiotic, overnight at 37 °C. A negative control of circular, uncut plasmid DNA was routinely included.

2.4 General Protein techniques

2.4.1 Protein expression

A single culture of the expression host transformed with a construct was inoculated into 5 ml LB broth supplemented with appropriate antibiotic and grown overnight at 37 °C. The culture was transferred to 50 ml LB broth and grown at 37 °C until an OD_{600} of 0.6 was obtained. To this, 0.4 - 2 mM IPTG was added and cells were incubated according to the optimised temperature as determined for each individual protein. The cultures were centrifuged at 6000 × g for 5 minutes at 4 °C and the supernatant was either discarded or precipitated with 20 % [w/v] trichloroacetic acid. Cell pellets were resuspended in lysis buffer (50 mM Tris pH 7.5, 0.3 M NaCl, 5 % glycerol [v/v]) and subjected to 6 cycles of sonication (20 seconds each). Following centrifugation, the resulting cell pellet and cell free lysate fractions were collected. The supernatant was sterilised through a 0.22 µm filter and used for His-tag purification. All fractions collected were analysed by electrophoresis on 12-15 % SDS-PAGE. Protein expressions with the parental vector as a control were routinely included. Expressions were scaled up to 2 L volumes and adjustments were made accordingly.

2.4.2 SDS-PAGE

Vertical SDS-PAGE gels were cast with a 10 - 15 % separating gel (1.5 M Tris-HCl [pH 8.8], 20 % [w/v] SDS, 30 % [w/v] acrylamide, 0.8 % [w/v] bis-acrylamide, 10 % [w/v] ammonium persulphate [Sigma], 0.1 % [v/v] TEMED [Fluka]) and a 4 % stacking gel (0.5 M Tris-HCl

[pH 6.8], 20 % [w/v] SDS, 30 % [w/v] acrylamide, 0.8 % [w/v] bis-acrylamide, 10 % [w/v] ammonium persulphate, 0.1 % [v/v] TEMED).

Samples were mixed with an equal volume of 2 × loading dye (80 mM Tris-HCl [pH 6.8], 10 % [v/v] mercaptoethanol, 2 % [v/v] SDS, 10 % [v/v] glycerine, bromophenol blue), vortexed and heated to 95 °C for 5-10 minutes. Samples were loaded on gels and electrophoresed at 60 V in 1 × running buffer (0.25 mM Tris-HCl, 2 M glycine, 1 % [w/v] SDS) for 30 minutes through the stacking gel. Electrophoresis continued through the separating gel at 120 V for ~ 2 hours. The gel was stained with coomassie stain (0,125 % [w/v] Coomassie blue R250, 50 % [v/v] methanol, 10 % [v/v] acetic acid) for 45 minutes and de-stained overnight with SDS destain (50 % [v/v] methanol, 10 % [v/v] acetic acid). The sizes of the proteins were determined according to their migration in the gel as compared to that of the protein molecular ladder used.

2.4.3 TCA precipitation of proteins

One third the volume of ice-cold 20 % [w/v] TCA was added to the culture supernatant, incubated overnight at 4 °C and centrifuged at 8000 × *g* for 25 minutes. The pellet was resuspended in sterile demineralised water and analysed by SDS-PAGE.

2.4.4 Histidine-tag chromatography

His-Bind resin was completely resuspended by gentle inversion. Two millilitres of the slurry was transferred to a purification column and packed by gravitational flow to a final bed volume of 1 ml. To charge and equilibrate the column the following sequence of washes was used;

- 2 ml sterile demineralised water
- 2. 5 ml of 1 × charge buffer (8 × = 400 mM NiSO4)
- 3 ml of 1 × binding buffer (8 × = 4 M NaCl, 160 mM Tris-HCl, 40 mM imidazole [pH 7.9])

After draining of the binding buffer, prepared extract was added to the column. The column was washed with 10 ml of 1 × binding buffer, 6 ml of 1 × wash buffer (0.5 M NaCl, 60 mM imidazole, 20 mM Tris-HCl [pH 7.9]), 6 ml 1 × elute buffer (0.5 M NaCl, 1 M imidazole, 20 mM Tris-HCl [pH 7.9]) and finally 6 ml of 1 × strip buffer (0.5 M NaCl, 100 mM EDTA, 20 mM Tris-HCl [pH 7.9]). For optimisation of purification, the imidazole concentration in each buffer was changed for each protein. The column was washed with sterile demineralised water and stored in 1 ml 20 % ethanol at 4°C.

The fractions in which the protein was eluted was transferred to a dialysis cassette and dialysed for a maximum of 2 days against 3 L of dialysis buffer (50 mM Tris-HCl [pH 7.5], 1 % [v/v] glycerol) at 4 °C. 50 ml of the buffer was retained for control purposes. Recovered fractions were stored at 4°C and used in a subsequent enzyme assays or for further FPLC analysis.

2.4.5 Bradford determination of protein concentration

Ten microliters of sample was mixed with 200 µl of Bradford's reagent and 790 µl of sterile dialysis buffer and incubated at room temperature for 20 minutes. Optical density measurements were performed at 595 nm and plotted against a 1-20 µg BSA standard curve (Bradford, 1976).

2.4.6 Enzyme assays using *p*-nitrophenyl esters

Nine hundred and seventy microliters of buffer (0.1 M NaCl, 0.1 M NaH_2PO_4, 1% [v/v] acetonitrile), 10 µl of 50 mM substrate (dissolved in acetonitrile) and 10 µl of 1 % [v/v] Triton-X 100 was pipetted into a 1 ml cuvette, mixed thoroughly by inversion and the absorbance at 405 nm over a period of 3 minutes was measured. This mixture was used as the blank and a new cuvette was used for each blank measurement. After addition of enzyme, the change in absorbance units per minute was measured for each substrate and enzyme tested and substrate specificity was determined. The volume of buffer was adjusted accordingly when increased volumes of enzyme or substrate were used and AU/min was measured. The Bradford assay was used to calculate enzyme concentration and the rate was calculated from the data obtained. The enzymes were also tested for optimal temperature, pH and thermal sensitivity.

Enzyme activity was calculated using the following equation;

$$\mathbf{A= £. C. L}$$

Where, ε is the extinction co-efficient of p-nitrophenol, A is the rate of the enzyme reaction based on *V*max and the volume of enzyme used in the 1 ml assay, L is the path length of light through the cuvette and has a value of 1.

Chapter 3: Metagenome sequencing and *in silico* gene discovery

3.1 Introduction

3.1.1 First generation sequencing

Dideoxynucleotide sequencing of DNA was first published in 1975 by Sanger *et al*. For over three decades, the Sanger chain termination method remained conceptually unchanged and the most commonly used DNA sequencing technique (Mardis, 2007; Morozova and Marra, 2008). The Sanger method relies on the synthesis of a complementary strand of DNA by DNA polymerase in the presence of dNTP's and ddNTP's. When the non-reversible synthesis terminators are incorporated into the growing oligonucleotide chain, synthesis terminates resulting in truncated products of various lengths (Sanger *et al*., 1977). Polyacrylamide gel electrophoresis is used to resolve products based on size, and the 3' ddNTP's reveal the DNA sequence (Hutchinson, 2007; Morozova and Marra, 2008). Initially, four separate reactions were required for synthesis of each template, until methods in fluorescence detection advanced, allowing for labelling of each of the ddNPT's with a different colour fluorescent dye (Smith *et al*., 1986). In 1986, the first report of an automated sequencing method using four different dyes coupled to dNTP's was used (Prober *et al*., 1987; Morozova and Marra, 2008). Improvements continued with the replacement of gel slabs with capillary electrophoresis arrays (Huang *et al*., 1992) and research in polymer chemistry led to the development of polydimethylacrylamide, allowing the re-use of capillaries for multiple runs, thereby increasing the efficiency of sequencing (Madabhushi, 1998). Finally, in 1998, Sanger sequencing was fully automated with the introduction of the ABI Prism 3700 machine.

From humble beginnings, the Sanger and Coulson 'plus and minus' sequencing method has generated a wealth of sequence data starting with the first genetic barcode for the 48.5 kb phage, Lambda (Sanger *et al*., 1982). The first highly ambitious sequencing project was discussed in 1985 and the human genome sequencing project was started in 1990. In 1995,

the first complete sequenced bacterial genome of *H. influenza* not only provided the first genetic blueprint for a living organism but also introduced the concept of the whole genome shot-gun sequencing (WGS) method (Fleischmann *et al*., 1995; Hutchinson, 2007). The shot-gun sequencing process starts with random shearing of genomic DNA, cloning of the resulting fragments and subsequent transformation into *E. coli.* Sequencing of clones is at random and results in a collection of reads which are assembled into a complete genome by computational matching of sequence alignments (Hutchinson, 2007; Pop, 2009). The genome of *Mycobacterium genitalium* was sequenced shortly after (Fraser *et al*., 1995), followed by a rapid increase in the number of bacterial genomes being sequenced and thus, the era of genomics was born (Mardis, 2007; Koonin and Wolf, 2008). The WGS method was used by a number of collaborative groups and the draft of the human genome was obtained in 2001 (Venter *et al*., 2001; Lander *et al*., 2001; Hutchinson, 2007). Other sequencing goals accomplished during this period included genomes of the model organisms, *Escherichia coli* (Blattner *et al*., 1997), *Methanococcus jannaschii* (Bult *et al*., 1996), *Bacillus subtilis* (Kunst *et al*.,1997), *Mus domesticus* and *Drosphilia melanogaster* (Adams *et al*., 2000).

3.1.2 Next generation sequencing (NGS) technologies

The advances in sequencing technology have greatly impacted genomics and genetic experimentation (Mardis, 2007; Metzker, 2010). NGS methodologies are currently being applied in a number of novel applications, previously unexplored by Sanger sequencing, primarily due to the ability of these new techniques to generate and process millions of sequence reads (Mardis, 2007; Morozova and Marra, 2008). These new techniques have been used mainly for standard sequencing and re-sequencing of genomes. Not only do these methods provide a substantial increase in cost effectiveness, but also generate large numbers of sequence reads in a short period of time while reducing the bias associated with conventional vector based cloning and amplification required to produce the large amounts of

DNA needed for first generation technology (Mardis, 2007; Morozova and Marra, 2008; Metzker, 2010). Some available instruments commonly used include the 454 Pyrosequencing instrument (Roche), the Solexa 1G analyser (Illumina, Inc) and the SOLiD system (Applied Biosystems). All three systems may utilise genomic fragments or mate-pair libraries as templates. A comparison of the three technologies is provided in Table 1 (Mardis, 2007; Morozova and Marra, 2008; Metzker, 2010). In this review, these technologies will be discussed briefly with a focus on the Solexa technology.

Table 3.1.1: Comparison of commonly used next generation sequencing systems (2010)

	Illumina	454 Pyrosequencing	SOLiD
Chemistry	Sequence by synthesis	Pyro-sequencing	Sequence by ligation
Application	Solid phase bridge amplification	Emulsion PCR	Emulsion PCR
Read length (bp)	30-100	200-300	35-50
Base pairs per run	1 Gbp	100- 450 Mbp	1-3 Gbp
Run time	4-9 days	3-7 hours	5-14 days
Biological application	Re-sequencing and variant discovery. Gene discovery in metagenomics. Whole exon capture.	*De novo* assembly of bacterial and insect genomes. 16S rRNA metagenomics. Medium exon capture.	Re-sequencing and variant discovery. Gene discovery in metagenomics. Whole exon capture.

3.1.2.1 Support oligonucleotide ligation detection (SOLiD)

This sequencing technology was released in 2007 and utilises a sequencing process catalysed by DNA ligase. DNA fragments are oligo adaptor- linked and coupled with magnetic beads which are decorated with complementary oligonucleotides. Each bead is subsequently amplified by emulsion PCR followed by covalent attachment of the beads to a glass slide (Mardis, 2007). Sequencing results from sequential rounds of hybridisation and ligation, with a mixture of fluorescently labelled probes and sequencing primers, followed by imaging and

probe cleavage (Morozova and Marra, 2008). Analysis of the colour resulting from two successive ligations provides the identity of the nucleotide, and each position is effectively probed twice, thereby improving accuracy (Hutchinson, 2007). Sequencing errors and polymorphisms are readily identified by the two-base encoding scheme (Harismendy *et al*., 2009). A disadvantage of this system is the long run times required for sequence generation (Petterson *et al*., 2009).

3.1.2.2 454 Pyrosequencing

This bioluminescense based method measures the order and intensity of light peaks created by the release of inorganic pyrophosphate and its subsequent enzymatic conversion to visible light, when complementary dNTP's are successfully incorporated into the growing chain by DNA polymerase (Metzker, 2010).

The system depends on emulsion PCR whereby DNA fragments, primer coated beads and other important components required for clonal amplification, are isolated in water micro reactors. Beads amplified in the oil-aqueous emulsion are distributed on PicoTitre plates with one bead per well (Margulies *et al*., 2005). It is important to note that some wells may not contain a bead and even if they do, usable sequence may not be provided (Petterson *et al*., 2009). Ligation of randomly sheared DNA fragments to linker sequences permits capture of individual molecules onto each bead, thereby permitting shot-gun sequencing of whole genomes without prior mate-pair library construction (Hutchinson, 2007).

Read lengths of over 400 bp may be obtained in short time periods using this 'sequence by synthesis' technology, however, the reagent costs for Pyrosequencing is high and increasing levels of noise can occur due to positive and/or negative frameshift events. Additionally, increased errors occur at DNA stretches with single nucleotide repeats (Metzker, 2010; Pop, 2009; Petterson *et al*., 2009).

3.1.2.3 Illumina, Solexa

In 2006, the 'sequence by synthesis' concept allowed for the generation of 32-40 bp sequence reads using the Illumina Genome Analyser (Bennet *et al*., 2005; Mardis, 2007). A microfluidic cluster station is used to add single-stranded adaptor-ligated DNA fragments to the surface of a glass flow cell, each with eight separate lanes to allow for simultaneous runs of independent samples (Mardis, 2007; Metzker, 2010). Hybridisation between the adaptors on the template DNA and complementary adaptors covalently attached to the flow cell surface occurs by a series of heating and cooling steps. Following incubation with reagents, an isothermal polymerase amplifies the fragments in discrete clusters (Mardis, 2007). Solid surface bridge amplification occurs, whereby single molecules attached to the surface bend over and hybridise to complementary adaptors, and clone-free DNA amplification ensues (Morozova and Marra, 2008). Approximately 1000 clonal copies of a single template are contained in each cluster. Reverse terminator nucleotides with chemically inactivated 3' OH ends are incorporated into the growing chain. A single base is incorporated per cycle, the unbound nucleotides are removed, imaging is performed and a subsequent cleavage step removes the fluorescent label thereby restoring activity at the 3' OH and allowing the incorporation of the next base (Mardis, 2007; Morozova and Marra, 2008). Figure 3.1.1 provides an overview of the sequence-by-synthesis technique employed in Illumina Solexa technology.

Following sequencing using this 'cyclic reversible terminator' (CRT) method, clusters are subjected to quality filtering to eliminate low quality reads (Metzker, 2010). One typical four day run can generate 40 – 50 million sequence reads and permits massive parallel sequencing of templates (Mardis, 2007). The Illumina systems tend to be more effective at reading homopolymer stretches than the 454 technology. Due to the generation of very short reads sequence repeats of short length cannot always be resolved. Additionally, substitution errors

have been noted as the most common error type, particularly when a 'G' base is previously incorporated (Dohm *et al*., 2008; Hutchinson, 2007; Metzker, 2010; Harismendy *et al*., 2009). Although limited by read length, quality and accuracy of sequence data is obtained by the high levels of redundancy and coverage (Morozova and Marra, 2008; Metzker, 2010).

Despite these issues, the Solexa genome analyser is said to produce a raw read accuracy of 98.5 % and currently dominates the market due to its extremely high throughput (Metzker, 2010, Petterson *et al*., 2009).

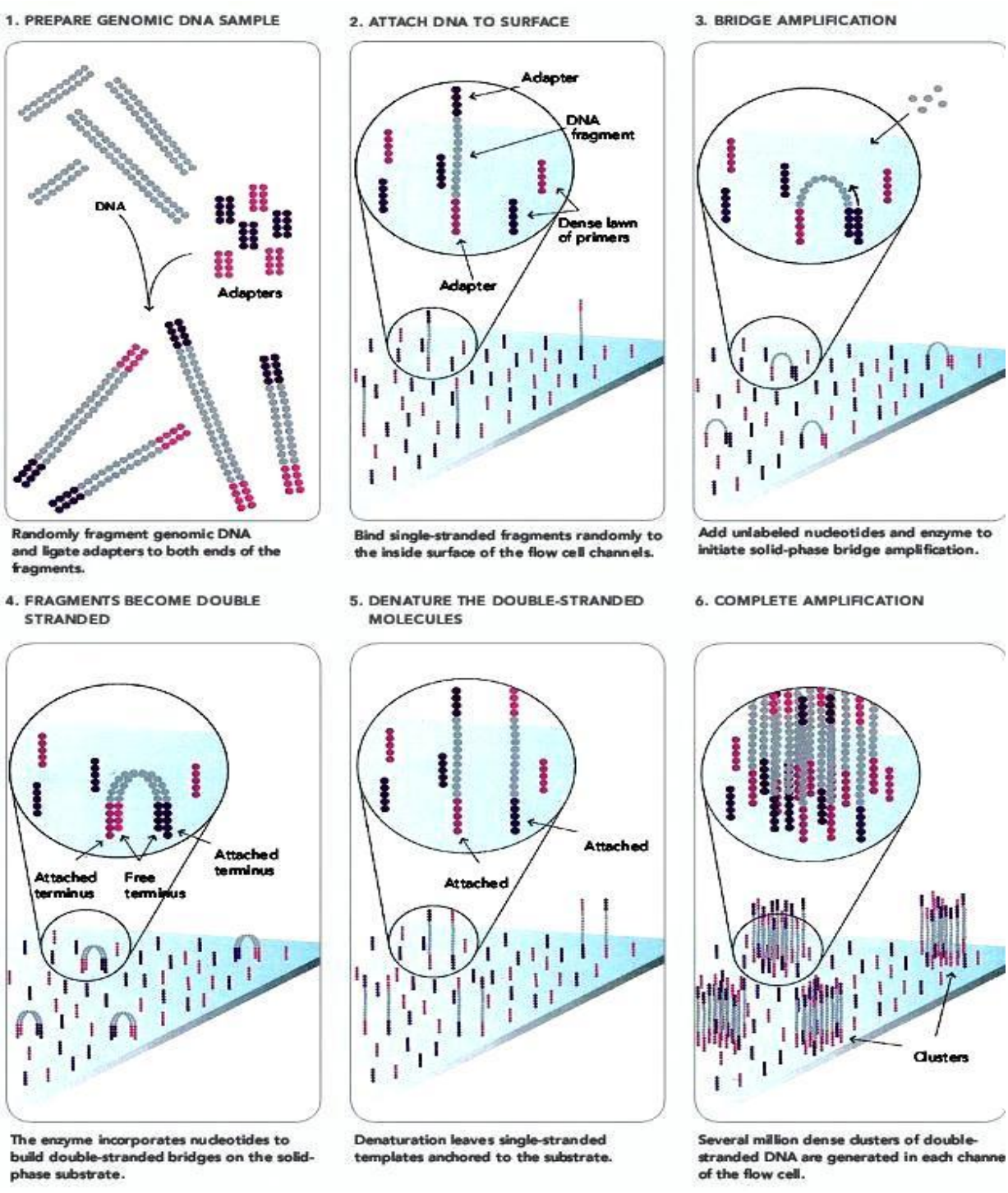


Figure 3.1.1: Overview of the sequence-by-synthesis technique employed in Illumina Solexa technology (Taken from www.illumina.com).

Next generation sequencing technologies have revolutionised many of the biological sciences disciplines by providing a cost effective means to obtain masses of sequence data. No single technology is without some disadvantage and the combination of short read methods with longer read length techniques may provide increased accuracy to aid in assembly of sequence data while remaining cost effective (Hutchinson, 2007). Ultimately, the objectives of each study must be clearly defined in order to employ a particular method of sequencing.

3.1.3 Sequence assembly

Once sequence reads have been generated, the next critical stage in genomic sequencing is the assembly of multiple overlapping segments into contigs (Zerbino and Birney, 2008). This is by no means a trivial task and is complicated by the existence of near identical repeats as well as the number and length of reads (Pop, 2009, Morozova and Marra, 2008). Assemblies are computationally obtained by two methods; *de novo* assembly or comparative assembly (Metzker, 2010).

De novo assembly approaches are used to reconstruct genomes that are dissimilar to known sequences of organisms and may be essential for characterisation of biological diversity, while comparative approaches assemble sequence data by mapping it to that of a closely related reference organism (Pop, 2009). This approach is essentially used for re-sequencing applications or for assembly of closely related organisms, such as strains of the same species (Pop, 2009). The reference sequence effectively acts as a guide to infer genomic structure, similar to the modelling of a protein sequence in order to obtain possible tertiary structure. Alignments of multiple reads may identify regions where possible insertion or deletions (InDels) may have occurred. In addition to read alignment, sequence quality data can be used to identify single nucleotide polymorphisms (SNPs). A major obstacle in this alignment approach occurs when reads do not exist in the reference (Metzker, 2010). Some examples of

comparative assemblers are ABBA and AMOScmp (Pop, 2009). The following methods which will be discussed are the three main strategies for *de novo* sequence assembly.

3.1.3.1 The 'Greedy' algorithm

These algorithms represent the simplest solution to computation of overlaps [pairwise alignments between sets of reads] (Bansal, 2005). During this process, contigs are constructed from individual reads in an iterative fashion, starting with reads that produce the best overlap whereby one read prefix shares very high sequence similarity to the suffix of another read (Raes *et al*., 2007; Pop, 2009). Overlap quality is locally optimised and the processing in this manner may cause misassembly of repeats. Some software programs that utilise some form of the 'greedy' approach includes CAP3 and TIGR Assembler [for Sanger data], and SSAKE and VCAKE [for short read sequence data] (Pop, 2009).

3.1.3.2 Overlap-layout-consensus (OLC)

The overlap-layout-consensus approach uses three distinct steps to allow for global analysis of read relationship. Firstly, reads are compared to produce pairwise overlaps, as with the 'greedy' approach. Secondly, an overlap graph is constructed whereby each read is considered as a node and if an overlap is detected, an edge connects the nodes (Zerbino and Birney, 2008). Hamiltonian paths are identified during this layout stage (Pop, 2009). These paths correspond to segments of the assembled genome, and ultimately, a single path that traverses each node once, using all reads, is provided as the input for the last step. Finally, consensus computation weighted by sequence quality is used to reconstruct the sequence (Churchill and Waterman, 1992). Some examples of software which rely on this approach include Newbler, Arachne, Celera Assembler and Edena (Pop, 2009; Zerbino and Birney, 2008). Edena has been used for sequence assembly of genomes by shotgun sequencing and is supplied with the 454 Pyrosequencer, and has also been successfully applied for short read data assembly (Hernandez *et al*., 2008).

3.1.3.3 Eulerian path

This computational method breaks down and lists all reads into a defined length, *k*, which are always odd to ensure that each generated *k-mer* is not its own reverse compliment (Zerbino and Birney, 2008). This *k-mer* spectrum is a particularly valuable output for shotgun experiments where all reads have equal length. With short reads, and therefore small values of *k*, the sequence is easier to interrogate (Pop, 2009; Zerbino and Birney, 2008). De Bruijn graphs contain the *k*-1 length prefixes and suffixes and renames them 'nodes' (Pop, 2009). If an adjacent *k*-1 *mer* has an exact overlap match of length *k*-2, then nodes are joined by an edge. The Eulerian path is constructed through the graph using every edge, which are themselves constructed from each *k-mer* within a set of reads (Pop, 2009; Zerbino and Birney, 2008). Although the Eulerian strategy is better suited to NGS technologies that generate short reads, the assembly process may be particularly sensitive to sequence errors subsequently creating false *k-mers* which increase both complexity and size of resulting graphs (Pevzner *et al*., 2001). However, short read sequencing technologies produce high coverage rates which allows for more valid assembly data (Morozova and Marra, 2008; Zerbino and Birney, 2008). In cases where sequences are generated using a combination of sequencing technologies, Eulerian based strategies may be the best solution. Velvet and Allpaths are two programs that effectively utilise the Eulerian strategy for assembly of short read sequences (Pop, 2009; Zerbino and Birney, 2008).

3.1.4 ORF calling

Genes are the basic hereditary units and provide regulation, structure and function to biological systems (Rouchka and Cha, 2009). Following genomic sequence reconstruction, the basis of further research is to predict ORFs, provide annotation and ultimately decipher probable function (Skolnick and Fetrow, 2000, Angelova *et al*., 2010). A prokaryotic gene is generally defined as the longest ORF for a given region of DNA and contains a start codon,

followed by a variable number of in-frame codons and a termination site (Angelova *et al*., 2010; Koonin and Galperin, 2003). The prediction of these coding regions is performed via computational gene calling. Tools for computational gene calling are divided into two categories and due to their complementary nature, both prediction approaches are commonly used in a study (Angelova *et al*., 2010). Those that rely on the intrinsic parameters of the individual DNA sequences, such as coding potential start and stop codons and promoter elements, are referred to as *ab initio* tools (Windsor and Mitchell-Olds, 2006; Guigó *et al*., 2000). Extrinsic tools for gene prediction are based on sequence similarity to genes within databases and are generally restricted to the collection of known genes (Hoff *et al*., 2008; Raes *et al*., 2007; Besemer and Borodovsky, 2005). Prokaryotic gene finding tools generally employ algorithms which require large numbers of experimentally determined gene sequences for training (Noguchi *et al*., 2006).

Intrinsic methods for gene calling in prokaryotes may present some challenges due to the small intergenic regions, overlapping of genes and the high frequency of occurrence of translation start codons (Angelova *et al*., 2010; Mathé *et al*., 2002). With respect to sequences derived from metagenomic projects, low sequence quality can produce contigs which contain frame shifts, thereby coding for an in-frame stop codon, which are not present in the true genomic sequence (Krause *et al*., 2006). In addition, sequences may contain a collection of sequence reads of varying length from unidentified organisms. Differences in contig length may negatively impact the quality of gene calling and ultimately functional prediction (Raes *et al*., 2007; Wooley *et al*., 2010). Sequence properties upon which intrinsic methods are based, may not be present in short contig stretches and in these cases, heuristically derived models with non-supervised training are well suited for gene calling in these fragmented anonymous sequences (Noguchi *et al*., 2006; Hoff *et al*., 2008).

3.1.4.1 *Ab initio* gene finders

Intrinsic methods for ORF prediction tend to display high sensitivity, but low specificity, and are generally not comparative tools (Windsor and Mitchell-Olds, 2006). FgenesB, GeneMark and GLIMMER are classical prokaryotic gene prediction tools which are used to predict ORFs which have no current database homologues (Raes *et al*., 2007; Besemer and Borodovsky, 2005). The coding potential of uncharacterised ORFs are estimated by building statistical Markov models to known coding regions (Koonin and Galperin, 2003). Hidden Markov Model (HMM) based techniques, such as those employed by GeneMark and GLIMMER, predict each base in a sequence with maximum probability using a state transition and proceeds to match a current nucleotide character with the predicted one (Salzberg *et al*., 1998). The state transition is derived for statistical training using known sample sequences (Bansal, 2005). Horizontally transferred DNA sequences which may otherwise affect the statistical model, due to differences in codon usage can be differentiated by Markov models (Angelova *et al*., 2010). The probable existence of an ORF is predicted in all six frames by utilising a Bayesian method to search for characteristic genetic features such as ribosome binding sites, gene overlaps, start and stop codons, as well as uninterrupted genes (Besemer and Borodovsky, 2005). Mono- and di- codon frequencies (to discriminate between coding and non- coding ORFs), GC content, codon usage and intergenic distance between ORFs are additional measures which can be utilised to increase prediction accuracy and are generally integrated into many existing ORF finding programs (Angelova *et al*., 2010; Noguchi *et al*., 2006). Extracted ORFs are then scored based on statistical model estimates of annotated genomes (Hoff *et al*., 2008). Dynamic programming computes the final ORFs from the different scores and the output generated by ORF prediction tools includes the gene boundaries, length, coding strand and the translated sequence (Besemer and Borodovsky, 2005; Hoff *et al*., 2008).

- GLIMMER

This *ab initio* tool uses Interpolated Markov Models (IMM) which are generalised Markov chains with variable order (Salzberg *et al*., 1998; Angelova *et al*., 2010). The number of probabilities that must be estimated increases exponentially as higher order models are used, ultimately requiring more training which is not always available for new sequences. GLIMMER may calculate the probability for all Markov chains from the 0-th to the 8-th order and the lower order models are used when highest order statistics do not provide sufficient data (Angelova *et al*., 2010, Koonin and Galperin, 2003). By using long ORFs as a training set, GLIMMER generally lengthens predicted ORFs. This can, however, be detrimental to accurate gene calling in AT rich organisms as the frequency of AT-rich stop codons is high.

- FgenesB

This software is available from SoftBerry. This tool is also a Markov chain based algorithm but unlike Genemark and GLIMMER, is able to predict tRNA and rRNA genes in the sequence. An automated version of FgenesB is available and not only predicts putative ORFs but also annotates them using homology based searches, accompanied by operon, promoter and terminator predictions (Angelova *et al*., 2010). When tested on the *Pseudomonas aeruginosa* genome, FgenesB performed better than GLIMMER and GeneMarkS by prediciting the largest number of correct genes.

However, no tool is able to find ORFs in a given DNA sequence with 100 % accuracy and it is therefore critical to improve current methodologies. Currently, the best approach is to use a combination of prediction tools (Angelova *et al*., 2010).

3.1.4.2 Homology based tools

Coding regions generally have lower evolution rates and this can be used to identify them by comparison to existing genes. Two methods based on similarity searches are local and global alignments. The BLAST family of programs are local alignment tools which are most commonly used for homology based searches. The most significant limitation to gene prediction using these tools is the lack of sequence similarity between organisms in the existing database (Angelova *et al*., 2010). These computationally expensive tools require massive amounts of extrinsic data and tend to be inadequate for the identification of novel genes (Hoff *et al*., 2008).

3.1.5 Functional annotation

Following the prediction of protein coding regions, functional annotation of proteins is required (Raes *et al*., 2007; Bansal, 2005). BLAST and its variations are the most commonly used algorithms for sequence similarity searches and pairwise alignments to identify homologues in a variety of databases and therefore possible gene function (Bansal, 2005; Sleator and Walsh, 2010). Amino acid comparisons tend to be more sensitive than nucleotide alignments as the nucleotide databases are large and error prone (Koonin and Galperin, 2003). Ultimately, analysis is only as good as the reference database used for functional annotation. Theoretically, if significant homology is shared between sequences, common ancestral evolution may be inferred. This 'guilt by association' approach then suggests functional transfer (Sleator and Walsh, 2010). However, fewer protein queries register to known homologous superfamilies and annotations are becoming increasingly limited (Raes *et al*., 2007; Sleator and Walsh, 2010). Another question remains unresolved in terms of defining homology. Sequences that are 99 % similar are definitely close relatives, but what is the threshold percentage at which they are not (Koonin and Galperin, 2003)?

There are several universal databases which contain sequences from all species and may either be simple archives, or expertly curated collections (Mulder *et al*., 2008). Archive databases add very little additional annotation information to the sequence records but do provide users access to sequences as quickly as possible. Curated databases greatly increase the value of sequence data by enrichment with additional information from validated sources, making them highly reliable. One of the most comprehensive databases which utilises manual intervention to create non-redundant records is the universal protein resource (UniProt) (Mulder *et al*., 2008). Specialised databases also exist and are devoted to particular proteins or protein families and vary both in scope and size of the curated data. The protein databank (PDB) is just one example of such a database and harbours the collection of resolved protein tertiary structures. Databases such as InterPro rely on integration of database resources, which each use various methods for annotation and curation. To predict inferred protein function, the classification of proteins into families is a valuable starting point. Protein signature methods are the best approach for this task as these signatures, generally a unique folding pattern of α-helices and β-sheets, are diagnostic conserved domains belonging to various protein families (Bansal, 2005; Mulder *et al*., 2008). Domain related databases include Pfam PROSITE, TIGRfam and SMART. InterproScan may be used to calculate protein matches from all integrated protein signature databases and provides a powerful tool for annotation and functional prediction. An entry not only contains high quality information on protein families but also links to Gene Ontology (GO) terms (Mulder *et al*., 2008). Best match techniques are not however capable of identifying all possible functions in multidomain proteins and may fail to provide annotation if the function is localised to specific regions, such as hydrophobic domains, or if the function is dependent on the presence of specific amino acid patterns (Bansal, 2005). Another method elucidating function from annotation is by clustering genes into Clusters of Orthologous Groups (COG's) which

are orthologous groups of paralogs from three or more phylogenetic lineages. Ultimately, two proteins from different lineages that are clustered in the same COG, are orthologs which have evolved from a common ancestor and may therefore retain the same function (Tatusov *et al*., 1997). Furthermore, genome synteny (the order of genes on a genome) may also be utilised as a method to infer function. This method is mainly used for comparisons between highly related organisms. While this method is successful for operons which are strongly conserved during evolutionary events, the order of homologous groups, particularly those which may function even if a subunit is lost, are generally not the same in two organisms as InDels and HGT events are not uncommon in microbial communities (Bansal, 2005; Koonin and Wolf, 2008). Reconstruction of metabolic pathways requires four steps; firstly, enzymes and functions in a newly sequenced genome must be identified by ortholog analysis. Secondly, analysis of promoter regions allows for the identification of gene groups which share a common promoter. Pairwise comparisons with multiple genomes then allows for derivation of gene groups in the genome and lastly, these gene groups may be connected using existing knowledge of biochemical pathways (Bansal, 2005).

The remainder of ORFs for which function cannot be predicted are categorised as 'conserved hypothetical' or 'hypothetical' genes. The former have homology to genes of unknown function while the latter have no known homologs (Sivashankari and Shanmughavel, 2006). Genomic studies highlight the need for functional characterisation of the pool of hypothetical proteins. Most notably, in the Sargasso sea metagenome, more than half the total number of genes were classified as hypothetical proteins (Rusch *et al*., 2007). In newly sequenced genomes, approximately 30-40 % of coding sequences fall into one of the two above mentioned categories, a phenomenon known at the '70 % hurdle'(Koonin and Wolf, 2008, Bork, 2000). Clearly, a bottleneck now exists which not only affects all downstream annotation of both existing and future genome sequences, but may also impair our 'complete'

understanding of biological systems. Most technologies in bioinformatics have been developed based on knowledge gained from experimental validation. A major threat to genomic based studies is the lack of experimental evidence for gene functions hampering the required improvements in bioinformatic strategies (Bansal, 2005). The implications of this knowledge gap are wide ranging, for instance, in model organisms, our understanding of metabolic processes remains incomplete (Bertin *et al*., 2008), and only once experimental determination of function is obtained, will the impact of hypothetical proteins during experimental manipulation or industrial utilization be revealed.

3.1.6 Comparative genomics and metagenomics

Prokaryotes, viruses and microeukaryotes dominate all of the earth's ecological niches and are vital for ecosystem functioning. These organisms are the primary source of nutrients and are a ubiquitous biomass essential to life (Wooley *et al*., 2010). Most ecological cycles that shape environments are due to the complex microbial communities that inhabit them. Furthermore, microbial communities tend to maintain genetic elements which may ensure efficient structure and functioning of bacterial assemblages in a particular ecological niche (Raes *et al*., 2007). The complex physiology of interacting organisms holds clues to detailed understanding of the vital roles that microbial communities play in environmental systems and is rapidly gaining interest in various fields of research (Bertin *et al*., 2008). 16S rRNA analysis for any given environment indicates that less than 1 % of microbes have been cultured using standard techniques (Bertin *et al*., 2008; Snyder *et al*., 2009). Metagenomics is defined as the culture independent, sequence-based or function-based analysis of the collective environmental DNA in a given habitat (Simon and Daniel, 2009). Metagenomic-based approaches are powerful tools for the analysis of the identity, genetic and functional potential of the 'yet to be cultured' microbial world, and may provide a different perspective of ecosystem functioning (Noguchi *et al*., 2006; Raes *et al*., 2007). Large scale environmental

sequencing projects produce massive amounts of DNA data and the analysis of these datasets provides a platform for discovery of functional novelty and biological diversity (Huson *et al.*, 2009; Raes *et al*., 2007). For example, the global oceanic sampling (GOS) study, an extension of the Sargasso Sea project, liberated 6.3 Gb of sequence, allowing for the prediction of over six million proteins (Hutchinson, 2007). Aside from the discovery of products which may be used in a variety of industrial applications, these studies also provide an understanding of ecosystem function and the specific genetic traits which allows organisms to thrive in diverse habitats (Bertin *et al*., 2008). Shotgun sequencing of metagenomic DNA has been applied to a number of environmental studies, most notably, the acid mine drainage biofilm and the Sargasso Sea (Noguchi *et al*., 2006). GC content variations allowed for accurate genomic reconstruction of the acid mine drainage community and the near complete genomes of two chemolithic microbes, *Leptospirillium* species and *Ferroplasma acidarmanus*, were obtained and the ecological roles of these organisms were elucidated (Tyson *et al*., 2004). While *Ferroplasma* species clearly exhibited a heterotrophic lifestyle, the *Leptospirillium* genome indicated that this species was a keystone organism for nitrogen fixation in the community (Riesenfeld *et al*., 2004). Furthermore, as expected, genes associated with the removal of toxic elements were present in all the acid mine drainage genomes (Handelsman, 2004).

A large and diverse pool of proteo-rhodopsin genes were discovered in the Sargasso sea study (Rusch *et al*., 2007) which increased the number of sequenced representatives of this protein family tenfold (Handelsman, 2004). Furthermore, uptake genes for a variety of phosphorous compounds were readily identified among the 1.2 million genes (Riesenfeld *et al*., 2004). Analysis of the Sargasso Sea data also provided evidence of 10 plasmids and multiple double-stranded DNA bacteriophages which not only act as vehicles of gene transfer but also genetic reservoirs (Venter *et al*., 2004). While genes involved in role categories such as

amino acid biosynthesis, energy metabolism, and transport and binding proteins were readily identified (Venter *et al*., 2004), the most astounding discovery was that over 700 000 genes in the dataset were conserved hypothetical proteins and the combined number of functionally categorised genes were less than the number of hypothetical proteins, again illustrating one limitation of current functional annotation in sequencing projects. However, these results also demonstrate the potential to discover vast quantities of novel bacterial functions.

Due to next generation sequencing technologies, current research is shifting focus from single genes to whole genomes and 'top down' exploratory investigations are becoming increasingly valuable (Mulder *et al*., 2008). Comparative genomics allows researchers to develop general conclusions on how an ecosystem may imprint on genome sequences. For example, GC content has been shown to vary greatly in the genomes of microorganisms from different environments. Higher G + C values are found in soils when compared to open sea water. This may possibly be due to the existence of fast evolving AT- rich organisms with small genomes in aquatic environments (Raes *et al*., 2007). In addition, it has been observed that parasites and intracellular symbionts tend to exhibit smaller genomes, enriched in AT frequency and which contain a large fraction of genes encoding for transcription, translation and replication machinery when compared to their free living prokaryotic counterparts (Koonin and Wolf, 2008). In addition, genome comparisons have shown the rich diversity even in closely related strains, lending strength to the concept of the pangenome; the core genomic elements present in all strains, with the remaining genetic complement being unique to an individual (Bertin *et al*., 2008). Comparative analysis of similar and /or diverse habitats provides researchers with the opportunity to discover general tends that may link community profiles to a set of conditions in an environment (Raes *et al*., 2007; Sivashankari and Shanmughavel, 2006). Additional information pertaining to protein families and cellular processes which are conserved for specific environmental adaptations may also be obtained

by comparative genomic studies (Raes *et al*., 2007). Some other applications of NGS technology includes variant discovery in re-sequenced genomes, direct RNA sequencing to catalogue transcriptomes, genome wide profiling and metagenomic based species classification and gene discovery (Metzker, 2010).

Aims and objectives

4. Assemble and annotate full fosmid sequences using both manual and automated approaches
5. Describe genetic potential contained within the sequences
6. Discuss key genes possibly linked to adaptation to environmental conditions experienced in Antarctic Dry Valley soils

3.2 Materials and methods

3.2.1 Sample preparation and sequencing

A large contig shotgun fosmid library was obtained in a previous study and functionally screened for lipolytic activity on tributyrin agar (Anderson, 2008). Fosmid DNA was extracted from clones LD1, LD2, LD3, LD4, LD6, LD7, LD9 and LD13. DNA concentration was measured by fluorimetery using the Quanti-iT ™ ds DNA BR assay kit and the Qubit ™ system (Invitrogen, USA) according to the manufacturer's specifications. Equal concentrations of each fosmid were pooled to a total of 5 µg and sequenced using the Solexa system (University of the Western Cape 2008). Each individual fosmid was also end-sequenced (University of Stellenbosch) using the primers provided in Table 2.1.2, in order to generate mate-pair data. Additionally, extracted fosmid DNA was subjected to HindIII/EcoRI restriction digestion and agarose gel electrophoresis in order to estimate insert size.

3.2.2 Contig assembly

Over 3 million short read sequences of 36 bp were assembled into larger contigs using *de novo* assembly in CLC genomics workbench and VELVET. Default assembly parameters were used in the CLC assisted assembly and three independent *k-mer* lengths (23-, 25- and 27- *k-mer*) were chosen for VELVET assembly. Contigs which resulted from both assemblies, were larger than 2 kb and matched fosmid end-sequences, were used to construct the draft insert sequences for four of the fosmids (LD4, LD6, LD7 and LD13). LD6 was removed from further analysis in this study and became the subject of a separate study (Dean Booysen, MSc). Regions which were not assembled into contigs constitute gaps in the draft genomic fragments and were finished by subsequent primer walking steps. Overlap assemblies were produced using Sequencher, version 4.10.1 (Gene Codes Corporation, USA) and background vector sequence was removed.

3.2.3 ORF calling

Two *ab initio* tools, GLIMMER (www.ncbi.nlm.nih.gov) and FgenesB (www.softberry.com), were used to predict ORFs in the assembled sequences. Results from gene calling through FgenesB using the archaeal/bacterial genetic code were used for manual annotations (whereby, predicted ORFs were individually submitted for BLASTp and InterProScan anaylsis). The three fully assembled fosmid sequences were also submitted to the IGS annotation engine for gene calling, annotation and curation in the MANATEE database (www.ae.igs.umaryland.edu).

3.2.4 Annotation

Predicted ORFs were automatically translated into protein sequences in all six reading frames using FgenesB. All predicted proteins were manually searched for homologs in the UniProt database (www.uniprot.com) and assigned COG categories using the NCBI COGnitor program (www.ncbi.nlm.gov). ORFs were automatically annotated and characterised into TIGRfam functional groups using the IGS pipeline. Overall GC content and sequence maps were obtained by DNAMAN. General functional data related to homology was obtained using curated databases such as InterPro, Pfam and extensive literature surveys. A summary of methods used in this study is provided in Figure 3.2.1.

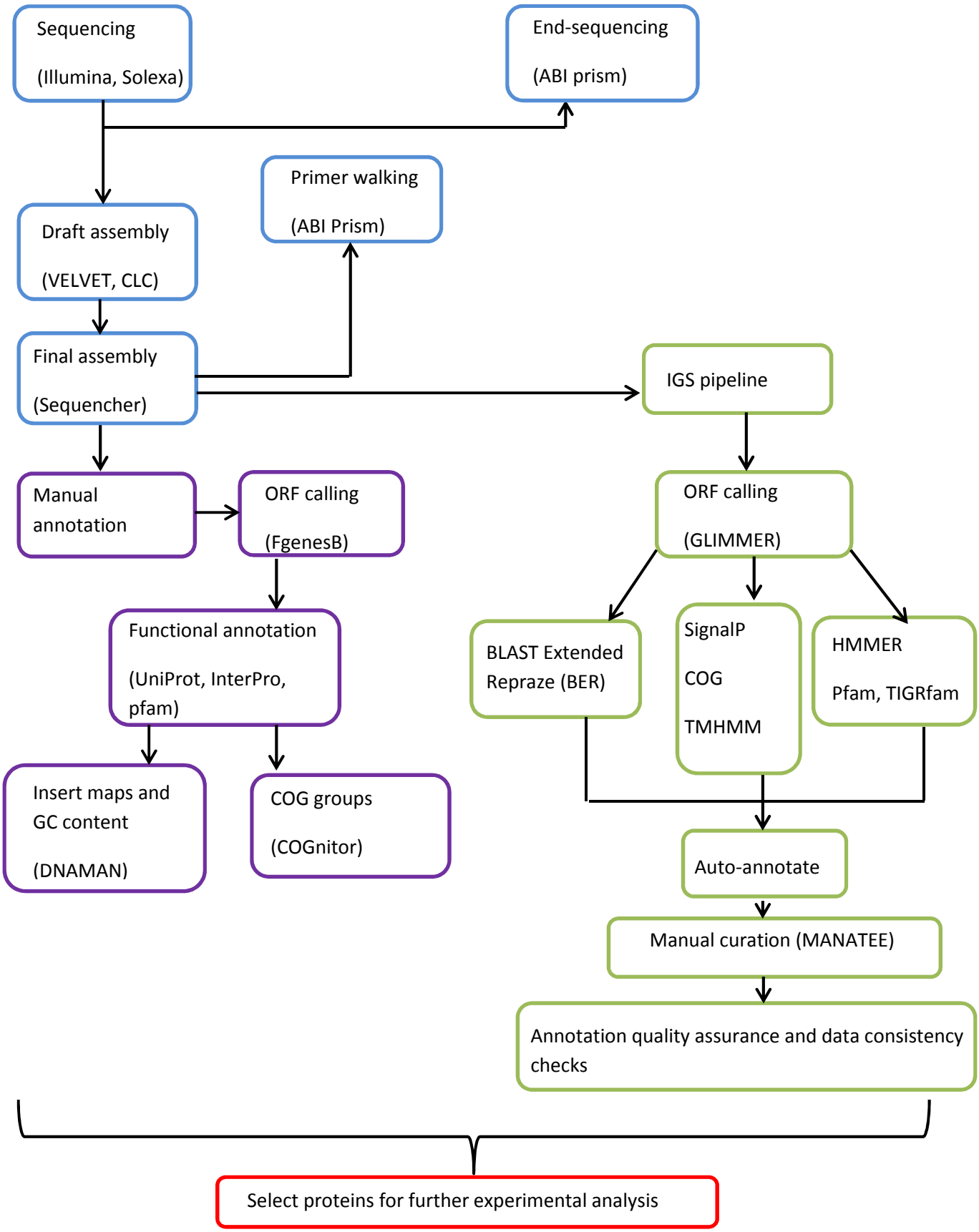


Figure 3.2.1: Summary of methods used in this study represented as a flow diagram. Blue blocks indicate the process of generating the full length fosmid sequences. Blocks in purple illustrate manual analysis, while those in green illustrate the automated pipeline methodology.

3.3 Results and discussion

3.3.1 Sequence assembly

Solexa sequencing of the fosmid clones generated 3.1 million reads of 36 bp per read. *De novo* contig assembly using the CLC Genomics workbench generated 66 contigs by matching over one million reads (Table 3.3.1.1). The largest contig obtained was 22 500 bp in length. Assembly generated from VELVET, using 23 *k-mer* lengths generated 166 contigs with the largest node 26 500 bp in length. One possible way to explain why all fosmid fragments were not assembled in their entirety, and why a large number of reads were not assembled into contigs, is based on the assembly programs themselves. Generally, due to the various ways that a sequence may be constructed, an exponential number of distinct paths can be found in an Eulerian assembly and just one of the possible paths would correspond to the correct sequence assembly. Additionally, breaking of reads into *k-mers* used to construct the de Bruijn graphs may result in a loss of information (Pop, 2009). Contigs assembled from both programs were used together with the end-sequences generated for each fosmid for draft assembly by aligning sequences in Sequencher using a minimum overlap of 20 bp and minimum sequence match of 95 %. The largest contigs from both assemblies aligned with each other as well as with the end sequences of clone LD13 (Figure 3.3.1.1)

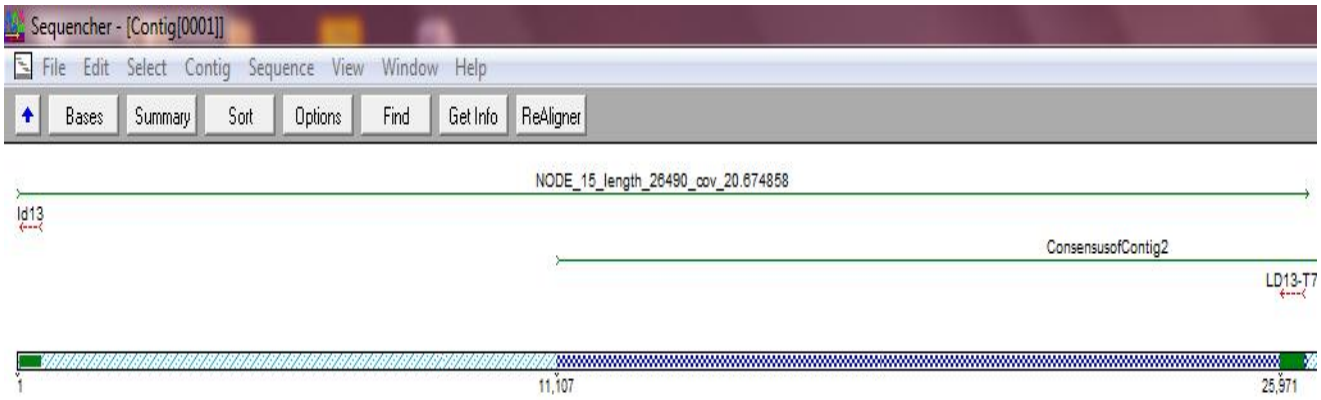


Figure 3.3.1.1: Fosmid assembly diagram of clone LD13. Nodes represent contigs generated from VELVET assembly and 'Consensusofcontig' represent CLC genomics workbench generated contigs. End sequences are shown in red with reverse sequence in small letters and forward sequence in caps. For clone LD13, Node_15 and Consensusofcontig2 matched to the forward and reverse end sequences with 100 % sequence identity.

In order to generate the complete insert sequences of clone LD4 and LD7, multiple primer walking steps were required to align various contig fragments (Figure 3.3.1.2 and Figure 3.3.1.3). Fortunately, in the draft assembly, mate-pair information allowed the identification of at least one contig matching to the forward and reverse end sequences of each clone. This facilitated primer walking in both directions on each clone. The three fosmid inserts were successfully assembled and the final length correlated well to the estimates determined by agarose gel electrophoresis. The sequences were subjected to nucleotide BLAST against the Refseq_genome database in an attempt to identify possible homologs. All clones showed less than 20 % coverage to any sequenced genome, low levels of homology were detected in clone LD4 (*Psychrobacter* spp) and LD7 (*Aeromicrobium* spp) and none could be detected for clone LD13. This may indicate that the three fosmids contain high levels of sequence novelty. Fosmid insert sizes and overall GC content are provided in Table 3.3.1.2.

Table 3.3.1.1: ***De novo*** **assembly report generated by CLC genomics workbench**

	Count	Average length (bp)
Reads	3 132 350	36
Matches	1 093 688	36
Not Matched	2 038 622	36
Contigs	33	3104

Table 3.3.1.2: Insert size estimated by agarose gel electrophoresis and true size of assembled fragments as well as GC content of the three fosmid clones evaluated in this study.

Fosmid name	Estimated size (bp)	True size (bp)	Overall GC content
LD4	38 400	37 425	42 %
LD7	32 000	32 536	62 %
LD13	26 500	26 420	47 %

Figure 3.3.1.2: Fosmid assembly diagram for clone LD4. Included are the various sequences (named according to the designed primer) generated from primer walking.

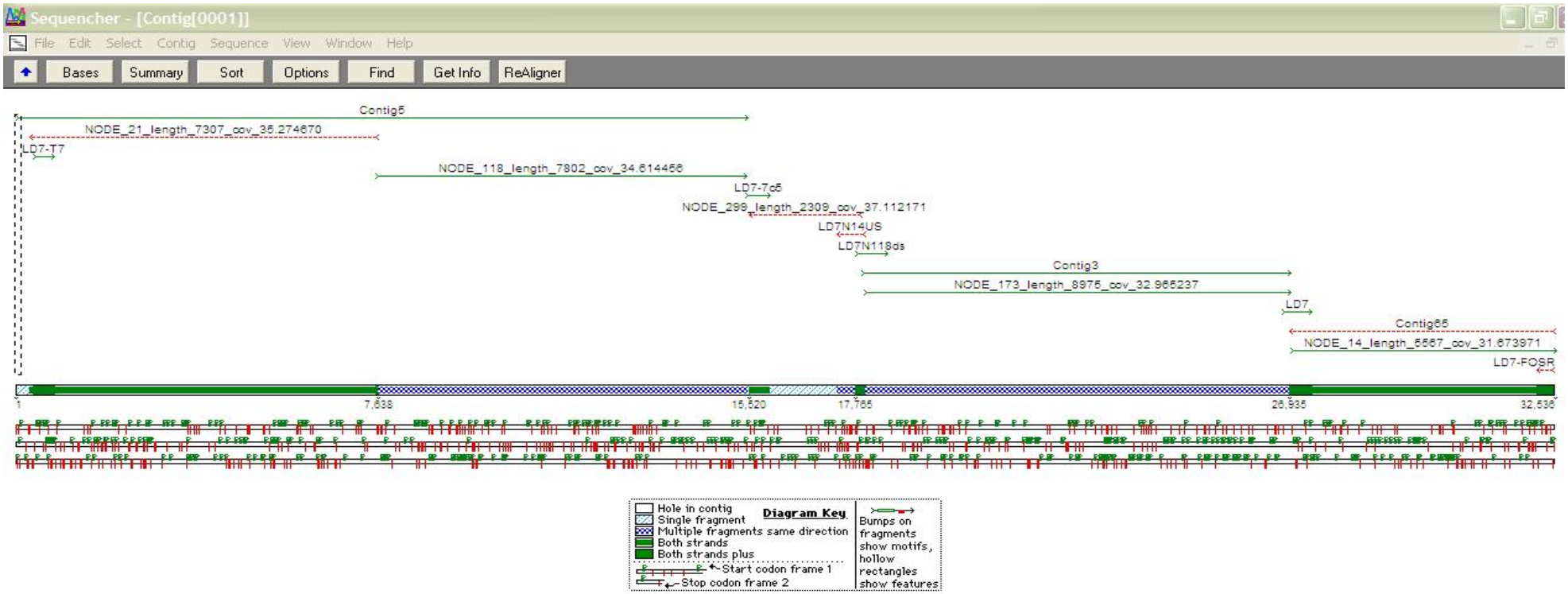


Figure 3.3.1.3: Fosmid assembly diagram for clone LD7. Included are the various sequences (named according to the designed primer) generated from primer walking.

3.3.2 Prediction of ORFs

Manual ORF calling in the fosmid sequences predicted 31, 37 and 30 coding sequences for clone LD4, LD7 and LD13 respectively. The automated pipeline predicted more coding sequences in LD4 and LD13 (37 ORFs and 33 ORFs respectively) but the same number in clone LD7. There are two possible explanations for these differences; firstly, the automated pipeline predicts ORFs using GLIMMER, while the manual prediction was performed using FgenesB. Although both tools are based on Markov models for prediction, training sets for the algorithms are different, which could result in a difference in the numbers of predicted ORFs. Secondly, the training set for GLIMMER is based on long coding regions and, due to the low GC content in clones LD4 and LD13, a higher proportion of AT- rich stop codons would occur thereby increasing the number of putative ORFs in these sequences. In addition, many of the ORFs predicted using GLIMMER showed increased gene length. This was expected and has previously been observed for this tool due to long ORF training data (Angelova *et al*., 2010). Based on the abovementioned results, and a comparison of these two programs showing a higher level of performance in FgenesB (Mavromatis *et al*., 2007), all further analysis and experimental validation was performed using FgenesB. The predicted ORFs, basic gene descriptors, homologs and possible function as well as COG categories for all fosmid sequences is summarised in Table 3.3.1.3 [A-D].

To discern species composition, each fosmid sequence was searched for genes which could provide some phylogenetic grouping in order to link possible genetic function to community members. None of the fosmid sequences in this study contained 16S rRNA sequences, but other phylogenetic descriptors such as the Elongation factor Tu, RecA/RadA , GreA/GreB, HSP70 and Nif genes (Simon and Daniel, 2009) were predicted in fosmid clones LD4 and LD13. Clone LD13 contained a putative GreA/GreB coding sequence (13ORF22) which allowed for phylogenetic grouping to the genus *Pseudomonas*. Clone LD4 contained an

HSP70 coding sequence (4ORF7) which shared 93 % homology to DnaK from *Psychrobacter* spp. Although phylogenetic clustering is purely speculative, based on overall sequence homology across the majority of predicted ORFs, a possible candidate genus can be assigned for the final fosmid sequence. Bacteria isolated and identified in Dry Valley soils tend to belong to a wide range of genera, including *Arthrobacter*, *Bacillus*, *Bacteroidetes*, *Corynebacterium*, *Cytophaga*, *Flavobacterium*, *Micrococcus*, *Planococcus*, *Pseudomonas*, *Psychrobacter* and *Streptomyces* (Adams *et al*., 2006; Cowan *et al*., 2004; Shravage *et al*., 2007; Smith *et al*., 2006). The presumptive phylogenetic affiliations made in this study are consistent with a recent study of bacterial diversity of soils beneath seal carcasses in the Antarctic Dry Valleys, which showed similar dominant microbial classes (Tiao *et al*., 2011).

Table 3.3.1.3: Complete list of ORFs, the gene length and orientation, the percentage amino acid sequence identity as well as cellular roles for fosmid sequences in this study. A] Clone LD4, B] Clone LD7, C] Clone LD13 and D] Description of COG domains and their function.

A

Feature name	Common name	Gene boundary	Coding strand	Homolog (e value)	Sequence identity (%)	TIGRfam role (Sub category role)	COG Category
4ORF1	RimN/SUA5/YrdC	1132-1767	+	Q4FPS7 ($1X10^{-76}$)	69	Transport and binding proteins and cellular processes	J
4ORF2	NAD epimerase	2057-3061	+	Q1Q7W5 ($1X10^{-130}$)	69	Energy metabolism (sugars)	R
4ORF3	Phosphoesterase	3265-4311	+	Q1Q7W6 ($1X10^{-127}$)	63	Enzymes of unknown specificity	R
4ORF4	Phosphoesterase	4431-5576	+	Q1Q7W7 ($1X10^{-152}$)	69	Enzymes of unknown specificity	R
4ORF5	Lipase	5551-6945	-	Q1Q7W8 ($1X10^{-143}$)	53	Enzymes of unknown specificity	I
4ORF6	GrpE	7432-8037	+	Q9LS16 ($2X10^{-83}$)	80	Protein fate (folding and stabilisation)	O
4ORF7	DnaK	8305-10245	+	Q4FPS9 (0)	93	Protein fate (folding and stabilisation)	O
4ORF8	OmpA-like	10682-11215	+	Q1Q7X3 ($1X10^{-71}$)	73	Cell envelope (other)	M
4ORF9	OmpA-like	11455-12003	+	Q1Q7X3 ($2X10^{-56}$)	60	Cell envelope (other)	M
4ORF10	NfeD-like	12249-12860	+	Q1Q7X4 ($6X10^{-60}$)	68	Unknown function (general)	NO
4ORF11	Band7 protein	13051-13932	+	Q1Q7X5 ($1X10^{-142}$)	92	Unknown function (general)	O
4ORF12	Conserved hypothetical	14244-14763	+	Q1Q7X6 ($3X10^{-75}$)	75	Enzymes of unknown specificity	H
4ORF13	Conserved hypothetical	14887-15660	+	Q1Q7X6 ($2X10^{-93}$)	69	Conserved hypothetical	
4ORF14	Oxidase	15729-17267	-	Q1Q7X7 (0)	85	Biosynthesis of cofactors, and carriers (heme, porphyrin and cobalamin)	H
4ORF15	Acetyltransferase	17527-18471	-	F2KEL7 ($6X10^{-50}$)	39	Protein fate (degradation on proteins and peptides)	Q
4ORF16	Hypothetical	18517-18840	+			Hypothetical	
4ORF17	Conserved hypothetical	19345-21066	+	D6KZV8 ($4X10^{-40}$)	22	Conserved hypothetical	
4ORF18	Mg chelatase	21351-22943	+	Q1Q7X9 (0)	85	Unknown function (general)	O
4ORF19	Formate tetrahydrofolate ligase	22956-24713	-	A5W1O4 (0)	80	Unclassified	F
4ORF20	Phosphoglycolate phosphatase	24697-25401	-	Q1Q7Y0 ($1X10^{-100}$)	80	Not predicted by IGS annotation engine	R
4ORF21	Conserved hypothetical	25577-26077	-	Q1Q7Y1 ($2X10^{-86}$)	92	Conserved hypothetical	
4ORF22	Hypothetical	26373-26723	+			Hypothetical	
4ORF23	Conserved hypothetical	27068-27490		Q1Q7Y2 ($6X10^{-60}$)	82	Conserved hypothetical	
4ORF24	Imidazole glycerol phosphate dehyratase	27820-28464	+	QIQ7Y3 ($1X10^{-113}$)	91	Amino acid biosynthesis (Histidine family)	E
4ORF25	Imidazole glycerol phosphate synthase	28633-29268	+	Q1Q7Y4 ($1X10^{-109}$)	89	Amino acid biosynthesis (Histidine family)	E
4ORF26	Conserved hypothetical	29459-29878	+	Q1Q7Y5 ($2X10^{-34}$)	56	Conserved hypothetical	
4ORF27	Aminotransferase	29909-31267	-	Q1Q7Y6 (0)	91	Transport and binding proteins	KE

4ORF28	Phosphoglycerate dehydrogenase	31552-32499	+	F5SQG6 ($1X10^{-153}$)	84	Enzymes of unknown specificity	E
4ORF29	Hypothetical	33296-33433	-			Hypothetical	
4ORF30	Reverse transcriptase	33583-35052	+	F5SQA1 (0)	63	DNA metabolism and mobile element	L
4ORF31	FtsX-like	35264-37096	-	Q1Q7Z0 (0)	81	Transport and binding proteins	S

B

Feature name	Common name	Gene boundary	Coding strand	Homolog (e value)	Sequence identity (%)	TIGRfam role (Sub category role)	COG function
7ORF1	*PspA*	129-632	-	E2SE62 ($6X10^{-36}$)	76	Regulatory (protein interaction)	I
7ORF2	MgtC	730-1425	-	AOQHD5 ($4X10^{-56}$)	54	Unknown function	S
7ORF3	Conserved hypothetical	1600-2190	+	AOK1G5 ($4X10^{-8}$)	33	Hypothetical	
7ORF4	CpaF	2392-3639	+	E2SE76 ($1X10^{-167}$)	74	Unclassified	N
7ORF5	TypeII/TypeIV Secretion protein	3777-4466	+	E2SE75 ($2X10^{-82}$)	67	Protein fate (secretion and trafficking)	S
7ORF6	TypeII sectretion protein	4739-5365	+	E2SE74 ($7X10^{-67}$)	62	Protein fate (secretion and trafficking)	S
7ORF7	Conserved hypothetical	5387-5554	+	E2SE73 ($2X10^{-17}$)	84	Conserved hypothetical	
7ORF8	TadE	5623-5943	+	E2SE72 ($5X10^{-23}$)	58	Cell envelope (surface structures)	
7ORF9	Conserved hypothetical	5940-6380	+	E2SE71 ($1X10^{-42}$)	60	Conserved hypothetical	
7ORF10	Conserved hypothetical	6377-6793	+	E2SE70 ($2X10^{-13}$)	38	Hypothetical	
7ORF11	Peptide chain release factorII	6861-7967	+	E2SE81 ($1X10^{-170}$)	80	Protein synthesis (Translation factor)	J
7ORF12	FtsE	8044-8733	+	E2SE84 ($1X10^{-107}$)	84	Cellular processes (Cell division)	D
7ORF13	FtsX	8742-9662	+	E2SE85 ($1X10^{-128}$)	76	Transport and binding	DR
7ORF14	Peptidase	9777-11108	+	E2SE86 ($1X10^{-123}$)	52	Protein fate (degradation)	M
7ORF15	SsrA	11142-11621	+	E2SE87 ($3X10^{-65}$)	76	Protein synthesis (other)	O
7ORF16	Amidohydrolase	11622-12494	+	E2SE88 ($1X10^{-114}$)	68	Enzymes of unknown specificity	R
7ORF17	UspA	12532-13056	+	E2SD75 ($3X10^{-35}$)	59	Cellular processes (adaptation to atypical conditions)	
7ORF18	Hypothetical	13587-13742	-				
7ORF19	Gamma glutamyl transferase	13675-15594	+	E1VS92 (0)	63	Biosynthesis of co-factors and carriers (glutathione)	E
7ORF20	Pterin-4-α carbinolamine Dehydratase	14613-15903	+	AORWH7 ($6X10^{-21}$)	49	Cellular processes (other)	
7ORF21	Conserved hypothetical	16082-16978	+	E2SDR3 ($2X10^{-78}$)	49	Conserved hypothetical	
7ORF22	TetR	16989-17693	-	E2SDR4 ($5X10^{-65}$)	63	Regulatory (DNA interactions)	K
7ORF23	Conserved hypothetical	17831-18337	+	A1SG67 ($1X10^{-15}$)	33	Hypothetical	

7ORF24	MerR	18408-19229	-	E2SDR5 ($4X10^{-63}$)	50	Transcription factor	K
7ORF25	UPF 0301	19187-19735	+	E2SDR6 ($1X10^{-56}$)	58	Conserved hypothetical	
7ORF26	Conserved hypothetical	19800-20114	+	E2SDR7 ($3X10^{-26}$)	68	Conserved hypothetical	
7ORF27	TypeII restriction subunit /SrmB	20104-21807	+	E2SDR8 (0)	74	Enzymes of unknown specificity	JKL
7ORF28	Carboxyl-esterase	21810-22757	+	($4X10^{-76}$)	52	Protein fate (degradation)	I
7ORF29	Aspartate oxidase	22760-23641	+	E2SDR9 ($8X10^{-92}$)	60	Conserved hypothetical	
7ORF30	Acetyltransferase	23625-24209	+	D3FAY9 ($7X10^{-58}$)	60	Enzymes of unknown specificity	R
7ORF31	Exosortase	24266-27118	+	E2SCE8 (0)	48	Protein fate (degradation)	
7ORF32	YeeE/YeeD	27384-28430	+	C5C7Q5 ($1X10^{-114}$)	61		R
7ORF33	YeeE/YeeD-like	28456-28680	+	E2MVY6 ($2X10^{-29}$)	71		
7ORF34	Luciferase	28805-29851	+	D5PGA8 ($1X10^{-136}$)	72	Enzymes of unknown specificity	C
7ORF35	Membrane protein	29879-30367	-		29	Hypothetical	
7ORF36	Pyrazinamidase	30550-31101	-	F5XQ79 ($9X10^{-56}$)	65	Transport and binding	Q
7ORF37	Nicotinate phosphoribosyl transferase	31104-32438	-	A3TF24 ($1X10^{-164}$)	68	Biosynthesis of co-factors and carriers (pyridine nt's)	H

C

Feature name	Common name	Gene boundary	Coding strand	Homolog (e value)	Sequence identity (%)	TIGRfam role (Sub category role)	COG function
13ORF1	DEAD helicase/ DbpA	327-1826	+	A4XU88 (0)	75	Transcription/ Enzymes of unknown specificity	JKL
13ORF2	Glutamine amidotransferase	2014-2793	+	A4XVH7 ($1X10^{-77}$)	57	Enzymes of unknown specificity	F
13ORF3	RmuC	2781-4334	+	F7NWG1 ($1X10^{-122}$)	51	Unknown function	S
13ORF4	Conserved hypothetical	4331-4777	-	F8H760 ($4X10^{-25}$)	44	Conserved hypothetical	
13ORF5	Hypothetical	4772-4975	+		51	Hypothetical	
13ORF6	WHy	4988-5485	-	A4X566 ($1X10^{-44}$)	54	Unclassified	
13ORF7	UPF 0225	5728-6294	-	F6ABQ0 ($3X10^{-43}$)	53	Transport and binding protein (unknown substrate)	S
13ORF8	DinG	6293-8437	+	F2MXB3 (0)	78	Enzymes of unknown specificity	L
13ORF9	β- lactamase	8466-9611	+	F4DX91 ($1X10^{-126}$)	56	Cellular processes (toxin production and resistance)	M
13ORF10	FolD	9702-10571	-	F2NOD9 ($1X10^{-126}$)	80	Transport and binding protein (unknown substrate)	H
13ORF11	Hypothetical	11305-11526	+			Hypothetical	
13ORF12	SirB	11610-12008	+	ELVLD4 ($9X10^{-17}$)	39	Regulatory function	
13ORF13	CcmA	12426-13064	+	A4XW30 ($3X10^{-83}$)	73	Protein fate (secretion and trafficking)	QPR
13ORF14	CcmB	13061-13732	+	QO2JV8 ($1X10^{-100}$)	83	Protein fate (secretion and trafficking)	O
13ORF15	CcmC	13834-14589	+	A6V819	85	Protein fate (secretion and	O

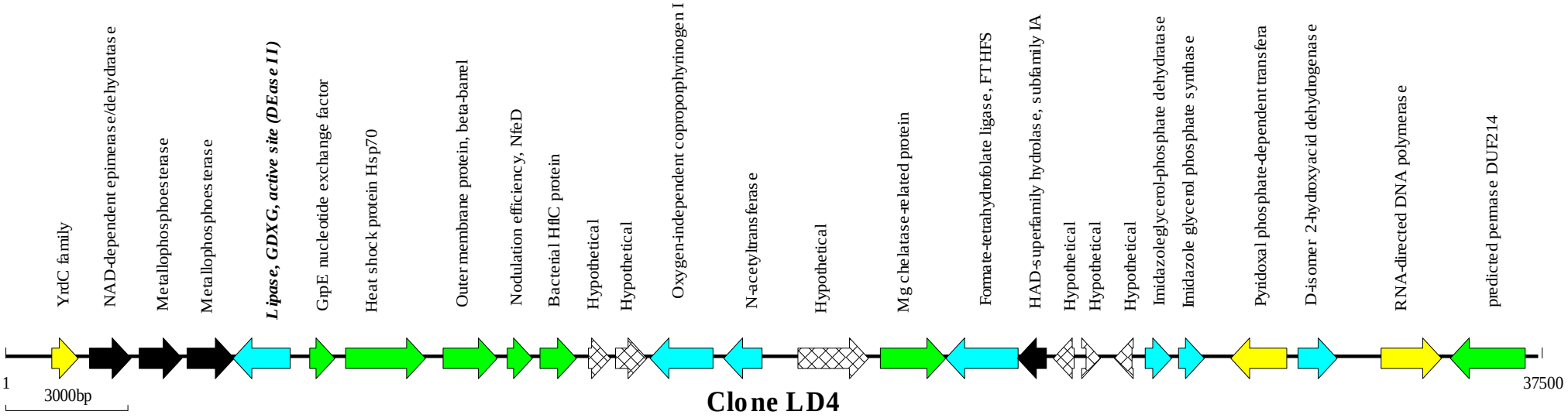
				($1X10^{-126}$)		trafficking)	
13ORF16	CcmE	14759-15226	+	QO2JW1 ($2X10^{-57}$)	74	Energy metabolism (electron transport)	O
13ORF17	CcmF	15223-17226	+	F8H2X5 (0)	77	Energy metabolism (electron transport)	O
13ORF18	CcmG/DsbE	17219-17758	+	F8H2X4 ($2X10^{-63}$)	66	Enzymes of unknown specificity	OC
13ORF19	CcmH	17755-18228	+	F8H2X3 ($4X10^{-48}$)	61	Transport and binding protein (unknown substrate)	O
13ORF20	CycL	18243-19478	+	F8H2X2 ($1X10^{-109}$)	51	Energy metabolism (electron transport)	R
13ORF21	Hypothetical	19608-19790	+			Hypothetical	
13ORF22	GreA/GreB	19801-20148	+			Transcription factor	K
13ORF23	Conserved hypothetical	20581-21333	+	C9QKI1 ($1X10^{-138}$)	95	Conserved hypothetical	
13ORF24	YkkB/invasion gene up-regulator	21444-21968	+	A6F2M1 ($1X10^{-68}$)	73	Enzymes of unknown specificity	J
13ORF25	Membrane protein	22035-22460	+	Q47ZD9 ($1X10^{-49}$)	65	Cell envelope	
13ORF26	Hypothetical	22548-23243	+			Hypothetical	
13ORF27	Conserved hypothetical/Reverse transcriptase	23324-24646	-	Q79RQ7 (0)	97	Mobile element	L
13ORF28	Conserved hypothetical	25468-25857	+	F7RRT1 ($1X10^{-11}$)	33	Hypothetical	
13ORF29	Transposase	26022-26417	+	QOA7P5 ($1X10^{-26}$)	46	Mobile element	
13ORF30	Conserved hypothetical	27147-27626	-				

D

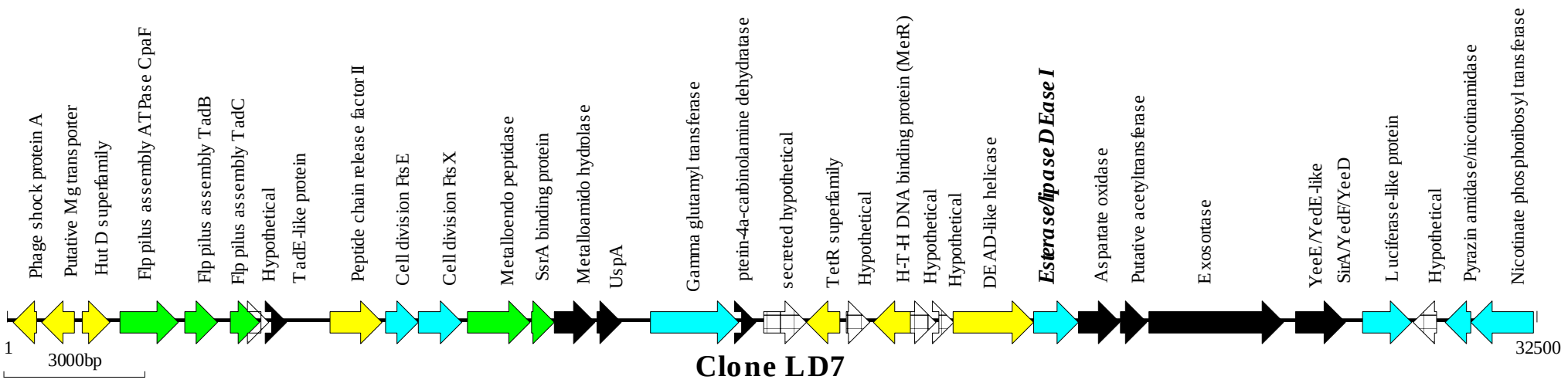
Code	COGs	Domains	Description
Information storage and processing			
J	**245**	10,572	Translation, ribosomal structure and biogenesis
K	**231**	11,271	Transcription
L	**238**	10,338	Replication, recombination and repair
Cellular processes and signaling			
D	**72**	1,678	Cell cycle control, cell division, chromosome partitioning
M	**188**	7,858	Cell wall/membrane/envelope biogenesis
N	**96**	2,747	Cell motility
O	**203**	6,206	Posttranslational modification, protein turnover, chaperones
Metabolism			
C	**258**	9,830	Energy production and conversion
E	**270**	14,939	Amino acid transport and metabolism
F	**95**	3,922	Nucleotide transport and metabolism

H	**179**	6,582	Coenzyme transport and metabolism
I	**94**	5,201	Lipid transport and metabolism
P	**212**	9,232	Inorganic ion transport and metabolism
Q	**88**	4,055	Secondary metabolites biosynthesis, transport and catabolism
Poorly characterized			
R	**702**	22,721	General function prediction only
S	**1346**	13,883	Function unknown

A


YrdC family
NAD-dependent epimerase/dehydratase
Metallophosphoesterase
Metallophosphoesterase
Lipase, GDXG, active site (DEase II)
GmpE nucleotide exchange factor
Heat shock protein Hsp70
Outer membrane protein, beta-barrel
Nodulation efficiency, NfeD
Bacterial HflC protein
Hypothetical
Hypothetical
Oxygen-independent coproporphyrinogen I
N-acetyltransferase
Hypothetical
Mg chelatase-related protein
Formate-tetrahydrofolate ligase, FTHFS
HAD-superfamily hydrolase, subfamily IA
Hypothetical
Hypothetical
Hypothetical
Imidazoleglycerol-phosphate dehydratase
Imidazole glycerol phosphate synthase
Pyridoxal phosphate-dependent transfera
D-isomer 2-hydroxyacid dehydrogenase
RNA-directed DNA polymerase
predicted pemase DUF214
1
3000bp
37500
Clone LD4

B


Phage shock protein A
Putative Mg transporter
Hut D superfamily
Flp pilus assembly ATPase CpaF
Flp pilus assembly TadB
Flp pilus assembly TadC
Hypothetical
TadE-like protein
Peptide chain release factor II
Cell division FtsE
Cell division FtsX
Metalloendo peptidase
SsrA binding protein
Metalloamido hydrolase
UspA
Gamma glutamyl transferase
ptein-4a-carbinolamine dehydratase
secreted hypothetical
TetR superfamily
Hypothetical
H-T-H DNA binding protein (MerR)
Hypothetical
Hypothetical
DEAD-like helicase
Esterase/lipase DEase I
Aspartate oxidase
Putative acetyltransferase
Exosortase
YeeE/YedE-like
SirA/YedF/YeeD
Luciferase-like protein
Hypothetical
Pyrazin amidase/nicotinamidase
Nicotinate phosphoribosyl transferase
1
3000bp
32500
Clone LD7

C

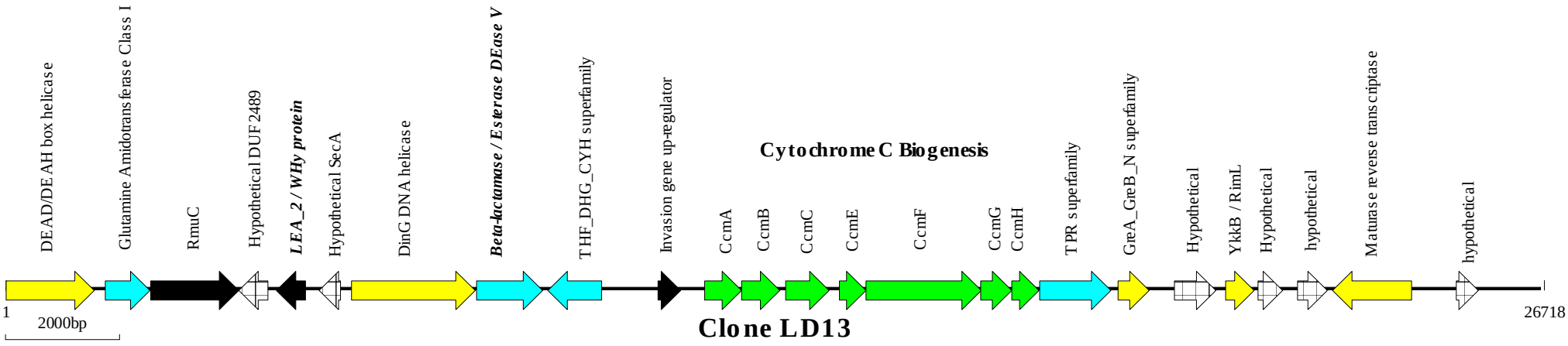


Figure 3.3.1.4: Linear ORF maps of the three fosmid sequences in this study; A] Clone LD4, B] Clone LD7, C] Clone LD13. Maps are generated in DNAMAN and ORF's are colour coded according to main COG categories. Information storage and processing Metabolism Cellular processes and signalling Poorly characterised Hypothetical

3.3.3 Functional annotation

Approximately 70 % of the predicted ORFs could be assigned putative function by homology-based protein sequence searches against the UniProt database. The remaining genes were classified as hypothetical or conserved hypothetical proteins with no sequence homology to proteins of known function. While several genes (*pspA* [7ORF1], *mgtC* [7ORF2], *uspA* [7ORF17], *hutD* [7ORF3]) were subunits of defined operons, only one complete operon was observed in clone LD13, encoding for the cytochrome C biogenesis pathway (13ORF13-13ORF20).

Abiotic factors which may affect organisms from the Antarctic Dry Valley soils include low water potential, low temperatures, increased levels of UV irradiation in summer months, alkaline pH due to the low buffering capacity of the soils and poor nutrient availability. In addition, microorganisms inhabiting this environment experience daily and seasonal fluctuations in environmental conditions. The sample used for this study was soil beneath a seal carcass, which provides a nutrient pool rich in lipids and proteins. Low temperatures increase the solubility of gases, which in turn results in increased reactive oxygen species and ultimately oxidative stress, which would adversely affect most cellular processes. In addition, the stability of toxic metabolites is increased while molecular diffusion rates, fluidity of membranes, and chemical reaction rates decrease with decreasing temperature (Piette *et al*., 2011). All cellular processes are affected by these conditions including protein synthesis, DNA metabolism, transcriptional regulation, and energy metabolism (Figure 3.3.3.1).

It was therefore not surprising to identified proteins involved in nutrient utilisation, cold, desiccation and oxidative stress responses in the fosmid sequences. In theory, all proteins on the three fosmid sequences could possibly be linked to adaptation in the hostile Antarctic environment. Genes encoding general cellular functions may be linked to adaptation by virtue

of the fact that they provide the fundamental requirements to cells inhabiting any environment. For example, the *mgtC* (7ORF2) protein on clone LD7 has been linked to magnesium transport into and out of cells, even though its precise role is not known. Magnesium is required by cells for a variety of reactions including the stabilisation of the rRNA tertiary structure (Kaczanowska and Rudén-Aulin, 2007). Therefore, if this cation is not transported in sufficient amounts into cells, these processes would not occur in an efficient manner, irrespective of the environmental conditions. It may therefore be hypothesised that all proteins would allow for some level of adaptation to the cold, if they functioned with high efficiency at lower temperatures. These parameters can only be investigated and verified by experimental analysis and this may not be possible, mainly due to problems encountered when expressing proteins in a heterologous host. Furthermore, the complete analysis of adaptation mechanisms cannot occur due to the high frequency of conserved hypothetical and hypothetical proteins. Proteomic studies conducted under various conditions have shown that a sizable portion of up-regulated genes have no functional homologs (Bakermans *et al*., 2007; Saunders *et al*., 2005; Guina *et al*., 2003; Goodchild *et al*., 2005; Qui *et al*., 2006). These studies clearly demonstrate that potential novel mechanisms may be employed by microorganisms to survive extreme conditions. For the purpose of this study, those implicated in confirmed adaptation-based functional roles will be discussed in more detail.

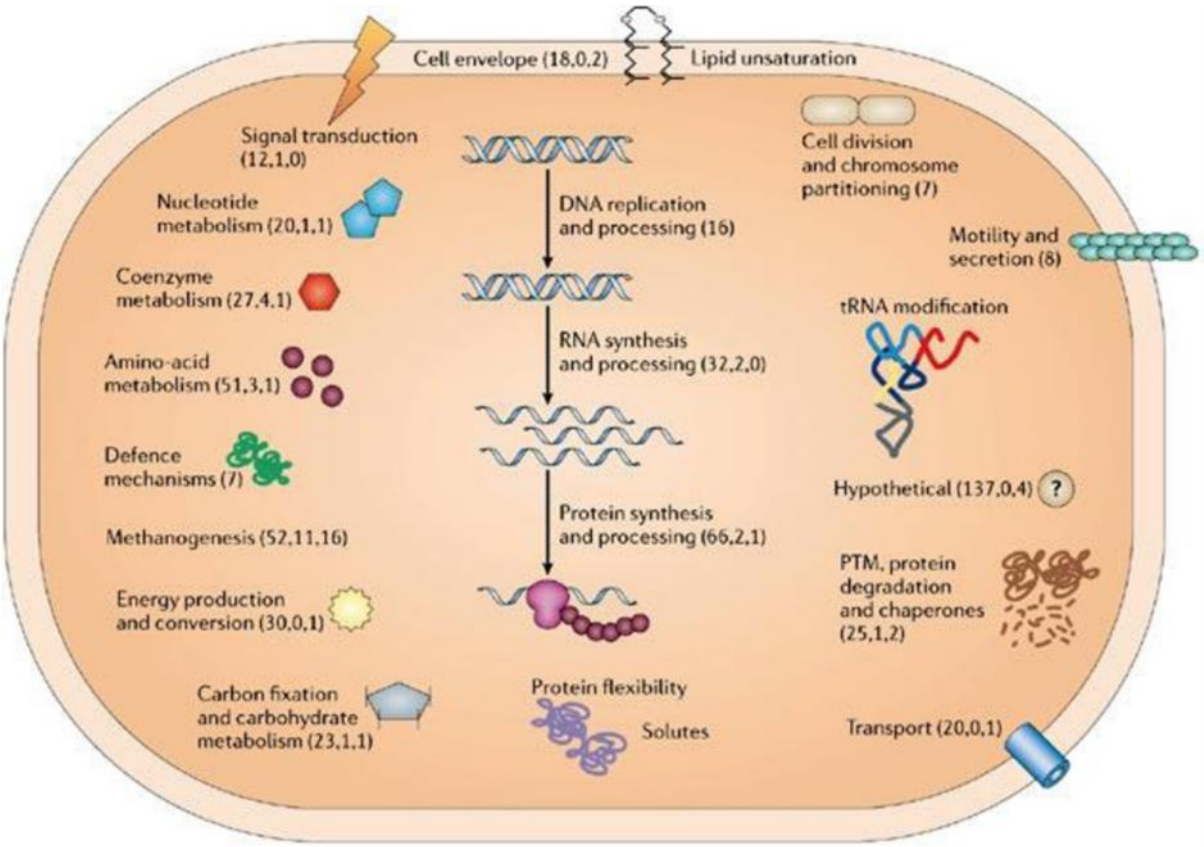


Figure 3.3.3.1: Summary of cellular processes affected by low temperatures (taken from Cavicchioli, 2006).

3.3.3.1 Proteins in COG categories for information storage and processing (Table 3.3.1.4)

Replication, transcription and translation processes may be severely affected in cold environments due to complex inhibitory structures which occur in nucleic acids under these conditions. Protein synthesis, folding and stabilization are additional cellular processes which are cold-sensitive and may restrict microbial growth if disturbed due to cold conditions (Piette *et al.*, 2011). Psychrophiles have clearly adapted coupled mechanisms to optimise replication, transcription and translation under extreme conditions.

Enzymes such as the DEAD box helicase (13ORF1, 7ORF27) and the *dinG* (13ORF8) helicase are members of the very large protein superfamily known as the 'ATPases associated with various cellular activities' or AAA+ proteins. This particular group of proteins is

commonly implicated in cold adaptation and those with ATPase activity can melt short DNA duplexes (Hébraud and Poitier, 1999; Rodrigues and Tiedje, 2008). Enzymes in this superfamily typically contain the characteristic Walker A and Walker B motifs which are essential for nucleotide-Magnesium interactions in P-loop NTPases (Snider and Houry, 2008). The DExH/D family of proteins are ATP dependent remodelling helicases which are involved in all known aspects of RNA metabolism (Kawaoka and Pyle, 2005). Additionally, studies of psychrotrophic microbes such as *Psychrobacter arcticus*, *Methanococcoides burtoni* and *Pseudoalteromonas haloplanktis* have shown that RNA helicases are induced to high levels at low temperatures and that these proteins are required for high capacity translation of structured mRNA's at low temperatures. The *dinG* helicase is involved in nucleic acid metabolism and is both SOS regulated and damage inducible. In *E. coli*, mutants of *dinG* show decreased survival following UV irradiation damage; however, *dinG* is not the sole protein used for repair of disrupted DNA (Voloshin *et al*., 2003). DNA binding proteins are also required for fine tuning bacterial expression in order to deal with frequently changing environmental conditions (Kaczanowska and Rudén-Aulin, 2007). *rimN* (4ORF1) and *dnaK* (4ORF7) are proteins which are important for ribosome assembly. *rimN* plays a role in 30S subunit assembly even though it does not possess any known helicase characteristics. Mutants deficient in *rimN* show accumulation of 17S rRNA leading to inefficient 30S ribosomal subunit assembly, and the mRNA for *rimN* is mainly expressed at low temperatures and may function as a chaperone to facilitate rRNA folding (Kaczanowska and Rudén-Aulin, 2007). This protein can bind double stranded RNA but its function is as yet unknown. Other proteins associated with ribosome assembly include DEAD box helicases, *srmB* and *dbpA* (13ORF1) and mutations in DEAD clearly affect ribosome assembly, due to defects in rRNA processing (Rodrigues and Tiedje, 2008). Assembly of the 50S subunit requires at least three modifications which occur during maturation. The methyltransferase

encoded by *rrmJ* modifies U2552 and when this gene is mutated, severe growth defects occur. This phenotype can, however, be rescued by two different GTPase enzymes (Kaczanowska and Rudén-Aulin, 2007). It can be hypothesised that the GTPases (7ORF11) in the fosmids may act in a similar way, thereby assisting in ribosome assembly.

Both reverse transcriptase (RT) genes and transposons were identified in the sequenced fosmids (4ORF30, 13ORF27, 13ORF29). RT's are RNA-dependent DNA polymerases and are considered to be rare, yet diverse elements within bacterial genomes. They can be classified into three groups; retrons, group II introns and diversity generating retroelements [DGR] (Inouye and Inouye, 1996). These enzymes are responsible for the production of cDNA which, in bacteria, corresponds to a small region of the retron genome (Inouye and Inouye, 1996).

Group II introns are selfish retroelements which appear to provide no advantage to the host organism but are able to persist in environments due to autonomous mobility (Simon and Zimmerly, 2008). DGR's on the hand are not mobile and may instead function to diversify DNA sequences (Medhekar and Miller, 2007). Genome analysis of the cold-adapted bacterium *Methanococcoides burtonii* showed large numbers of mobile genetic elements, which have been further implicated as a possible strategy for cold-adaptation (Karpinets *et al*., 2010; Casanueva *et al*., 2010).

How could the presence of these elements contribute to adaptation in the Dry Valley soils? If an RT element is a DGR, it may mediate a trophism switch which is the case in the microbe *Bordetella pertussis*, where cells alternate between pathogenic and free-living stages (Simon and Zimmerly, 2008). Antarctic organisms, when confronted by nutrient limitations, may use a similar strategy to parasitize other community members for survival. Even a selfish element may assist communities in extreme environments by increasing the amount of HGT

events, possibly transferring adaptive genes in the process (Konstatinidis *et al*., 2009). Additionally, the fusion of C-terminal extensions to RT domains may contribute to diverse and novel biochemical activities (Simon and Zimmerly, 2008). Clustered regularly interspaced short palindromic repeats (CRISPR), which are associated with RT's, have been shown to provide phage resistance to bacterial cells. For example, in the bacterium *Lactococcus lactis*, two systems which encode RT related genes are implicated in defence against phage infection (Simon and Zimmerly, 2008). In an environment where cellular processes are already under strain, phage infection would inflict massive metabolic load on the microbial community. Mechanisms to overcome any additional burden would contribute greatly to the overall fitness of a population. By coupling RT activity with DNA recombination and repair events, genome integrity may also be maintained.

Table 3.3.1.4: Proteins involved in information storage and processing identified in fosmid fragments.

Feature name	Common name	Functional role	Interpro domain
4ORF1	RimN/YrdC/SUA5	Predicted rRNA maturation factor, RNA chaperone	IPR006070
4ORF30	Reverse transcriptase	RNA-directed DNA polymerase, recombination, mobile element	IPR000477
7ORF11	Peptide chain release factor II	GTP-binding protein, mediates UAA and UGA protein termination	IPR020853
7ORF22	TetR	Transcriptional regulation, helix-turn-helix DNA binding motif, control level of susceptibility to antibiotics and detergents	IPR001647
7ORF24	MerR	Transcriptional regulation, helix-turn-helix DNA binding motif	IPR000551
7ORF27	SrmB/DEAD-like helicase	RNA metabolism, gene expression, ribosome biogenesis	IPR014001

13ORF1	DEAD/ DbpA	RNA helicase, ATP dependant nucleic acid unwinding	IPR000629
13ORF 3	RmuC	DNA recombination and repair, function unknown	IPR003798
13ORF8	DinG	DNA /RNA helicase, DNA repair and replication	IPR014013
13ORF22	GreA/GreB	RNA polymerase transcription elongation	IPR018151
13ORF24	YkkB/invasion gene upregulator SirA	Function unknown	IPR016181
13ORF27	Reverse transcriptase	RNA-directed DNA polymerase, recombination, mobile element	IPR000477

3.3.3.2 Proteins in COG categories for cell processing and signalling (Table 3.3.1.5)

Following translation events, protein folding and stabilisation are required to achieve a biologically active conformation. Some determinants of protein folding include burial of hydrophobic regions and interactions, such as H-bonding, van der Waals forces, salt bridges, disulphide bonds and aromatic interactions between residues in the protein (Piette *et al*., 2011). The rate of folding of polypeptides is reduced at lower temperatures making this process susceptible to errors. To combat misfolding and aggregation of proteins, chaperones are generally up-regulated and have a significant contribution to cold adaptation (Piette *et al*., 2011). The first chaperone that interacts with approximately 70 % of nascent polypeptides is the trigger factor (TF). DnaK (4ORF7) and its co-chaperones interact with longer chain polypeptides to assist folding, by cycles of ATP-binding and release (Piette *et al*., 2011; Rodrigues and Tiedje, 2008). In cases where proteins cannot be 'saved', amino acid salvage pathways exist and should be efficient processes both in cold and nutrient-poor conditions. The same theory can be applied to nucleic acid salvage when secondary structures in mRNA cannot be resolved and are targeted for degradation.

Chaperone interactions tend to vary among organisms in cold environments. In *P. haloplanktis*, cold-shock TF is overexpressed while most other heat shock protein (HSP) chaperones are suppressed (Piette *et al*., 2011). Similar observations have been made in *E. coli* and the induction of HSP's reduces the viability of cells grown at low temperatures. In contrast, growth of *P. arcticus* at low temperatures shows up-regulation of the GroEL/GroES chaperones and suppression of TF. In *S. alaskensis*, two DnaK-DnaJ-GrpE clusters are suggested to function independently at both low and high temperatures. Considering that diverse and distinct strategies are utilised by microbes in extreme environments, it is not surprising that these mechanisms may be species specific. In the fosmid sequences, chaperone encoded genes were identified. These chaperones are interesting for further study to determine, for example, whether they assist in protein folding at cold temperatures or if they provide protection from temperature fluctuations and heat denaturation events particularly in the summer months.

Post translational maturation of proteins by insertion of prosthetic groups may be required for the conversion of primary gene products into the mature active form. Co-factor incorporation is linked to the folding and ultimately, stability of polypeptides (Thöny-Meyer, 1997). Cytochromes function as electron transfer proteins in both aerobic and anaerobic respiration. They not only contribute to post translational modification events but also to energy production processes; a valuable function in cold survival. A proton gradient is created by the transport of protons and electrons through the cytoplasmic membrane which when coupled to oxidation of reduced substrates, drives ATP formation (Thöny-Meyer, 1997). CcmE, CcmH, CcmG, and CcmI are considered to be periplasmic Cytochrome C chaperones which may also be involved in pathways relating to pilus assembly, protein secretion and outer membrane protein assembly (Thöny-Meyer, 1997). This operon is well conserved in clone LD13 with

the *ccmABC* and *ccmEFG* clusters typically observed in the γ-proteobacteria (Thöny-Meyer, 1997).

ssrA (7ORF15) and *smpB* (7ORF27) are two proteins that function together to rescue ribosomes stalled on defective mRNA by terminating translation and recycling ribosomal subunits (Nonin-Lecomte *et al*., 2009). SsrA enters the ribosome only when the A-site is empty and contains a stop codon tag. SmpB is an RNA binding protein which binds to SsrA with a high degree of affinity and allows for stable association with the 70S ribosomes (Karzai *et al*., 1999). This trans- translation system allows for survival of bacterial species under adverse conditions, such as growth limitations due to protein synthesis inhibitors (*Synechocystis* spp), increased ethanol concentrations and high temperatures [*Bacillus subtilis*] (Withley and Friedman, 2003). Temperature sensitivity and reduced mobility are phenotypes commonly observed in mutants defective in *ssrA* (Karzai *et al*., 1999). These elements have been identified in the fosmid sequences in this study and may function in a similar manner, to provide another mechanism of adaptation related to post-transcriptional modification. Other proteins for cold adaptation include those for cell division, identified in clone LD7 (7ORF4 – 7ORF6). In *E. coli* cells, mutations of these proteins have been implicated in cold-sensitive phenotypes (Sturgeon and Ingram, 1987).

Horizontal gene transfer and conjugation events are both valuable contributors to genome plasticity and may provide bacteria with fitness in the face of changing environmental conditions (Cascales and Christie, 2003). Type II secretion coupled with Type IV pili (7ORF5 and 7ORF6) is associated with motility and adhesion as well as conjugation. Motility is an important mechanism for adaptation, allowing microorganisms to move to an area which may be more suited to survival. Bacteria may utilise this system to inject proteins or nucleoproteins into other cells or into the extracellular medium (Hazes and Frost, 2008).

Growth of the bacterium *Legionella pneumophila* at low temperatures is promoted by Type II protein secretion, and is most likely due to enhanced protein and/or substrate diffusion across membranes (Söderberg *et al*., 2004). In addition to this system, proteins such as DsbA (13ORF18) and OmpA (4ORF8 and 4ORF9) can be implicated in cold adaptation. Firstly, OmpA provides junction stabilisation between mating cells. Furthermore, this protein may facilitate uptake of nutrients by counteracting low diffusion rates of solutes which occurs at low temperatures. A proteomic study of *Shewanella livingstonensis* showed that OmpA was indeed induced by cold- stress in this organism (Kawamoto *et al*., 2007). The disulphide oxidoreductase, DsbA, functions in mate pair and protein stabilisation by promoting disulphide bond formation where required (Hazes and Frost, 2008).

Another link with the Type II-Type IV secretion system can be made to the phage shock protein A (7ORF1). These two secretion systems have been shown to induce expression of the Psp operon (Lloyd *et al*., 2004). This operon was originally identified in *E. coli* and provided cells with protection against phage infections. The cell envelope controls the influx and efflux of various molecules and provides an ion permeable barrier for the establishment of the proton motive force. Conditions which alter this force include extremes of temperature and osmolarity, altered cytoplasmic membrane properties and biofilm formation (Darwin, 2005). *pspA* is the effector protein for this operon and is known to be involved in membrane functioning and reduction in proton motive force dissipation during a variety of stress conditions, including cold (Lloyd *et al*., 2004). This may assist organisms in Antarctica to adapt to the cold conditions.

Table 3.3.1.5: Proteins involved in cell processing and signalling identified in fosmid fragments.

Feature name	Common name	Functional role	Interpro domain
4ORF6	GrpE	Protein protection for aggregation during stress, co-chaperone for DnaK, nucleotide exchange factor	IPR000740
4ORF7	DnaK	Molecular chaperone, protein protection for aggregation during stress	IPR012725
4ORF8-4ORF9	OmpA outer membrane protein	Multi-functional role, can act as a porin for solute and small molecule exchange, stabilisation of mating cells during conjugation	IPR000498
4ORF18	Mg chelatase	AAA+ protein superfamily, catalyses insertion of Mg^{2+} into protoporphyrin IX for synthesis of (bacterio) chlorophyll	IPR004482
7ORF4	CpaF ATPase domain/ Type II secretion	Involved in general secretory pathway for protein export	IPR001482
7ORF12	FtsE ATP binding domain	ABC transporter, localises to cell division site along with the permease partner	IPR005286
7ORF13	FtsX permease	Localises to cell division site along with the ABC transporter partner, may transport lipids across the inner membrane	IPR003838
7ORF14	Peptidase M23	Amino acid utilisation, proteolytic activity, zinc metallopeptidase	IPR016047
7ORF15	SrmB	SsrA binding protein, ribosome rescue, accuracy and fidelity of protein synthesis	IPR023620
13ORF9	β- lactamase/ esterase	Hydrolysis of β-lactam ring in penicillin and cephalosporins, serine hydrolases	IPR001466
13ORF13-13ORF20	Cytochrome C biogenesis units	Transport, assembly and post translational modification of Cytochrome C. Haem linked to	IPR005895 IPR003544 IPR003557 IPR004329

	aerobic respiration, anerobic respiration with nitrate as electron acceptor. DsbA is a thioredoxin for disulphide bonds in proteins. Periplasmic chaperone activity in some sub units	IPR003568 IPR005746 (DsbA) IPR005616 IPR017560

3.3.3.3 Proteins in COG categories for metabolism (Table 3.3.1.6)

In many natural environments, substrates required for energy, growth and biochemical reactions may be limiting. Survival under these conditions therefore depends on efficient scavenging for scarce resources (Rodrigues and Tiedje, 2008). Even in nutrient rich niches, the ability to utilise a particular resource is vitally important for colonisation and growth.

Slow diffusion rates coupled with an increase in water viscosity in cold environments may hinder the adequate acquisition of nutrients (Rodrigues and Tiedje, 2008). In addition, species occupying the same niche are in competition with each other for the same nutrient pool. A primary strategy to overcome these issues is the production of enzymes with improved enzyme-substrate interaction and reduced activation energy requirements (Rodrigues and Tiedje, 2008). The genetically encoded features of enzyme adaptation have been well documented and will not be discussed in detail here. The key to this long term adaptation strategy is high specific activity at low temperatures, mainly due to conformational flexibity at the active site (D'Amico *et al*., 2006; Caviccioli *et al*., 2002; Gianese *et al*., 2001; Georlette *et al*., 2004).

Lipolytic genes (4ORF5; 7ORF28; 13ORF9) identified in the three fosmid sequences analysed in this study have two possible roles in cold adaptation. Firstly, maintenance of membrane fluidity is essential at low temperatures and the involvement of esterases and lipases in lipid metabolism would contribute to the modification of fatty acid composition in membrane lipids (Morgan-Kiss *et al*., 2006). Secondly, in relation to nutrient acquisition, seal

derived substances constitute a valuable nutrient pool and are the main source of carbon and nitrogen. Microbes which adequately utilise this rich resource would have a competitive advantage to become highly active and successful in the environment. Clearly, the lipolytic genes would play an essential role in this process. Based on this, these proteins are split between two role categories; cell processing and signalling, and metabolism.

An additional role for adaptation can be speculated for the β-lactamase of clone LD13 (13ORF9). This protein is annotated either as an antibiotic resistance gene, or an esterase. The clone is known to have esterolytic activity (See chapter 4) but may be promiscuous and exhibit additional activities. Mineral soils in the Dry Valleys contain both cosmopolitan and indigenous fungi (e.g. *Aspergillus*, *Penicillium*, *Alternaria*) and yeasts (e.g. *Candida*) (Adams *et al*., 2006; Cowan *et al*., 2004). These organisms have the potential to produce antibiotic secondary metabolites, such as β-lactam containing compounds, which would provide a competitive advantage. The ability to counteract the deleterious effects of these bactericidal agents would be highly beneficial to bacterial community members. Another adaptive mechanism found in natural environments is enzyme promiscuity itself. The presence of genes encoding multifunctional proteins allows organisms to utilise enzymes for a variety of processes, thereby possibly reducing the energetic cost of multiple enzyme production with a single function. Another example of multifunctional activity in the fosmid LD13 is the FolD protein (13ORF10), which catalyses C-N bonds of various substrates.

Table 3.3.1.6: Proteins involved in metabolism identified in fosmid fragments.

Feature name	Common name	Functional role	Interpro domain
4ORF5	Lipase	Lipid metabolism	IPR002168
4ORF10	NfeD	Function unknown. Part of peptidase superfamily S49 as non-peptidase members	IPR002810
4ORF11	Band 7 protein/ HflC	Integral membrane protein. Possible role in cation conductance, function not described	IPR001107
4ORF14	Coprophorphyrinogen III Oxidase /HemN	Replaces HemF function under anaerobic conditions, tranformations during porphyrin biosynthesis, linked to (bacterio) chlorophyll	IPR004558
4ORF15	N-acetyl transferase/ transglutamase	Facilitates transfer of acetyl group from acetyl-coA to a range of arylamines and hydrazines	IPR001447
4ORF19	Formate tetrahydrofolate ligase	Multifunctional protein, transfer of one carbon units essential in many biosynthetic pathways	IPR000559
4ORF24	Imidazoleglycerol-phosphate dehydratase	Catalyses seventh step in histidine biosynthesis	IPR000807
4ORF25	Imidazoleglycerol-phosphate synthase	Fifth step in hisidine biosynthesis provides substrate for *de novo* purine biosynthesis. Links amino acid and nucleic acid synthesis pathways	IPR010139
4ORF28	D-isomer specific 2-hydroxyacid dehydrogenase	Oxidation and reduction processes. Specific for D-isomer of substrate	IPR006140
4ORF27	Aminotransferase	Pyridoxal-phosphate dependent enzyme. Important in carbon and nitrogen metabolism	IPR004839
7ORF1	Phage shock protein A	Negative regulator of Psp operon, maintains proton motive force. Interacts directly with PspF transcriptional activator	IPR007157
7ORF19	Gamma-glutamyl transpeptidase	Transfer of glutamyl moiety of glutathione forming glutamate. Drug and xenobiotic	IPR000101

		degradation. Autolytic peptidase of MEROPS family	
7ORF28	Lipase/esterase of α/β hydrolase fold	Lipid metabolism	IPR013094
7ORF34	Luciferase	Flavin monooxygenase. Oxidises long-chain aldehydes and releases energy in the form of visible light	IPR011251
7ORF36	Isochorismatase-like	Conversion of isochorismate to pyruvate	IPR000868
7ORF37	Nicotinate phosphoribosyl transferase	Catalyses first step in NAD salvage pathway	IPR015977
13ORF2	Glutamine amidotranferase/GMP synthase	GATase enzymes catalyse the formation of GMP from Xanthosine 5'-phosphate and glutamine for pyrimidine synthesis.	IPR017926
13ORF10	Tetrahydrofolate dehydrogenase	Multifunctional enzyme, acting on carbon and nitrogen bonds in substrates other than peptides	IPR000672

3.3.3.4 Proteins not assigned COG categories

In carbon starved bacteria, *de novo* production of macromolecules must occur in order to maintain endogenous metabolism (Nyström and Gustavsson, 1998). In growth arrested cells, oxidised proteins and macromolecules may increase oxidative stress within cells. The universal stress protein is induced under conditions of nutrient starvation, antibiotic and oxidant exposure, while being repressed under temperature extremes (Kvint *et al*., 2003). In *E. coli*, 6 Usp paralogs exist (A,C,D,E,F,G), and cold shock represses UspA (7ORF17) expression due to a reduction in the levels of the alarmone ppGpp (Kvint *et al*., 2003). Although this particular protein does not show a clear role in cold adaptation, it may mark proteins for degradation and is important to oxidative stress protection, a severe abiotic stress in Antarctic Dry Valley soils.

HutD (7ORF3) is part of the histidine uptake and utilisation operon. This is a typical system dedicated to amino acid metabolism (Zang and Rainey, 2007). HutD is required for efficient utilization of histidine as a sole carbon and nitrogen source and is suggested to play a regulatory role, possibly as governor to Hut transcription. HutD controls the intracellular levels of the Hut inducer, urocanate, thereby preventing excessive ammonia concentrations which may be potentially harmful to cells (Zang and Rainey, 2007). This protein may have an adaptive function by not only detoxifying harmful chemicals but also by allowing cells to metabolise the seal-derived amino acids for energy and protein synthesis which are in turn derived from further degradation of glutamate (Zang and Rainey, 2007). The latter function connects this pathway with two other genes in clone LD7, the gamma glutamyl transferase (7ORF19) and helicase/ SrmB (7ORF27). Transfer of the gamma-glutamyl subunit of glutathione to an acceptor molecule produces glutamate. Glutamate is a known osmoprotectant and compatible solute. In addition, the gamma-glutamyl cycle for synthesis and degradation of glutathione also functions in drug and xenobiotic degradation (Siest, 1992). A further indirect effect of this protein is linked to the helicase proteins, which contain higher levels of histidine residues, important for function.

Another protein found on clone LD13 shows potential in both cold adaptation and desiccation resistance. The WHy protein (13ORF6) is part of the hydrophilin/ dehydrin group of proteins which have been found to be up-regulated during both of these conditions. However, the function of these proteins remains largely unknown. WHy is discussed in detail in Chapter 5.

3.4 Conclusion

Sequence assembly of four fosmid clones together with gene calling and annotation was successfully accomplished, and has been described in this chapter. Furthermore, key functions possibly related to adaptation were identified. This study provides a number of testable hypotheses to link genetic and metabolic potential to environmental functioning. The success of this particular section of work relied on determining strengths and weaknesses of freely available bioinformatic tools, and using a combination of approaches in order to elucidate gene functions for given fragments. As sequencing technologies advance and become less expensive, researchers can expect an explosion in novel sequence space, and ultimately the discovery of currently unknown functions, activities and adaptation mechanisms. Results obtained in this study show consistency with other –omic based studies which regularly assign genes associated with low temperature survival to three major cellular processes; transcription, translation and ribosome maintenance, maintenance of membrane integrity and fluidity, as well as energy metabolism (Reva *et al*., 2006; Methé *et al*., 2005; Piette *et al*., 2011; Allen *et al*., 2009; Cacace *et al*., 2010). In addition, this study also highlights the need to assess the potential contained in the vast amount of 'hypothetical' proteins identified from such research.

Chapter 4: From genetic potential to function – Lipolytic genes.

4.1 Introduction

4.1.1 Lipolytic enzymes

Lipid macromolecules play significant physiological roles in all living systems by participating in energy storage and cell signalling processes (Gilham and Lehner, 2005; Hasan *et al*., 2006, Sangeetha *et al*., 2011), and are essential structural components of the cellular membranes of prokaryotes and eukaryotes. The turnover of lipid biomass within cells and in the extracellular milieu is mediated by lipolytic enzymes (Hasan *et al*., 2006). Lipolytic enzymes are ubiquitous in nature being widely distributed in plants, animals and microbes, and include lipases (E.C 3.1.1.3) and esterases (E.C 3.1.1.1). Lipases exhibit a broad substrate range with preference for longer chain fatty acid molecules > C10. They can hydrolyse triglycerides to diglycerides, monoglycerides, glycerol and fatty acids. Esterases preferentially catalyse the hydrolysis of short chain esters (< C10) which are partly soluble in water (Arpigny and Jaeger, 1999; Jaeger *et al*., 1999; Fojan *et al*., 2000; Gilham and Lehner, 2005; Sangeetha *et al*., 2011). In addition to the valuable biological role that these enzymes play in the synthesis and hydrolysis of lipid biomolecules, they are also recognised as important biocatalysts with applications in a variety of industrial and biotechnological processes (Rosenau and Jaeger, 2000; Sangeetha *et al*., 2011).

Three dimensional structures of both enzyme classes have been solved, demonstrating a definite order of α- helices and β-sheets, known as the α/β hydrolase fold (Figure 4.1.1). The α/β hydrolase fold is the stable scaffold for numerous hydrolase enzymes, including the serine carboxypeptidases (E.C 3.4.16.X), dehalogenases (E.C 3.8.1.X) and acetyl cholinesterases (E.C 3.1.1.7) (Jaeger *et al*., 1999; Angkawidjaja and Kanaya, 2006). Enzymes classes of the α/β hydrolase fold family are expected to contain novel structural features which surround the central core thereby imparting a range of differing substrate specificities (Nardini and Dijkstra, 1999). This structural fold consists of 8 β-sheets, with

β3 and β8 connected by α- helices packed on either side of the central parallel β-sheet (Figure 4.1.1) (Jaeger *et al*, 1999; Bornscheuer, 2002). In addition to the common folding pattern, lipolytic enzymes also exhibit a triad of catalytic residues composed of Ser-Asp-His that constitute the active site. The serine nucleophile is positioned in the nucleophilic elbow, generally located in a conserved pentapeptide sequence, GXSXG, located in the middle of the gene (Nardini and Dijkstra, 1999; Arpigny and Jaeger, 1999). It has, however, been shown (Upton and Buckley, 1995; Arpigny and Jaeger, 1999) that not all lipolytic enzymes contain this GXSXG consensus motif. For instance, the family II GDSL esterases/lipases contain a GDS(L) motif located closer to the N-terminal of the protein and represent examples of a catalytic diad, where the third residue is not readily identified (Upton and Buckley, 1995; Jaeger *et al*., 1999). The geometry of this region contributes to the formation of the oxyanion binding site which is required for stabilisation of negatively charged transition states formed during hydrolysis (Nardini and Dijkstra, 1999).

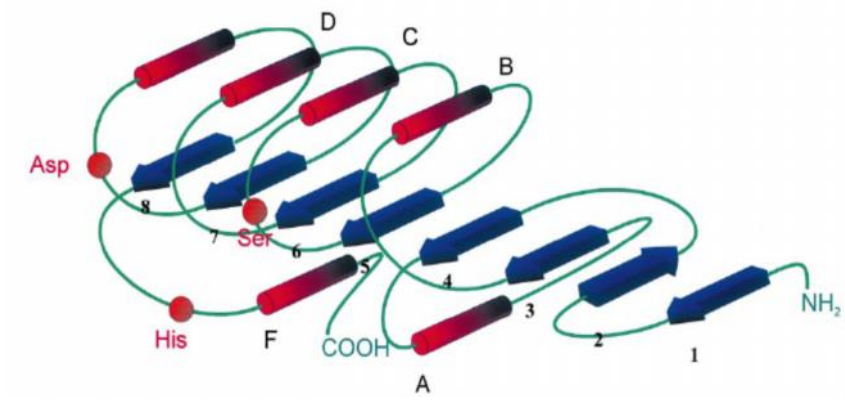


Figure 4.1.1: The canonical structure of the ~/β hydrolase fold. α-Helices (A-F) are shown as red cylinders and β-sheets (1-8) as blue arrows. Solid orange circles indicate the topological position of active site residues (Catalytic serine after β-5, Asp/Glu after β-7 and the histidine residue in the loop region between β-8 and α-F). Taken from Bornscheuer, 2002.

Lipases and esterases are generally distinguished from each other on the basis of substrate specificity and interfacial activation; the observed phenomenon of increased enzymatic activity in the presence of lipid: water interfaces (Jaeger *et al*., 1999; Verger, 1998). In the case of most lipases, a movable lid may exist with at least one of the oxyanion residues. In these cases, the correct orientation of the binding site is only attained when the protein is in an open conformation, during which time the active site becomes available for substrate binding (Nardini and Dijkstra, 1999).

The process of interfacial activation occurs due to the presence of a lid, consisting of a single or double helix, or a loop region, which covers the active site in the absence of a lipid: water interface. The presence of hydrophobic substrates initiates conformational rearrangement of the active site, making catalytic residues more accessible and thereby increasing activity of the lipase (Jaeger *et al*., 1994; Verger, 1998). However, a number of exceptions exist and these lipases display neither the 'lid' nor interfacial activation, thereby making this criteria for distinction untenable (Verger, 1998). Esterases follow classical Michaelis-Menten kinetics, where activity is a function of substrate concentration and the maximal rate is achieved at substrate saturation (Jaeger *et al*., 1994; Bornscheuer, 2002). Lipases (with some exceptions) exhibit increased enzymatic activity on emulsions (insoluble substrates) when compared to monomeric (soluble) solutions of the same substrate (Jaeger *et al*., 1994; Verger, 1998).

4.1.2 Bacterial lipolytic families

Arpigny and Jaeger (1999) were the first to classify 53 bacterial lipolytic enzymes into 8 families based on their fundamental biochemical properties, characteristic signatures in amino acid sequence and X-ray crystallography data.

4.1.2.1 Family I

Enzyme members in this family are generally defined as 'true' lipases and are further divided into several sub-families comprising mostly *Pseudomonas*, *Bacillus* and *Staphylococcus* lipases (Sangeetha *et al*., 2011). Subfamilies I.1 and I.2 contain the previously described Pseudomonad group I and II lipases which are encoded together with their respective chaperones (Lipase dependent foldase) in an operon unit (Jaeger *et al*., 1994; Rosenau and Jaeger, 2000). Enzymes in the subfamily I.2 are generally larger than those in subfamily I.1 due to an additional sequence insertion, resulting in a double anti-parallel β-strand at the molecule surface (Arpigny and Jaeger, 1999). Both subfamily enzymes are secreted via the bacterial type II pathway. This is a secretion (sec) dependant terminal branch of the general secretory pathway and is closely related to the biogenesis pathway for type IV pili (Henderson *et al*., 2004; Thanassi and Hultgreen, 2000). Another distinguishing feature of subfamilies I.1 and I.2 is the presence of two aspartic residues involved in a Ca^{2+}-binding site. Additionally, two disulphide bridge-forming cysteine residues are conserved and, together with the aspartate residues, are believed to assist in active site stabilization (Kim *et al*., 1997).

Lipases belonging to subfamily 1.3 have a large molecular mass, no disulphide bridge-forming cysteine residues, and the notable absence of an N-terminal signal peptide. These proteins do, however, have a C-terminal secretion signal and are translocated via the Type I protein secretion pathway (Henderson *et al*., 2004; Thanassi and Hultgreen, 2000).

Subfamily I.4 comprises the *Bacillus* lipases. Enzymes from *B. subtilis* and *B. pumillus* are considered to be the smallest lipases and share very little sequence identity to other true lipases. Additionally, an alanine residue replaces the first glycine amino acid in the GXSXG motif (Arpigny and Jaeger, 1999). *Geobacillus* lipases dominate subfamily I.5, while a group of large enzymes with alkaline pH optima and thermotolerant characteristics

occur in the subfamily I.6 lipases. Those members that originate from *Staphylococcus* species, are secreted as pro-peptides and, only once translocated across the cellular membrane, are processed into mature protein by specific proteases (Arpigny and Jaeger, 1999).

4.1.2.2 Family II

Enzymes belonging to this family are characterised by the signature GDSL conserved motif, which houses the catalytic serine residue. Furthermore, these enzymes have four strictly conserved residues, Ser-Gly-Asn-His, in four conserved blocks; I, II, III and IV, respectively. Each block plays an essential role in the catalytic function of these enzymes (Akoh *et al*., 2004). Due to the absence of the nucleophile elbow and a different tertiary fold structure, these enzymes are not members of the α/β hydrolase-fold superfamily but rather belong to the SGNH hydrolase superfamily (Akoh *et al*., 2004; Jaeger *et al*., 1999; Verger, 1998). These enzymes also share very little sequence homology to true lipases and some members exhibit protease, arylesterase and thioesterase activity (Arpigny and Jaeger, 1999; Bornscheuer, 2002; Akoh *et al*., 2004). Another characteristic feature of these enzymes is a covalently bound C-terminal translocator unit, which consist of 250-300 amino acid residues. This autotransporter domain forms a β-barrel pore in the outer membrane of Gram negative bacteria and allows for translocation of the linked passenger domain via the Type V secretion pathway (Jacob-Dubuisson *et al*., 2004). Proteins secreted via autotransporter systems typically contain N-terminal signal peptides allowing for their targeting and transport across the inner membrane via the GSP system (Henderson *et al*., 2004; Jacob-Dubuisson *et al*., 2004). Once transported across the inner membrane, the pro-protein exists as a periplasmic intermediate, where partial folding of the autotransporter may occur. This intermediate protein is also accessible to periplasmic enzymes (Henderson *et al*., 2004).

4.1.2.3 Family III

Enzymes from family III display the typical canonical fold of α/β-hydrolases. These extracellular enzymes contain the conserved catalytic triad and share approximately 20 % amino acid homology with mammalian platelet-activating factor acetylhydrolase (PAF-AH) (Arpigny and Jaeger, 1999). PAF-AH's are intracellular phospholipases which generally contain sequence insertions, suggesting the presence of a movable loop.

4.1.2.4 Family IV

Bacterial lipolytic enzymes in this family share high levels of sequence similarity to mammalian Hormone Sensitive Lipases (HSL). A signature HGGG motif comprising part of the oxyanion hole is strictly conserved and occurs at the N-terminus. This motif is involved in stabilisation of the oxyanion hole via hydrogen bond interactions and is believed to assist in the hydrolysis of ester bonds in tertiary alcohols due to a larger active site which accommodates the alcohol moiety (Virk *et al*., 2011; Hencke *et al*., 2003; Atomi, 2004). HSL family esterases have been utilised in industry for plant cell wall degradation as well as a variety of esterification reactions, particularly in the resolution of racemic mixtures (Panda, 2005). Enzymes from thermophiles, mesophiles and psychrophiles are represented in this family (Arpigny and Jaeger, 1999; Sangeetha *et al*., 2011).

4.1.2.5Family V

These enzymes share high sequence homology to a variety of bacterial non-lipolytic proteins, such as dehalogenases and haloperoxidases. These enzymes contain both the catalytic triad and signature α/β-hydrolase fold as well as a PTX_4GX_2A motif which precedes the active site aspartate residue (Arpigny and Jaeger, 1999).

4.1.2.6 Family VI

This family represents the smallest known esterases with a molecular mass ranging from 22 – 26 kDa. Dimerised forms of these enzymes are the active biocatalysts and these enzymes do show a high level of homology to eukaryotic calcium independent phospholipases.

4.1.2.7 Family VII

This group of esterases have high molecular mass, usually in excess of 55 kDa and share amino acid identity with acetylcholine esterases. One member of this family is a plasmid encoded esterase from *Arthrobacter oxydans* and has been shown to be active against phenylcarbamate containing herbicides (Sangeetha *et al*., 2011).

4.1.2.8 Family VIII

The enzymes represented in this family contain an N-terminal consensus motif that differs from all other lipolytic enzymes. This SXXL motif is shared with members of the bacterial β-lactam antibiotic resistance proteins. Interestingly, some studies on these enzymes have shown promiscuous activity on both lipid substrates as well as β-lactam antibiotics (Rashamuse *et al*., 2009). Other reports, however, demonstrate esterolytic activity but not cleavage of the β-lactam ring of compounds such as loracarbef and cefamandole (Eland *et al*., 2006).

4.1.3 Biotechnological application of lipolytic enzymes

According to a global strategic business report, the worldwide market for industrial enzymes is set to exceed $ 2.9 billion within the next two years (Sangeetha *et al*., 2011). This market is dominated by carbohydratases, proteases and lipases. In industrial processes, a high level of control over the products being manufactured can be achieved by exploiting the specificity exhibited by enzymes. In addition, increased global awareness of environmental and economic issues have become key drivers for the use of enzymes over chemical catalysts in industry due to the fact they are biodegradable, contribute minimal biological oxygen demand (BOD) in waste streams and reduce unwanted side reactions in manufacturing procedures. Enzymes of microbial origin are useful in many industrial processes due to their stability, higher product yield and sustainable supply (Hasan *et al*., 2006; Joseph *et al*., 2008). Furthermore, microorganisms adapted to extreme environmental

conditions potentially produce enzymes which may exhibit the unique desired characteristics required for harsh industrial processes (Hasan *et al*., 2006).

Lipolytic enzymes exhibit unique regio- and enantio- selectivity properties and are able to catalyse a variety of reactions such as esterification, acidolysis, aminolysis and trans-esterification (Figure 4.1.3.1) (Sangeetha *et al*., 2011; Joseph, 2008). In addition, these enzymes display a broad substrate range, are stable in organic solvents, show no co-factor dependence for catalytic hydrolysis and may be purified in large quantities (Hasan *et al*., 2006). They are valuable catalysts with a wide range of applications in the food, detergent, pharmaceutical and organic chemistry industries. For the purposes of this review, some applications will be discussed. For further reading, a number of key papers extensively review this topic (Hasan 2006, Jaeger and Eggert, 2002, Gandhi 1997, Vileneuve *et al*., 2000, Jaeger *et al*., 1997, Joseph *et al*., 2007).

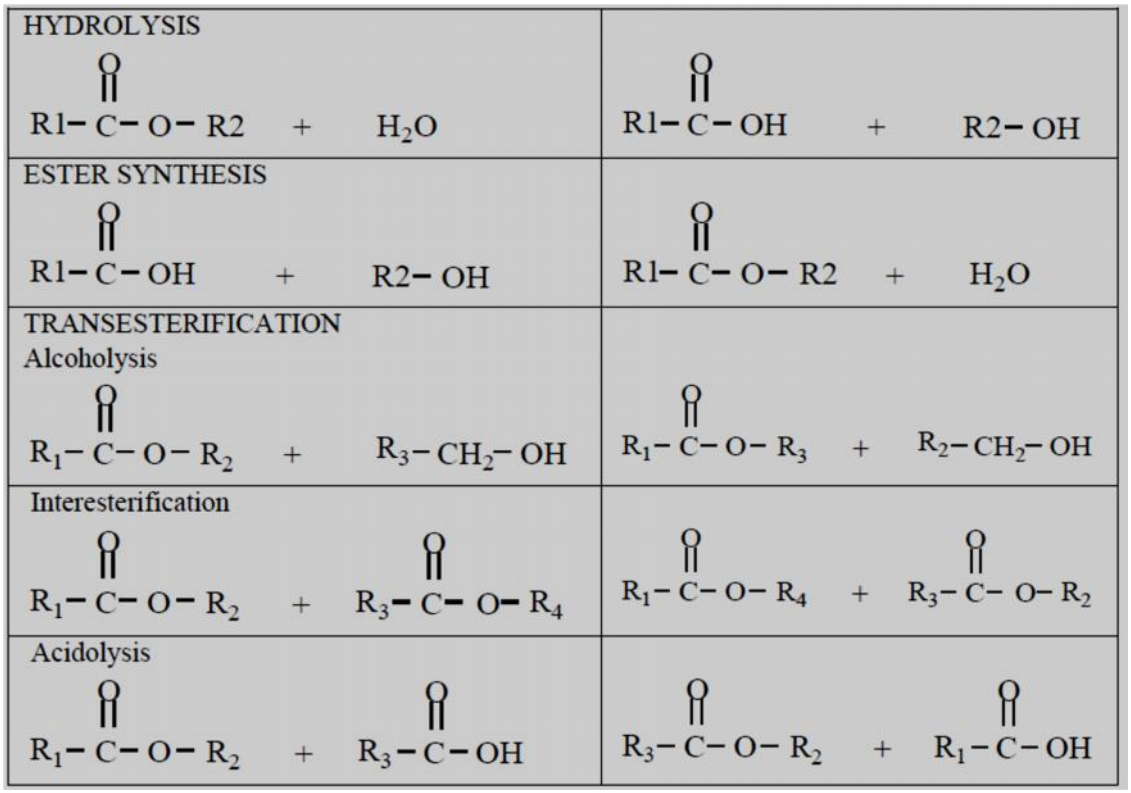


Figure 4.1.3.1: General reactions catalysed by lipases and esterases.

4.1.3.1 Lipolysis

Lipolysis is the catabolism of fats or esters into the constituent acid and alcohol/glycerol in the presence of water. Microbial carbohydrate esterases are employed in the hydrolysis of pectin or xylan in plant cell walls to liberate ferulic acid and increase the potential yield of C_5 and C_6 sugars from lignocellulose feed stocks (Aurilia *et al*., 2008; Bornscheuer, 2002). In the leather industry, residual fats and protein debris associated with hair and hides are removed through the action of lipases (Hasan *et al*., 2006). In activated sludge and aerobic waste treatment, thin layers of fat must be removed to permit oxygen transport. Lipases are used to degrade the lipid-rich liquid that is skimmed from these systems (Gandhi, 1997).

Lipases are also utilized extensively in the food industry. Conventional chemical processing of fats and oils requires harsh temperature and pressure conditions which result in undesirable side reactions such as decolourisation, odour and oxidation of fatty acids (Jaeger *et al*., 1994; Gandhi, 1997; Jaeger and Eggert, 2002). Accelerated flavour development occurs when free fatty acids, soluble peptides and amino acids are formed in the maturation stages of a dairy product (Hasan *et al*., 2006). As such, hydrolysis of milk fat in dairy products results in flavour enhancement in certain cheeses, particularly soft cheese. Lipases impart rich creamy flavours to coffee whiteners, caramels and toffees, and chocolate (Gandhi, 1997). One of the most important applications of lipases in industry is the resolution of racemic mixtures for the synthesis of chiral building blocks for pharmaceuticals and agrochemicals (Jaeger *et al*., 1994; Gandhi, 1997; Hasan *et al*., 2006). In the pharmaceutical industry, enantiomeric forms of particular drug intermediates are key factors for efficacy. For example; ketoprofen is a non-steroidal anti-inflammatory drug known to inhibit production of prostaglandins. It is the (S) enantiomer which is responsible for the desired effects and, as such, lipases can be tested for the production of the correct enantiomer (Kang *et al*., 2011). Other applications that

utilise the hydrolytic power of lipases and esterases include bioremediation of petroleum based hydrocarbons and biodiesel production (Hasan *et al*., 2006; Gandhi, 1997).

4.1.3.2 Ester synthesis

Lipolytic enzymes can catalyse the reverse reaction and in the process, liberate water. In low water activity systems, the normal hydrolytic equilibrium can be reversed in favour of esterification reactions (Jaeger *et al*., 1994; Sharma *et al*., 2001). Acidolysis, interesterification and alcoholysis reactions give rise to acids, esters or alcohols instead of water (Gandhi, 1997).

This ability of lipases is important in oleochemical processes where less useful fats may be converted to more nutritionally valuable ones (Hasan *et al*., 2006). Interesterification reactions have been applied for the conversion of palm oil into cocoa butter, a high value product used in food, confection, pharmaceuticals and the cosmetic industry (Gandhi, 1997; Sharma *et al*., 2001; Hasan *et al*., 2006).

4.1.4 Cold-active lipolytic enzymes

Permanently cold habitats exert high selective pressure on the resident microbial population. Organisms colonising these environments have developed adaptative strategies facilitating survival under extreme physiochemical conditions. In the cell membrane, for example, tailoring acyl chains and increasing the ratio of polyunsaturated fatty acids reduces the phospholipid melting point. This increases membrane fluidity of cells, allowing for appropriate exchange of solutes between cells and the external medium (Nichols *et al*., 1993; Gerday *et al*., 1997). Since low temperatures can slow down or even inhibit biochemical reactions, enzymes are under strong selective pressure for adaptation under conditions of extreme cold (D'Amico *et al*., 2002; Gerday *et al*., 2000).

Lipolytic enzymes produced by cold-adapted microbes have evolved structural features which confer a high degree of flexibility around the active site. This results in low activation enthalpy and high specific activity at low temperatures. This flexibility is often a result of multiple sequence and structural changes, outlined in Figure 4.1.4.1.

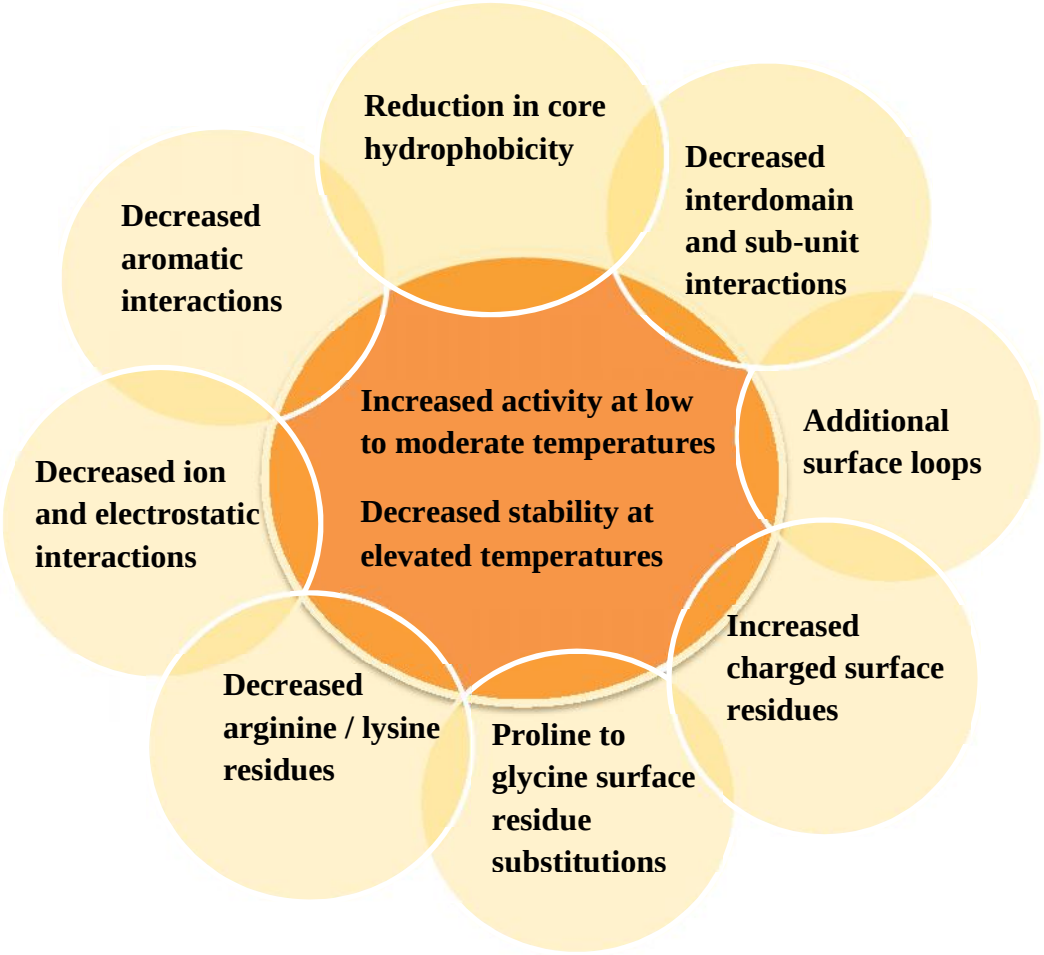


Figure 4.1.4.1: Structural modification for increased enzyme flexibility include modifications such as a lower number of arginine residues compared with lysine, low proline content, particularly in loop regions, increased clustering of glycine residues, a small number of salt bridges and aromatic-aromatic interactions, a decrease in the number of ionic interactions and hydrogen bonds as well as weakening of hydrophobic clusters (Joseph *et al.*, 2007; Joseph *et al.*, 2008; Rodrigues and Tiedje, 2008; Russell, 2000; Feller and Gerday, 2003).

4.1.4.1 Applications of cold-active lipolytic enzymes

Cold-adapted enzymes generally exhibit high catalytic activity at low temperatures when compared to their mesophilic homologs (Feller, 2003). The use of cold-adapted enzymes in certain processes may eliminate unwanted side reactions, and reduces energy consumption and environmental impact (Joseph *et al*., 2008). For example, rapid inactivation of heat liable enzymes increases mechanical resistance of fabrics, while cold washing reduces wear and tear (Hasan, 2006). The use of cold adapted lipases is particularly important in processes where products are highly susceptible to heat degradation (Kumar *et al*., 2011). In bioremediation schemes, seasonal fluctuations influence the effectiveness of the enzyme preparation for pollutant degradation, but a preparation of both mesophilic and psychrophilic enzymes may enhance the process due to combined activity over a wider temperature range (Joseph *et al*., 2007).

4.1.5 Aims and objectives

5. Identify possible lipolytic genes from fully assembled fosmid clones
6. Clone candidate genes, verify their respective enzyme activities and confirm the accuracy of bioinformatic predictions
7. Overexpress and kinetically characterise lipolytic enzymes
8. Perform comparative genomic study to identify possible sequence and structure modifications for cold-adaptation

4.2 Materials and methods

4.2.1 Bioinformatic analysis

Open Reading Frame prediction software (SoftBerry and Glimmer) was used in combination with BLASTp (Altschul *et al*., 1997) and UniProt searches to annotate fosmid sequences. Putative lipolytic genes were subsequently identified in the sequenced fosmid clones. Preceding the cloning of these genes, ORFs were analysed for restriction sites (DNAMAN), the presence of N-terminal signal peptides (SignalP 3.0) (Emanuelsson *et al*., 2007), and rare codon content (RareCodon Caltor). Protein sequences were searched against the UniProt database for homologous relatives. Multiple sequence alignments using ClustalW (Larkin *et al*., 2007) were used to determine conserved catalytic regions in the genes. InterproScan was used to find matches to the predicted protein based family domains (Finn *et al*., 2008).

The translated nucleotide sequences were used for homology modeling using the Swiss-protein modeler program (Schwede *et al*., 2003) PDB-sum was used to assess the accuracy of the models by generating individual Ramachandran plots (Lovell *et al*., 2002). Models were superimposed into the defined templates using the PyMol program.

For phylogenetic analysis, sequences representing 8 lipolytic families were retrieved from the NCBI database and aligned together with lipolytic gene sequences from the fosmid clones, using ClustalW. A neighbour joining tree was constructed using the CLC genomics workbench with 1000 replicates.

For comparative genomic studies, several homologous mesophilic or psychrophilic homologous sequences were retrieved from the UniProt database. General protein characteristics such as amino acid content, Grand Average Hydropathy (GRAVY) and

aliphatic index were predicted using the ProtParam tool (Expasy). The percentage of amino acid residues believed to contribute to psychrophilicity, were compared manually.

4.2.2 Transposon mutagenesis

Transposon mutagenesis was performed using the HyperMu™ <KAN-1> Insertion kit (Epicentre Biotechnology, USA) according to the manufacturer's specifications. Briefly, 600 ng of fosmid DNA was added to a reaction containing 1 × HyperMu reaction buffer, 12.5 ng HyperMu <KAN-1> transposon, and 0.5 U of HyperMu MuA transposase. The mixture was incubated for 2 hours at 37 °C and stopped by adding 1 × HyperMu stop solution, followed by heating at 70 °C for ten minutes. Two microliters of the mixture was electroporated into competent GeneHog *E. coli* cells and, following recovery at 37 °C for an hour, transformants were plated onto tributyrin agar supplemented with kanamycin and chloramphenicol. Plates were incubated at 37 °C overnight and transferred to 4 °C for a further 5 days. Plates were monitored daily for tributyrin hydrolysis. Clones which did not produce clearance zones were selected, fosmid DNA was extracted and were sequenced using the HyperMu primer pair.

4.2.3 Sub-cloning lipolytic genes

Putative lipolytic genes were amplified from the corresponding fosmid template using 30 cycles of PCR with specific primer pairs (Chapter 2, Table 2.1.2) in the Applied Biosystems Thermocycler Gene Amp® 2700. PCR conditions, as well as the expected molecular mass of the four amplicons are provided in Table 4.2.1. All primers contained NdeI and XhoI recognition sites. Purified PCR products were digested with both restriction enzymes and a small aliquot was visualised on a 1 % [w/v] agarose gel. Digested products were purified and the concentration of DNA determined. Amplified gene products were subsequently ligated into pET21a, pET28a or pCold plasmid vectors digested with the same restriction enzymes.

Recombinant DNA was used to transform electrocompetent GeneHog *E. coli* cells. Following a two hour incubation at 37 °C, the mixture was plated onto tributyrin agar supplemented with the appropriate antibiotic. Plates were incubated at 37 °C overnight, moved to 4 °C and monitored for several days for the formation of hydrolysis zones. Randomly selected clones were also screened via colony PCR, followed by agarose gel electrophoresis.

Clones with tributyrin hydrolysing activity were selected and plasmids extracted. Plasmid DNA was subjected to restriction enzyme analysis with NdeI and XhoI and sequenced with T7- promoter and terminator primers to verify gene orientation and sequence. All constructs were stored as 20 % [v/v] glycerol stocks at -80 °C.

Table 4.2.1: PCR conditions for amplification of predicted lipolytic genes.

Gene	Annealing temperature (°C)	Taq polymerase utilised	Expected size (bp)
***DEase*I**	69 °C	PrimeStar™ *Taq* (Takara)	948
***DEase*II**	67 °C	PrimeStar™ *Taq* (Takara)	1395
***DEase*IV**	58 °C	Phusion™ *Taq* (Finnzymes)	1074
***DEase*V**	67 °C	PrimeStar™ *Taq* (Takara)	1146

⁂ A standard protocol was used for amplification with the following conditions; Initial denaturation 98 °C for 2 minutes, followed by 30 cycles of denaturation at 98 °C for 10 seconds, annealing at defined temperature for 30 seconds, and extension at 72 °C for 30 seconds. The final elongation step was performed at 72 °C for 2 minutes.

Plasmids from sequence-verified clones were used to transform chemically competent *E. coli* Rosetta (DE3) pLysS cells. Transformation mixtures were incubated in SOC broth for 2 hours at 37 °C and plated onto tributyrin agar supplemented with the appropriate antibiotic. Following an overnight incubation at 37 °C, clones were monitored for hydrolysing activity at 4 °C. Hydrolysis positive clones were used in subsequent small scale expression trials.

4.2.4 Protein expression

In order to determine the expression profile of the proteins, overnight cultures were used to seed 50 ml LB broth supplemented with chloramphenicol and carbenicillin. Growth of these cultures at 37 °C was monitored until an OD_{600} ~0.4 – 0.6 was obtained. Each culture was then split into two equal volumes, one was treated as an uninduced control and the other induced with 0.4 – 1 mM IPTG. Both cultures were grown at 16 °C for 2 days. Cultures were centrifuged at 6000 × *g* for 10 minutes and the culture supernatant precipitated with 20 % [w/v] TCA at 4 °C for 24 hours. The extracellular fraction was obtained by another round of centrifugation at 8000 × *g* for 10 minutes and stored at 4 °C. Cell pellet fractions were resuspended in sonication buffer (30 mM Tris-HCl [pH 8.5], 300 mM NaCl, and 10 % [v/v] glycerol) and sonicated for 6 cycles of 30 seconds each. The soluble and insoluble fractions were separated by centrifugation at 6000 × *g* for 10 minutes. Aliquots of the extracellular- , insoluble- and soluble- fractions were mixed with an equal volume of 2 × SDS loading buffer and analysed by SDS-PAGE. Controls of parental vector in the *E. coli* strain were included. Soluble cell free extract containing the protein of interest was purified via metal affinity chromatography (Chapter 2).

4.2.5 HPLC analysis

Fifty millilitres of *DEase*I and *DEase*II overnight cultures in the Rosetta expression host were centrifuged at 6000 × *g* for 10 minutes. Cell pellets were resuspended in 4 ml of buffer (50 mM Na_2HPO_4) and subjected to six × 30 second cycles of sonication on ice. The cell suspension was incubated with 5 mM methyl ferulate in water, for two hours. Samples were centrifuged at 13 000 × *g* and decanted into HPLC tubes. Samples were maintained at 40 °C to prevent substrate precipitation and run on a C_{18} column at a flow rate of 0.8 ml/min with the mobile phase (35 % [v/v] methanol, 0.1 % [v/v] trifluoroacetic acid). The appearance of product and disappearance of substrate was observed and the AU/min measured.

4.2.6 Histidine- tag chromatography

Following charging and equilibration of His-Bind columns, the prepared extract was added. Buffer conditions required optimisation for the binding of both *DEase*I and *DEase*II proteins. *DEase*I was purified to homogeneity after the addition of 6 ml elution buffer (20 mM Tris-HCl [pH 7.9], 500 mM NaCl, 250 mM imidazole).

Purified protein was dialysed against 3 L of buffer (50 mM Tris-HCl [pH8.5], 1 % [v/v] glycerol) for 2 days at 4 °C. Fifty millilitres of dialysis buffer was stored at 4 °C and used as a control in assays. Proteins were quantified according to the Bradford assay using BSA as a standard.

4.2.7 Enzyme assays

Substrate preference towards *p*-Nitrophenyl esters of varying chain length was determined enzymatically by measuring the release of *p*-nitrophenol due to hydrolysis. Absorbance at 405 nm was measured continuously for 1–3 minutes using the Cary 50 UV-Vis spectrophotometer (Agilent Technology, USA). One unit of enzyme activity is defined as the amount of enzyme required for hydrolysis of 1 µmol *p*-nitrophenyl substrate per minute at 25

°C. In order to determine the temperature optimum of the enzyme, standard assays were performed using *p*-nitrophenyl propionate (C3) as substrate over a range of temperatures (5 – 45 °C). The pH optima was determined by measuring activity on *p*-nitrophenyl octanoate (C8) in the following buffers [mM]; sodium acetate (pH 5.5), MES (pH 5.5 and 6.5), MOPS (pH 6.5 and 7.5), sodium phosphate (pH 7.5 and 8.5), Tris-HCl (pH 7.5, 8.5 and 9.5) and CAPS (pH 9.5 and 10).

Thermal liability of enzyme was determined by measuring the residual enzyme activity following incubation of the enzyme at 25, 35 and 45 °C for time intervals of 5, 10, 30 and 60 minutes. The effect of NaCl concentration was determined by measuring enzyme activity in sodium phosphate buffer (pH 7.5) with increasing amounts of sodium chloride, from 1 M to 4 M. Blank reactions were included with each measurement. All assays were performed in triplicate.

4.2.8 FPLC analysis

Purified fractions of enzyme were subjected to size exclusion chromatography analysis using FPLC (Äkta, Amersham Biosciences, USA). Samples were loaded onto the HiLoad ™ Sephadex 75 HR 10/30 column (GE Healthcare) and fractionated with buffer (50 mM Tris-HCl [pH 7.5]) at a flow rate of 0.5 ml/min. All protein standards were used at a concentration of 1 mg/ml and included Cytochrome C (12 kDa), carbonic anhydrase (29 kDa) and BSA (66 kDa). Retention times were recorded and a standard curve was used to predict the size of the enzyme.

4.3 Results and Discussion

4.3.1 Bioinformatic analysis

In a previous study, several clones were obtained from a metagenomic library from Antarctic Dry Valley soils with tributyrin hydrolysing activity (Anderson, 2008). Fosmids were successfully extracted from these clones and sequenced at the University of the Western Cape Solexa sequencing platform. Full sequence assembly and annotation of three fosmid clones (LD4, LD7 and LD13) allowed for the prediction of putative lipolytic genes in each of the fosmids. Clones LD7 and LD13 each contained one ORF corresponding to possible lipolytic activity (*DEase*I and *DEase*V respectively). *DEase*III, from fosmid LD4, was the subject of another study and is therefore not included here. Interestingly, three adjacent ORFs with probable activity linked to the hydrolysis of ester containing substrates were detected in clone LD4. The architecture of the three ORFs, designated *DEase*II, *DEase*IV and *DEase*VI, as they are found on clone LD4, is indicated in Figure 4.3.1.1. ORF size and GC content were predicted and are provided in Table 4.3.1.1. The annotated sequences were subjected to a range of bioinformatic analyses. Translated nucleotide sequences were searched against the NCBI, InterproScan (Figure 4.3.1.2) and Uniprot databases, and the closest homologue as well as family motifs were identified (Table 4.3.1.2). The predicted molecular mass for the proteins *DEase*I, *DEase*II, *DEase*IV, *DEase*V and *DEase*VI were 34-, 52-, 40-, 45-, and 43-kDa respectively. All putative sequences were analysed for the presence of N-terminal signal peptides (Figure 4.3.1.3) as well as the occurrence and frequency of rare codons (Table 4.3.1.3). Interestingly, all putative genes contained a significant number of rare codons as well as codon sequence repeats. For the genes *DEase*I, *DEase*IV and *DEase*VI, the average percentage of rare codons was in excess of 11 %, while values of 6.9 and 6.8 were observed for *DEase*II and *DEase*V respectively. Similarly, an increased number of repeats of these rare codons was also observed for the former three genes. Of the three lipolytic genes situated

adjacent to each other on fosmid LD4, the two Metallophosphoesterase genes (*DEase*IV and *DEase*VI) shared a similar percentage of rare codons (11 %) with a high frequency of isoleucine and leucine codons in particular. When compared to the neighbouring gene, *DEase*II, the predominant rare codons were leucine, proline and arginine. This observation demonstrates the differences in codon frequency and usage among genes in the same genome even when the genes are in close proximity. Considering that *DEase*IV and *DEase*VI were annotated as Metallophosphoesterase enzymes, they were excluded from further bioinformatic analysis in this study.

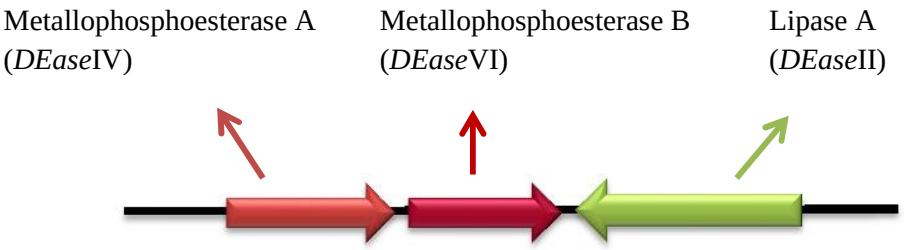


Figure 4.3.1.1: Diagrammatic representation of the architecture of the putative lipolytic operon from clone LD4.

Table 4.3.1.1: The full length gene size and the GC content of predicted lipolytic genes

Gene name	Gene length (bp)	GC content (%)
***DEase*I**	948	65
***DEase*II**	1395	47
***DEase*IV**	1074	43
***DEase*V**	1146	52
***DEase*VI**	1146	43

Table 4.3.1.2: Amino acid identity to the closest match using BLASTp (November, 2011) as well as predicted molecular weight (MW) of predicted lipolytic genes.

Gene name	Fosmid clone	% aa identity	Closest match	Protein MW
***DEase*I**	LD7	52	*Actinobacterium*	33.5 kDa
***DEase*II**	LD4	57	*Psychrobacter* spp	52 kDa
***DEase*IV**	LD4	64	*Psychrobacter* spp	40 kDa
***DEase*V**	LD13	55	*Pseudomonas* spp	41.5 kDa
***DEase*VI**	LD4	69	*Psychrobacter* spp	43 kDa

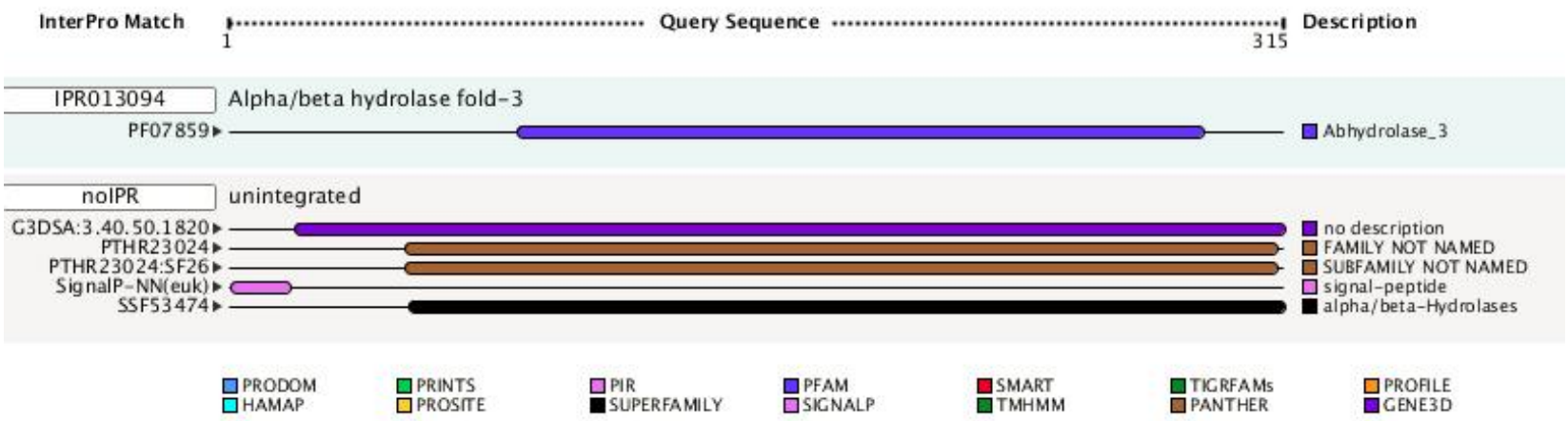
Table 4.3.1.3: Rare codon analysis of lipolytic genes. Numbers indicate the frequency with which each rare codon occurs in the gene sequence.

Amino acid	Rare codon ⁂	Frequency *DEase*I	Frequency *DEase*II	Frequency *DEase*IV	Frequency *DEase*V	Frequency *DEase*VI
Arginine	CGA	2	1	3	0	1
	CGG	8	2	3	5	1
	AGG	0	0	0	1	0
	AGA	1	3	3	0	4
Glycine	GGA	5	3	3	2	3
	GGG	3	3	4	8	4
Isoleucine	AUA	2	2	13	2	8
Leucine	CUA	1	6	3	5	13
Proline	CCC	7	6	2	3	2
Threonine	ACG	8	6	5	2	3
Repeated and/ or consecutive rare codons *		CUA- CGG CGG- CCC CGG- AUA CCC- GGA	CUA- CGG	ACG- AUA GGG- AGA AUA- ACG CCC- AUA	AUA- CCC CCC- GGA	AUA- CCC AGA- AUA- AUA CUA- AGA CUA- GGG- CGG CUA- CUA

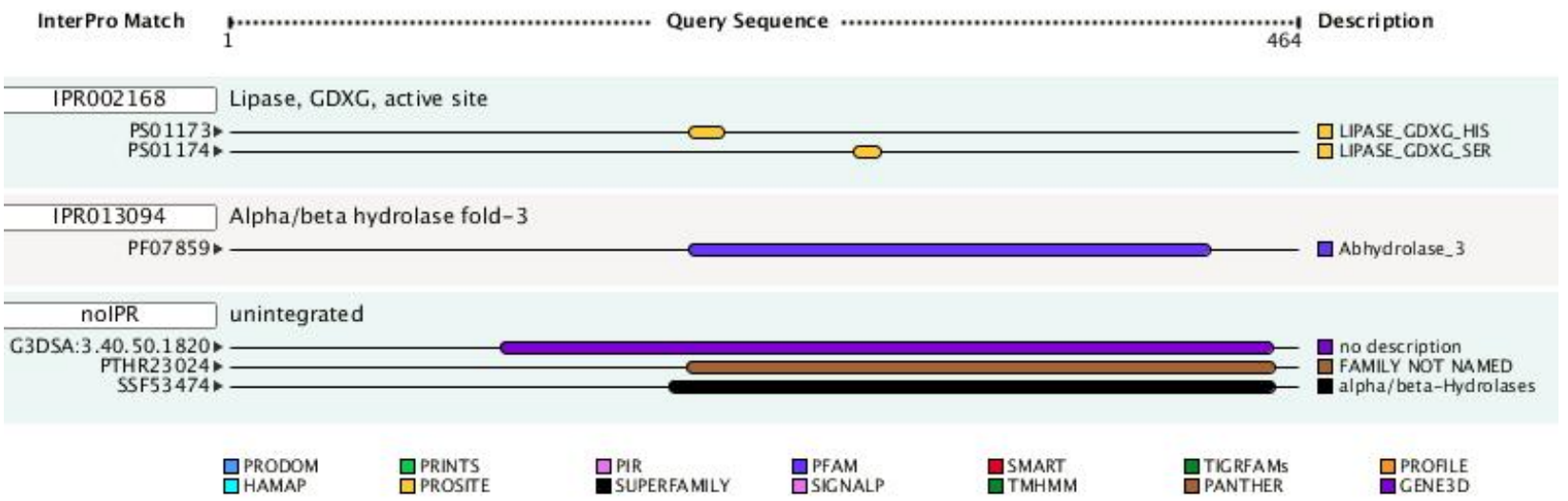
⁂ Rare codons occurring in *E. coli.*

* Repeated and/ or consecutive rare codons occur once in each of the five gene sequences.

A


InterPro Match
Query Sequence
Description
1
315
IPR013094
Alpha/beta hydrolase fold-3
PF07859
Abhydrolase_3
noIPR
unintegrated
G3DSA:3.40.50.1820
PTHR23024
PTHR23024:SF26
SignalP-NN(euk)
SSF53474
no description
FAMILY NOT NAMED
SUBFAMILY NOT NAMED
signal-peptide
alpha/beta-Hydrolases
PRODOM
HAMAP
PRINTS
PROSITE
PIR
SUPERFAMILY
PFAM
SIGNALP
SMART
TMHMM
TIGRFAMs
PANTHER
PROFILE
GENE3D

B


InterPro Match
Query Sequence
Description
1
464
IPR002168
Lipase, GDXG, active site
PS01173
PS01174
LIPASE_GDXG_HIS
LIPASE_GDXG_SER
IPR013094
Alpha/beta hydrolase fold-3
PF07859
Abhydrolase_3
noIPR
unintegrated
G3DSA:3.40.50.1820
PTHR23024
SSF53474
no description
FAMILY NOT NAMED
alpha/beta-Hydrolases
PRODOM
HAMAP
PRINTS
PROSITE
PIR
SUPERFAMILY
PFAM
SIGNALP
SMART
TMHMM
TIGRFAMs
PANTHER
PROFILE
GENE3D

C

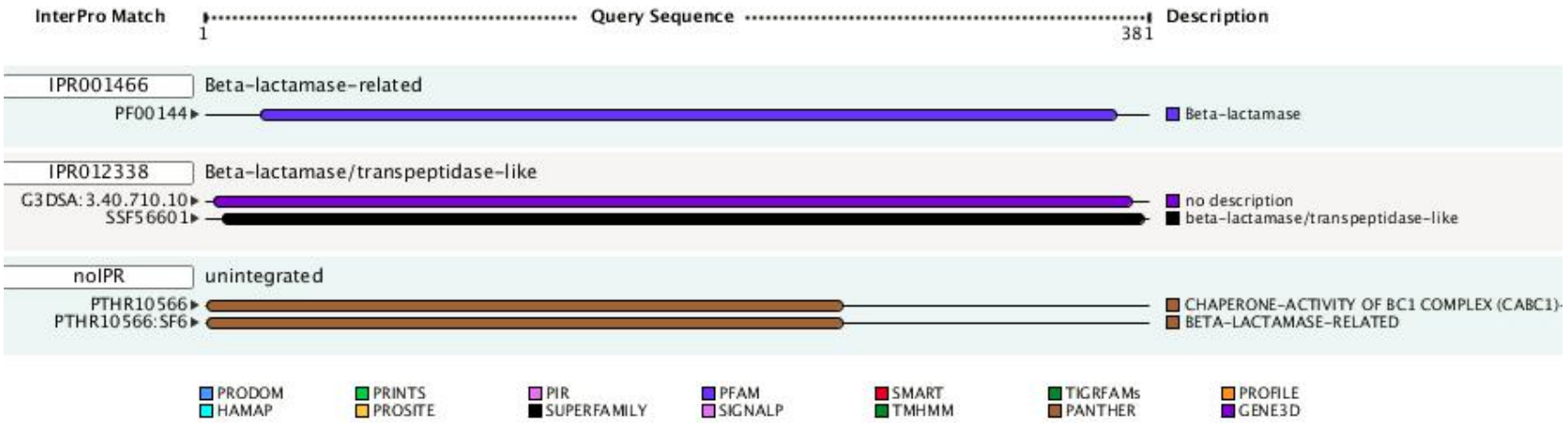
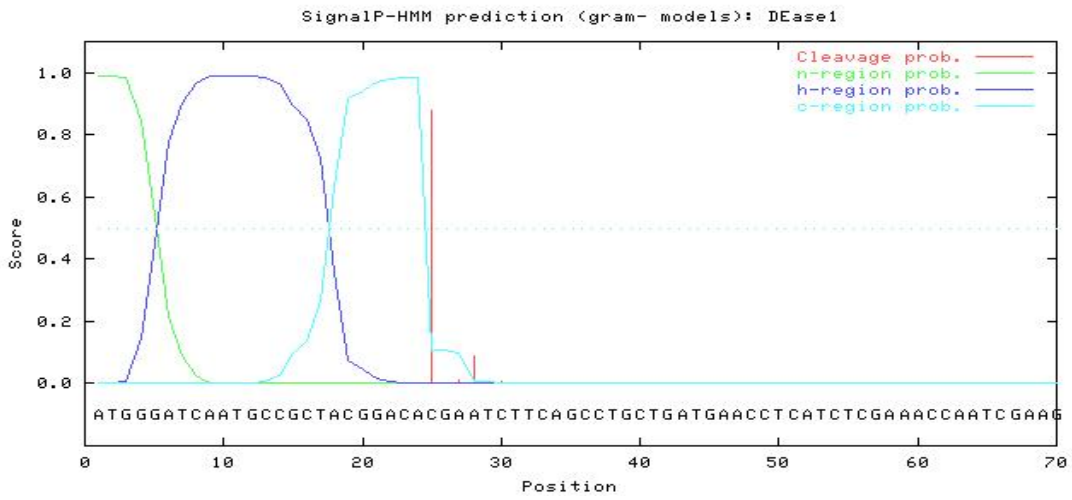


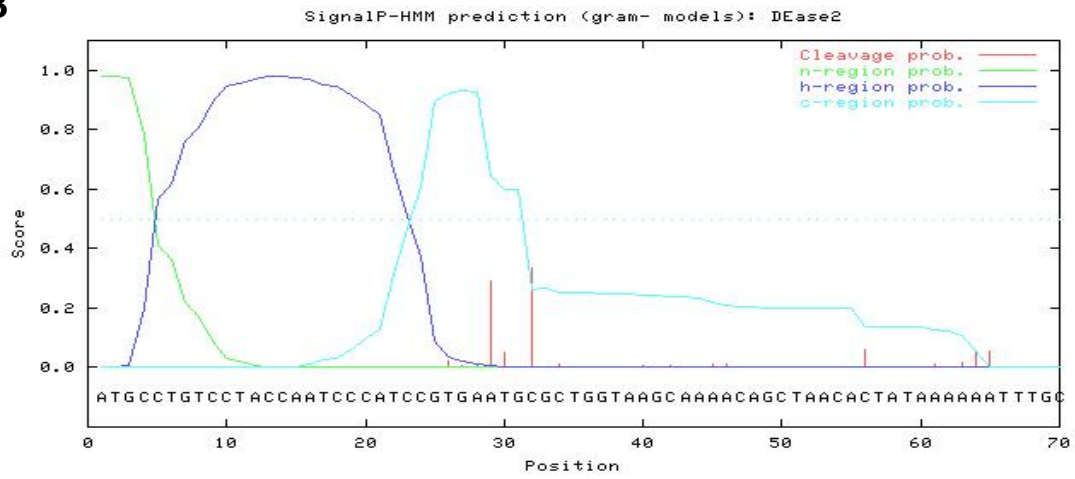
Figure 4.3.1.2: InterproScan results showing protein domain hits for predicted lipolytic ORFs. A; *DEase*I, B; *DEase*II and C; *DEase*V

A



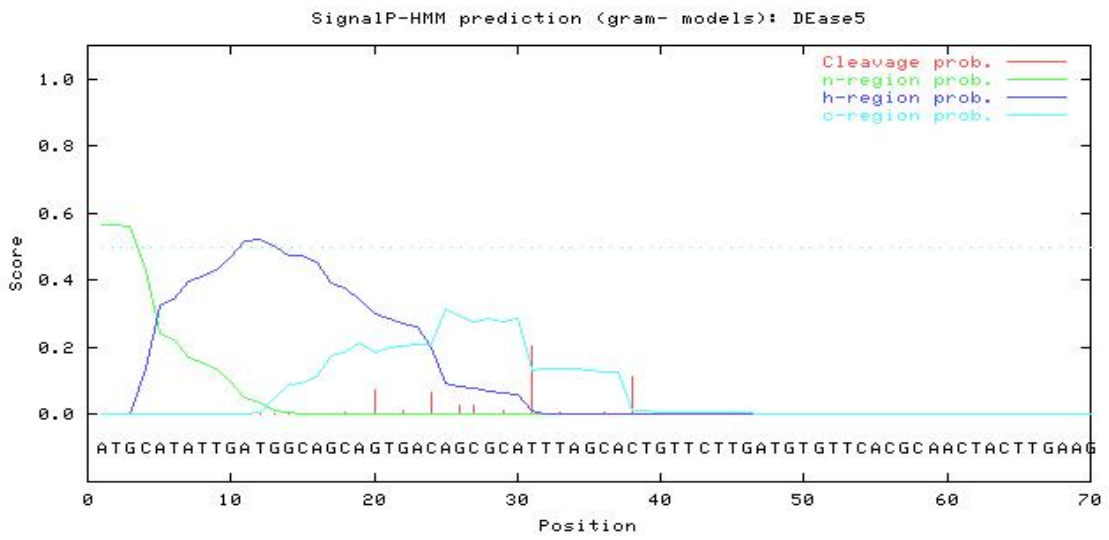
- *Signal peptide probability: 0.991*
- *Max cleavage site probability: 0.880 between pos. 24 and 25*

B



- *Signal peptide probability: 0.980*
- *Max cleavage site probability: 0.334 between pos. 31 and 32*

C



- *Signal peptide probability: 0.567*
- *Max cleavage site probability: 0.204 between pos. 30 and 31*

Figure 4.3.1.3: SignalP results for putative lipolytic clones. Numbers highlighted in green indicate the probability that an N-terminal signal cleavage site occurs. Low scores indicate low probability of a signal peptide site. A; *DEase*I, B; *DEase*II and C; *DEase*V.

The top hits identified by BLASTp analysis were used to construct individual multiple sequence alignments. Conserved residues were identified in all genes (Figures 4.3.1.4 – 4.3.1.6).

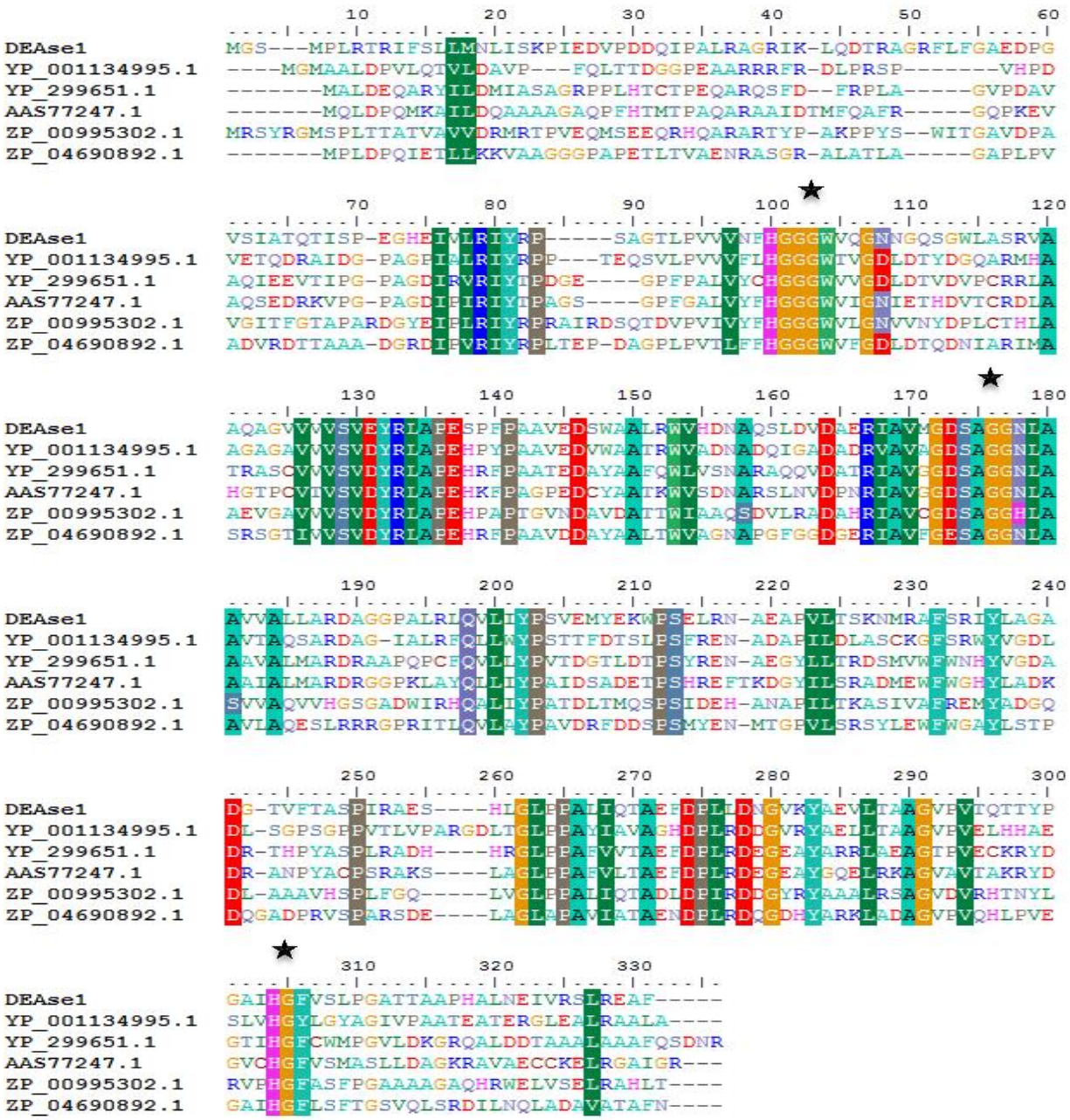


Figure 4.3.1.4: **Multiple sequence alignment of *DEase*I with the five top hits identified by BLASTp analysis against the NCBI non-redundant database. Conserved residues are indicated by ★. Accession numbers in the figure denote the following; AAS77247: lipase/esterase [uncultured bacterium], YP_299651: Alpha/beta hydrolase [*Mycobacterium gilvum*], YP_001134995: LipH [*Janibacter* sp.], ZP_00995302: Putative lipase [*Streptomyces ghanaensis*], ZP_04690892: Lipase [*Streptomyces ghanaensis*].**

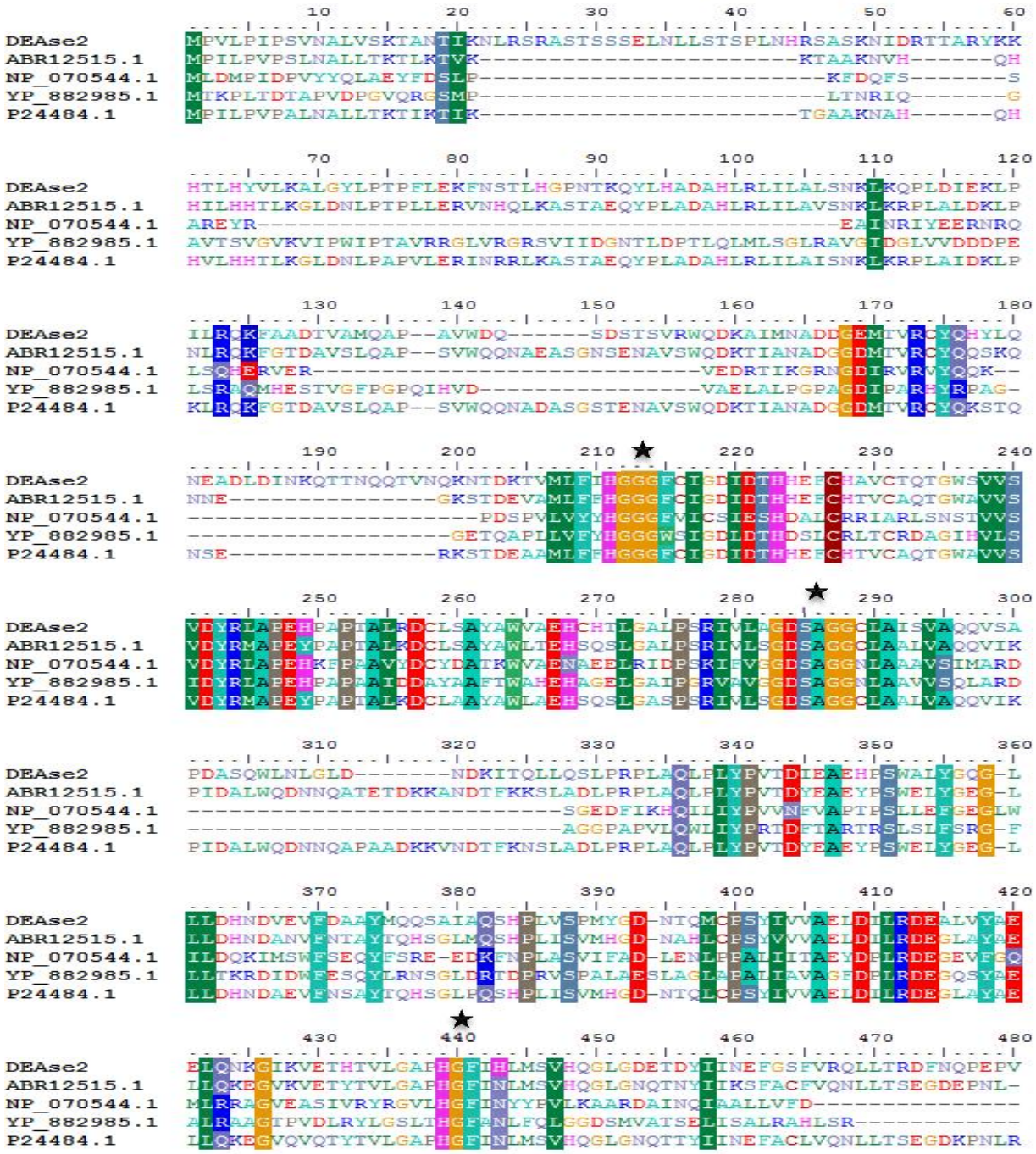


Figure 4.3.1.5: Multiple sequence alignment of *DEase*II with the four top hits identified by BLASTp analysis against the NCBI non-redundant database. Conserved residues are indicated by ★. Accession numbers in the figure denote the following; ABR12515: Lipase [*Psychrobacter* sp. 2-17], NP_070544: Carboxylesterase [*Archaeoglobus fulgidus*], YP_882985: Alpha/beta hydrolase [*Mycobacterium avium*], P24484: LIP2 [*Moraxella* sp. (strain TA144)]

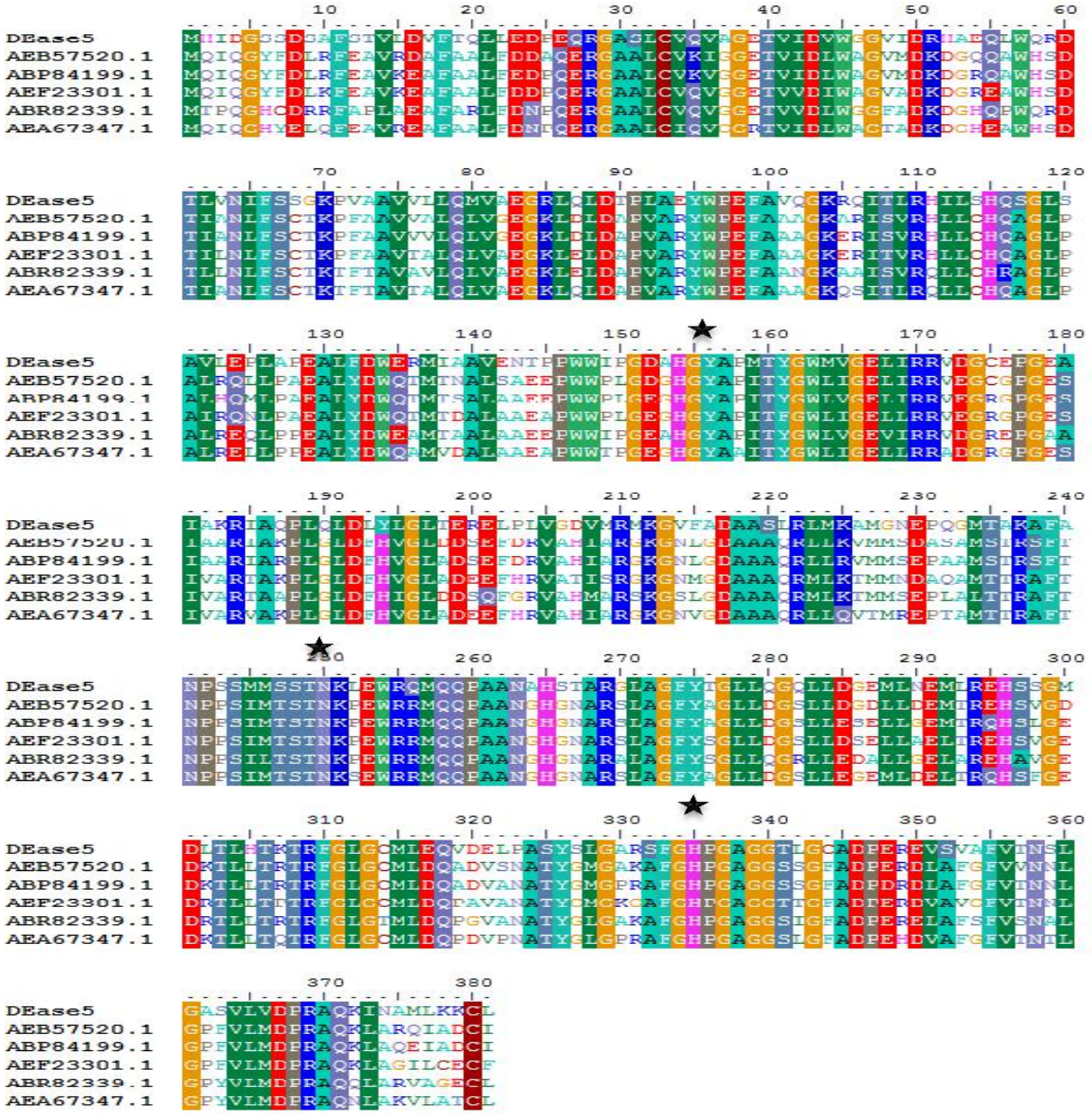


Figure 4.3.1.6: **Multiple sequence alignment of *DEase*V with the five top hits identified by BLASTp analysis against the NCBI non-redundant database. Conserved residues are indicated by ★. Accession numbers denote the following; AEA67347: Esterase III [*Pseudomonas brassicacearum*], ABR82339: Probable esterase [*Pseudomonas aeruginosa*], AEF23301: Beta-lactamase [*Pseudomonas fulva*], ABP84199: Beta-lactamase [*Pseudomonas mendocina* ymp], AEB57520: Beta-lactamase [*Pseudomonas mendocina*].**

Additionally, phylogenetic analysis of genes *DEase*I, *DEase*II and *DEase*V, together with other representative sequences of the 8 lipolytic families, revealed that *DEase*I and *DEase*II clustered within the HSL family (family IV) of lipases and esterases while *DEase*V clustered in Family VIII. This analysis corresponds well to the conserved motifs identified in individual multiple sequence alignments (Figure 4.3.1.7).

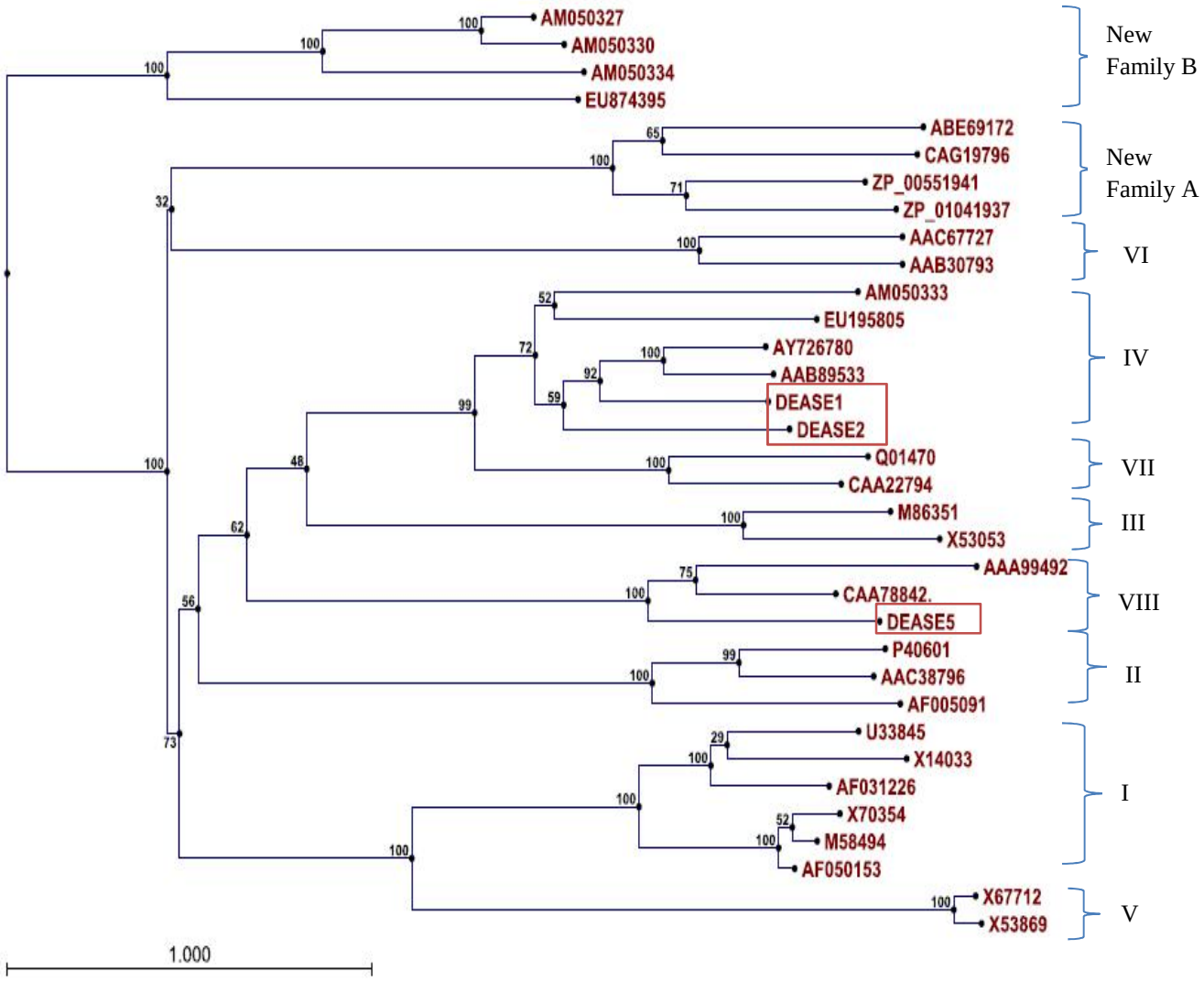


Figure 4.3.1.7: Neighbour joining tree showing phylogenetic positions of *DEase*I, *DEase*II and *DEase*V within selected lipase/esterase families based on conserved sequence motifs of bacterial lipolytic enzymes. The phylogenetic tree was constructed using the CLC genomics workbench software. Bar, 1 substitution per amino acid site.

Bootstrap values lower than 50% are also shown. Accession numbers in the figure denote the following; AM050327: Acetyl xylan esterase [phagemid clone pBKR.17], AM050330: Subtilisin-like serine protease [phagemid clone pBKR.35], AM050334: Acetlyxylan esterase [phagemid clone pBKR.44], EU874395: CHA3 esterase gene [Uncultured bacterium clone], ABE69172: Probable lipase [uncultured bacterium pFosLip], CAG19796: Hypothetical protein/ probable esterase [*Photobacterium profundum*], ZP_00551941: Lipase, class 3 [*Desulfuromonas acetoxidans*], ZP_01041937: Lipase family protein [*Idiomarina baltica*], AAC67727: Predicted Lysophospholipase esterase [*Chlamydia trachomatis*], AAB30793: Serine esterase [*Arthrospira platensis*], AM050333: [phagemid clone pBKR.43], EU195805: pBSAT1 esterase gene [Uncultured bacterium clone], AY726780: estE1 carboxylesterase gene [Uncultured archaeon clone], AAB89533: Carboxylesterase (estA) [*Archaeoglobus fulgidus*], Q01470: Serine esterase [*Arthrobacter oxydans*], CAA22794: Putative carboxylesterase [*Streptomyces coelicolor*], M86351: triacylglycerol acylhydrolase (lipA) [*Streptomyces* sp], X53053: lipase 1 [*Moraxella* sp], AAA99492: Carboxylic ester hydrolase [*Arthrobacter globiformis*], CAA78842: Esterase A [*Streptomyces anulatus*], P40601: Lipase 1 [*Photorhabdus luminescens*], AAC38796: Outer membrane esterase [*Salmonella enterica*], AF005091: esterase (estA) [*Pseudomonas aeruginosa*], U33845: Alkaline lipase [*Proteus vulgaris*], X14033: Lipase [*Pseudomonas fragi*], AF031226: Lipase (lipA) [*Pseudomonas fluorescens*], X70354: lipA [*Burkholderia glumae*], M58494: LipA lipase [*Burkholderia cepacia*], AF050153: Triacylglycerol lipase [*Pseudomonas luteola*], X67712: Lipase [*Psychrobacter immobilis*], X53869: Lipase 3 [*Moraxella* sp.].

Threading the query sequence onto a template with a known crystal structure allowed the prediction of tertiary structural models. The protein sequences of *DEase*I and *DEase*II modelled to templates from thermophilic organisms, namely pEstE from *Pyrobaculum calidifontis* [2wir] (40 % sequence identity) and EstE1, a carboxylic esterase from an uncultured thermophilic archeon [2c7b] (24 % sequence identity), respectively (Figure 4.3.1.8). In the case of *DEase*II, a second model was constructed using sequence residues 144 – 450 with 29 % sequence identity to the monomer protein, AFEST, from *Archeoglobus fulgidus* [1jji] (Figure 4.3.1.9). Residues 8 – 379 of *DEase*V modelled to EstB, a novel esterase containing the β-lactamase fold, from *Burkholderia gladioli* [1ci9] with 19.2 % sequence identity (Figure 4.3.1.10). Analysis of Ramachandran plots indicated that constructed models for *DEase*I, *DEase*II and *DEase*V were relatively accurate, with 1.5 %, 1.8 % and 2.2 % of residues in disallowed regions, respectively. This could be expected as templates for modelling are known crystal structures deposited in the protein data bank (PDB) and relative to the number of protein sequences available, the number of tertiary structures is limited. Folds or smaller domains which are novel cannot be modelled with high levels of certainty without the correct reference crystal structure. Templates for modelling in this study originated from thermotolerant or thermophilic microorganisms because there are more of these templates than from cold-adapted representatives. This can be explained by the substantial attention to various enzymes from these sources due to biotechnological and industrial interest (Gomes and Steiner, 2004).

Upon further analysis of the homology models, an unusual extension made up of two antiparallel β-sheets appeared to extend over the active site of *DEase*II. The initial speculation, that this structure may be a lipolytic lid, is unlikely for two reasons. Firstly, the known lid-like structures of lipases consist of a series of α-helices and not β-sheets as was observed in *DEase*II. Secondly, the lid structure generally originates from the N-terminus

but, in *DEase*II, the structure extends from the centre of the gene. This modification is not surprising, considering that the α/β hydrolase fold enzymes offen depict unique differentiation within a common fold and therefore, a striking ability for evolution and adaptation (Nardini and Dijkstra, 1999). Catalytic mechanisms are generally maintained even with insertions or deletions of amino acid residues and/ or extra domains. These insertions or deletions are not exclusively located at C- or N- terminal regions of the protein but may also occur at the C edge of β strands (Nardini and Dijkstra, 1999). Some interesting observations with respect to sequence and protein structure have been made in recent years. In 2005, Rusnak and co-workers characterised the enzyme AFL from *Archeoglobus fulgidus* and, based on enzymatic assays, this protein displayed typical esterase substrate specificities. However, in 2009, the crystal structure of the protein was resolved and the authors noted the presence of the N-terminal lipase lid as well as a unique C-terminal extension composed of a series of β-sheets. This prompted further enzymatic analysis and mutation studies which confirmed that AFL exhibited classical interfacial activation and that the C- terminal domain was responsible for the binding of long chain triacylglycerols. AFL was re-classified as a true lipase based on these observations, even though the sequence-based analysis did not cluster it in family I (Chen *et al*, 2009). The opposite may also occur; the gene sequence for the protein Est53 from *Thermotoga maritima* was shown to cluster in family I.2 with homology to the *Burkholderia* lipase. Additionally, following purification of the recombinant protein and characterisation using *p*-Nitrophenyl substrates, it was found that enzymatic activity was dependent on the presence of Ca^{2+} ions, a typical biochemical feature of this subfamily. However, Est53 demonstrated no activity toward long chain fatty acid molecules and was eventually classified as a carboxyl esterase (Kagugawa *et al*, 2007).

Figure 4.3.1.8: Homology model of *DEase*I built by the Swiss model server (Red) and superimposed (PyMol) onto the template 2C7B (Green).

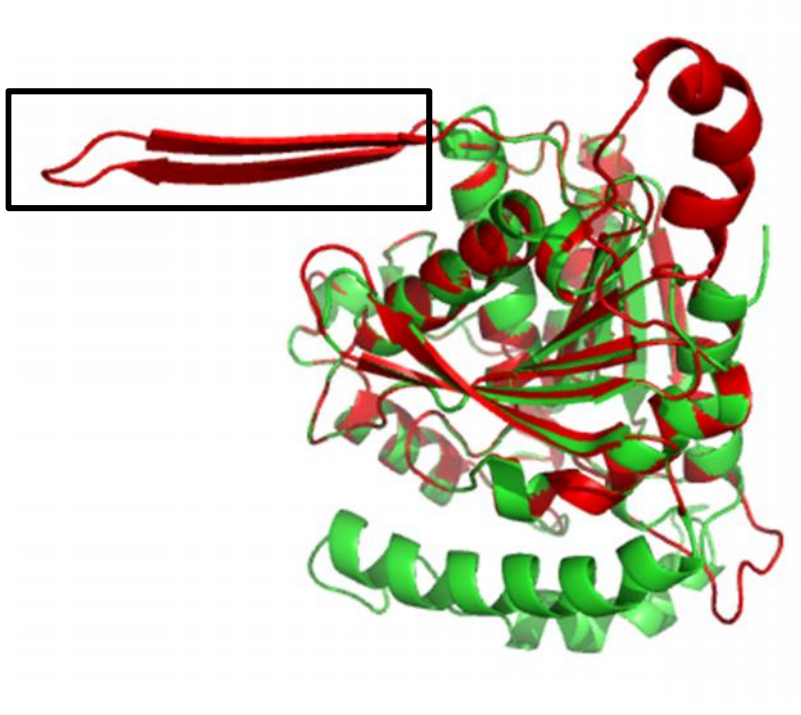


Figure 4.3.1.9: Homology model of *DEase*II built by the Swiss model server (Red) and superimposed (PyMol) onto the template 1JJI (Green). Within the black box is the protein region of the β-sheet extension which does not occur in the template.

Figure 4.3.1.10: Homology model of *DEase*V built by the Swiss model server (Red) and superimposed (PyMol) onto the template 1CI9 (Green).

4.3.1.1 Comparative genomics

No single set of parameters can distinguish psychrophilic, mesophilic or thermophilic proteins. However, comparative genomic studies have revealed some trends in amino acid composition, protein sequence and structure, GRAVY and disorder which may be used to assess probable psychrophilicity of a protein (Metpally and Reddy, 2009; Grzymski *et al*., 2006). These analyses must be used with caution as not all proteins utilise the same strategies to attain the increased level of flexibility required for cold-adaptation (Smalås *et al*., 2000; Gianese *et al*., 2002; Casanueva *et al*., 2010).

In this study, the lipolytic enzymes identified, as well as selected psychrophilic or mesophilic homologs, were assessed for amino acid composition, GRAVY and Aliphatic index (Table 4.3.1.4 – 4.3.1.6). The sequence and structural properties were investigated by examining the frequency of occurrence of amino acids believed to be general indicators of cold-adaptation.

For example; an increase in glycine residues may contribute to increase in protein flexibility and stability (Goldstein, 2007). Similarly, an increased proportion of threonine residues may also be advantageous for cold-adaptation, due to increased interaction of this residue with polar solvent such as water (Jahandideh *et al*., 2008).

Features which may be involved in temperature adaptation that are investigated here, and would indicate psychrophilicity include; decreased hydrophobic character, decreased percentage of arginine, glutamic acid, proline, leucine, lysine, phenylalanine, asparagine and tyrosine, increase in the amino acid residues glutamine, alanine, aspartic acid, serine, cysteine, threonine, isoleucine, glycine and valine, all of which may be common to cold-adapted proteins (Casanueva *et al*., 2010; Jahandideh *et al*., 2007; Metpally and Reddy, 2009; Ayala-del-Río *et al*., 2010; Gianese *et al*., 2002; Grzymski *et al*., 2006; Goldstein, 2007 as well as references therein).

In the case of *DEase*I, changes of this nature were not observed when compared to the mesophilic homologs. Differences which may point to cold-adaptation in this protein include a decrease in proline and arginine and an increase in serine residues. However, based on this comparative analysis alone, the sequence composition of *DEase*I closely resembles that of mesophilic proteins and includes an increase in lysine and leucine residues coupled to a decrease in alanine, glycine and valine, and a positive GRAVY value (0.104) which indicates overall hydrophobicity (Table 4.3.1.4).

For *DEase*II, similar amino acid profiles to psychrophilic homologs were detected, particularly for phenylalanine, isoleucine, asparagine, and serine. In addition, these residues followed a similar trend to averages calculated for mesophiles and psychrophiles by Metpally and Reddy (2009). For example, a decrease in the number of asparagine residues was observed for *DEase*II (3.17 %) and the psychrophilic homologs, which were all lower than

the average calculated for mesophilic organisms (4.7 %). In addition, an increase in alanine, glycine, proline, arginine and valine was observed and these values exceeded the average predicted for psychrophilic organisms (Table 4.3.1.5). The percentages of lysine, tyrosine and leucine decreased in *DEase*II when compared to the psychrophilic homologs and mesophilic averages. All the above mentioned factors, including a hydrophilic GRAVY value (-0.216), would be expected for a cold-adapted protein and it may therefore be hypothesised that *DEase*II could, in fact, function as a psychrophilic protein.

Table 4.3.1.4: Amino acid composition calculated in ProtParam (Expasy) for *DEase*I and four mesophilic homologs. E6TCC6: Esterase/lipase (*Mycobacterium*) sp.; D52NTO: Esterase/lipase/thioesterase (*Streptomyces ghanaensis*); A3TL42: Probable lipase LipH (*Janibacter* sp); E9UY50: Carboxylesterase (Nocardioidaceae).

	DEaseI	E6TCC6	D52NTO	A3TL42	E9UY50
Ala	9.05	16.23	15.06	15.22	15.03
Cys	1.72	0.32	0	0.62	0.33
Asp	6.47	8.77	8.01	6.83	4.25
Glu	4.09	3.9	4.17	3.11	5.88
Phe	2.37	2.27	3.53	1.86	3.59
Gly	4.74	9.42	9.62	7.45	9.15
His	4.53	2.60	1.6	4.35	2.94
Ile	474	2.6	3.53	4.66	3.92
Lys	4.09	0.32	0.96	0.62	1.31
Leu	12.07	10.06	9.62	6.83	8.17
Met	1.94	0.65	1.28	1.86	2.61
Asn	4.74	0.97	2.56	1.55	1.96
Pro	5.60	8.44	8.01	7.14	8.17
Gln	6.03	2.6	2.88	3.73	1.96
Arg	3.45	6.82	7.05	7.45	6.21
Ser	6.9	4.22	5.13	5.28	5.88
Thr	6.47	5.52	5.45	5.9	5.56
Val	6.47	9.42	7.69	9.94	8.17
Trp	1.29	1.62	1.28	1.55	1.63
Tyr	3.23	3.25	2.56	4.04	3.27
GRAVY	0.104	-0.002	-0.072	-0.062	0.032
Ai	98.54	92.92	88.62	88.85	85.88

Table 4.3.1.5: Amino acid composition calculated in ProtParam (Expasy) for DEaseII and three psychrophilic homologs. Averages determined by a comparative genomic study are included (Metpallay and Reddy, 2009). Q796C8: Cold-active esterase (*Psychrobacter* sp. Ant300); Q1Q7W8: Alpha/beta hydrolase fold-3 (*Psychrobacter cryohalolentis* strain K5); P24484: Lipase 2 (*Moraxella* sp. strain TA144).

	*DEase*II	Q769C8	Q1Q7W8	P24484	Average for psychrophiles⁂	Average for mesophiles ⁂
Ala	13.65	9.75	8.70	11.55	9.2	8.1
Cys	0	2.50	2.07	1.85	1.0	1.0
Asp	4.44	7	6.83	6	5.5	5.1
Glu	5.4	4	5.59	4.39	5.8	6.3
Phe	3.17	2.75	3.73	2.08	4.1	4.4
Gly	8.57	7.50	6	6	7.1	6.7
His	1.90	2.25	4.35	3.93	2.1	2.1
Ile	5.08	5.50	6.42	4.39	6.6	6.8
Lys	1.59	4.75	4.76	4.85	5.2	6.2
Leu	9.84	11.50	11.39	12.01	10.3	10.7
Met	1.90	1.75	2.48	1.39	2.5	2.5
Asn	3.17	3	3.93	4.85	4.3	4.7
Pro	7.3	5.25	4.97	5.77	4	4
Gln	3.17	4.25	5.18	5.77	4.3	4.5
Arg	6.35	2.75	3.11	3	4.6	4.6
Ser	6.67	6.0	6.63	6	6.8	6.1
Thr	4.76	5.75	5.59	5.77	5.5	5.2
Val	9.21	6	4.55	5.77	6.9	6.6
Trp	1.59	1.25	1.24	1.39	1.2	1.2
Tyr	2.2	3.5	2.48	3.23	2.9	3.3
GRAVY	-0.216	-0.090	-0.185	-0.210	N/D	N/D
Ai	93.36	93.45	91.35	92.24	N/D	N/D

N/D – not determined in the study.

⁂ Indicates average amino acid usage determined by Metpallay and Reddy (2009).

Amino acid variations that may indicate cold-adaptation of *DEase*V includes a decrease in phenylalanine, tyrosine and arginine as well as a GRAVY value indicative of hydrophilicity. In addition, the percentage of valine and serine residues increased when compared to the mesophilic homologs. Amino acid changes which would indicate the propensity for function at higher temperatures includes a decrease in alanine, aspartic acid, glycine and threonine,

and an increased, or similar percentage of, lysine, leucine, glutamic acid, proline and asparagine residues, when compared to mesophilic homologs (Table 4.3.1.6).

Table 4.3.1.6: Amino acid composition calculated in ProtParam (Expasy) for *DEase*V and four mesophilic homologs. F4DX91: Beta-lactamase (*Pseudomonas mendocina*); F6ABP2: Beta-lactamase (*Pseudomonas fulva*); Q5K6J9: Esterase (*Pseudomonas fluorescens*); F2K8N1: Esterase III (*Pseudomonas brassicacearum*).

	*DEase*V	F4DX91	F6ABP2	Q5K6J9	F2K8N1
Ala	10.24	13.39	13.39	13.39	13.65
Cys	1.31	1.57	1.57	1.31	1.31
Asp	4.99	7.87	5.51	5.25	5.25
Glu	6.56	4.2	6.3	6.56	5.77
Phe	2.89	4.46	4.72	3.94	4.2
Gly	9.19	10.76	10.76	10.24	10.5
His	2.36	2.36	2.36	2.89	3.15
Ile	3.15	3.67	2.89	2.10	2.62
Lys	2.89	3.15	3.15	2.36	2.36
Leu	12.07	10.76	10.24	12.07	11.55
Met	472	3.15	3.15	2.1	2.36
Asn	2.36	2.62	2.89	2.1	2.62
Pro	5.25	4.72	5.51	5.51	5.51
Gln	5.25	3.94	3.67	3.94	4.99
Arg	5.25	6.04	5.77	6.3	5.77
Ser	6.3	4.72	2.89	3.41	2.89
Thr	4.72	3.94	6.3	7.35	6.56
Val	6.82	4.72	5.25	4.99	4.72
Trp	2.1	2.1	2.1	2.1	2.1
Tyr	1.57	1.84	1.57	2.1	2.1
GRAVY	-0.064	-0.111	-0.127	-0.160	-0.154
Ai	89.40	83.39	79.79	83.12	82.62

The aliphatic index (Ai) of a protein is defined as the relative volume occupied by aliphatic side chains (alanine, valine, isoleucine, and leucine) (Sleator and Walsh, 2010). This value is generally correlated to protein thermostability (Ikai, 1980). High Ai values (>80) may indicate that the proteins are active over a wide range of temperatures. All three lipolytic enzymes investigated in this study show Ai values >85. This may indicate an adaptive feature

for survival in environments where temperature fluctuations are prevalent, such as the Antarctic Dry Valley soils.

While comparative analysis may generate interesting trends, the data may not always be indicative of psychrophilicity. In this study, comparative analysis based on sequence shows that both *DEase*I and *DEase*V would most likely be classified as psychro-tolerant proteins, while *DEase*II is most likely to be a truly psychrophilic enzyme. Clearly, this property should be validated by functional characterisation of individual enzymes. However, results from this analysis must be interpreted with caution as these trends are general observations and it is highly likely that even closely related protein families will not exhibit the same sequence modifications, nor will psychrophilic proteins exhibit all of the characteristics described.

4.3.2 Cloning of lipolytic genes

In order to confirm retention of the hydrolytic activity initially observed, fosmids were individually transformed into the Epi300 *E. coli* host and screened for tributyrin hydrolysing activity at 4 °C. All sequenced clones retained the lipolytic activity thereby confirming that activity was linked to the fosmid and each cloned insert rather than an artifactual activity of the *E. coli* heterologous host.

Three ORF's with homology to lipolytic genes occurred in fosmid LD4, two of which (*DEase*IV and *DEase*VI) were annotated as metallophosphoesterases (MPE's) while the third ORF showed homology to Lip2 from *Moraxella* TA144. Metallophosphoesterases (E.C. 3.6.1.53) belong to the calcineurin-like phosphoesterase superfamily (Pfam 00149), which contain two well characterised enzyme groups (monophosphoesterases and diphosphoesterases) which are involved in a wide variety of cellular functions including metabolic and regulatory pathways (Aravind and Koonin, 1998; Koonin, 1994). Members of this superfamily display a conserved GD/GNH signature in a structural β-α-β-α-β fold and

therefore may retain some similarity in their reaction mechanism (Koonin, 1994). Substrate targets for these enzymes include phosphorylated serine and threonine residues in proteins, polynucleotides, various phosphoesters and phospholipids (Koonin, 1994). Although many MPE's have been biochemically characterised, no natural biological substrates have been identified for many of these enzymes (Tyagi *et al*., 2009). Although no reports of tributyrin hydrolysis catalysed by MPE's currently exist in literature, the possibility of novel functions conferred by this large and diverse group of proteins cannot be overlooked. The possibility that either *DEase*IV or *DEase*VI may form zones of hydrolysis on lipid containing agar plates in this study required further investigation. Transposon mutagenesis was utilised to identify the ORF linked to hydrolytic activity (Figure 4.3.2.1). All knock out mutants selected for sequencing showed that transposon insertion in the gene sequence for *DEase*II was responsible for the loss of activity hence demonstration that in this study, both *DEase*IV and *DEase*VI accorded with the conventional characteristics of MPE's *i.e.* they did not possess tributyrin hydrolysing activity. However, in order to support the results obtained from random mutagenesis studies, *DEase*IV was included in further sub cloning experiments.

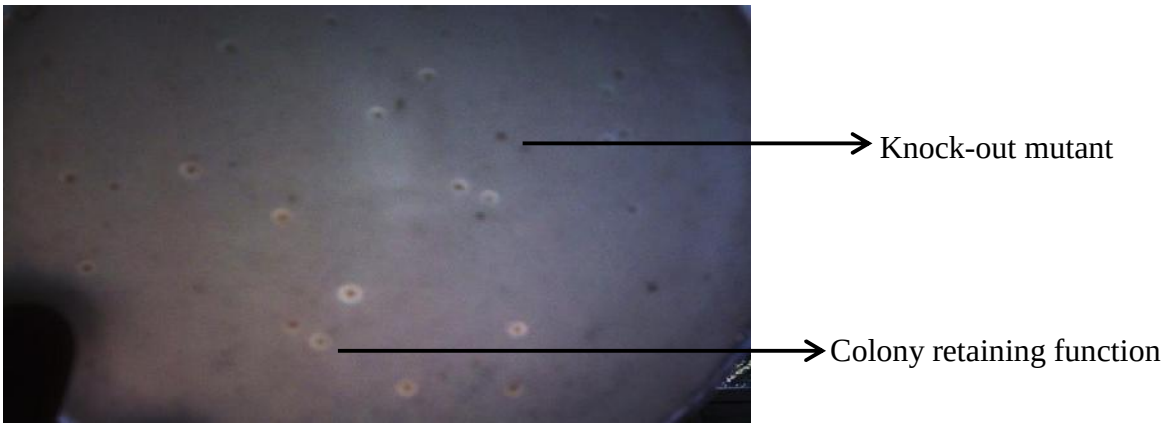


Figure 4.3.2.1: Functional screening of knock-out clones generated by transposon mutagenesis of clone LD4. Zones of clearing around colonies indicate retention of function while colonies with no zones are most likely mutated in the gene (s) responsible for tributyrin hydrolysis.

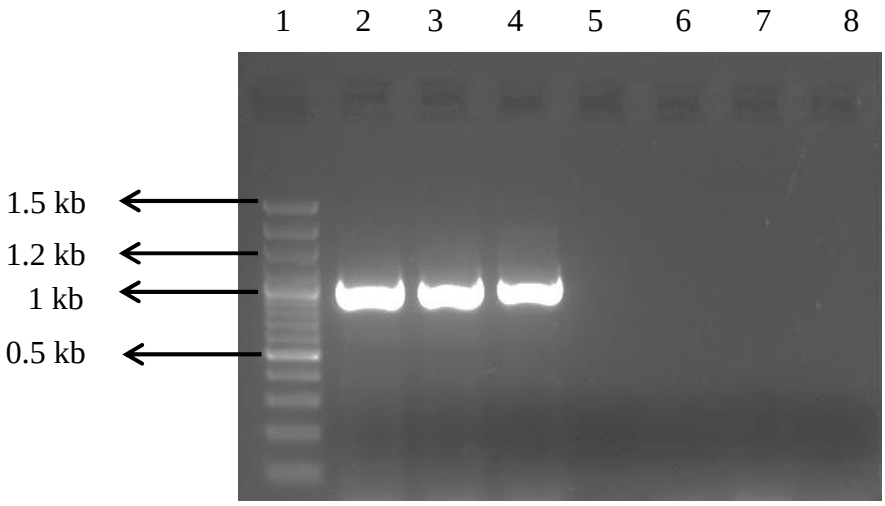


Figure 4.3.2.2: PCR amplification of the 950 bp *DEase*I gene product from clone LD7 using specific primer pairs. Lane 1) GeneRuler 1 kb plus™ DNA molecular mass marker (Fermentas). Lane 2 - 4) Amplicons of *DEase*I. Lane 5) Negative control. Lane 6) Reverse primer control. Lane 7) Forward primer control. Lane 8) Control reaction using *E. coli* genomic DNA.

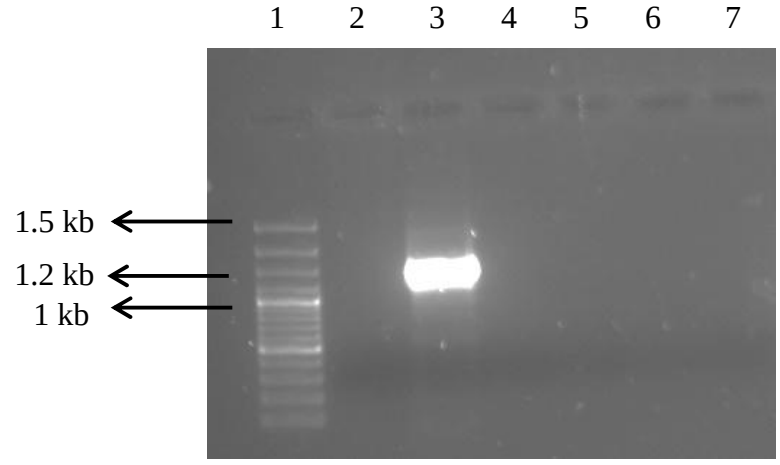


Figure 4.3.2.3: PCR amplification of the 1.3 kb *DEase*II gene product from clone LD4 using specific primer pairs. Lane 1) GeneRuler 1 kb plus™ DNA molecular mass marker (Fermentas). Lane 3) Amplicon of *DEase*II . Lane 4) Negative control. Lane 5) Reverse primer control. Lane 6) Forward primer control. Lane 7) Control reaction using *E. coli* genomic DNA.

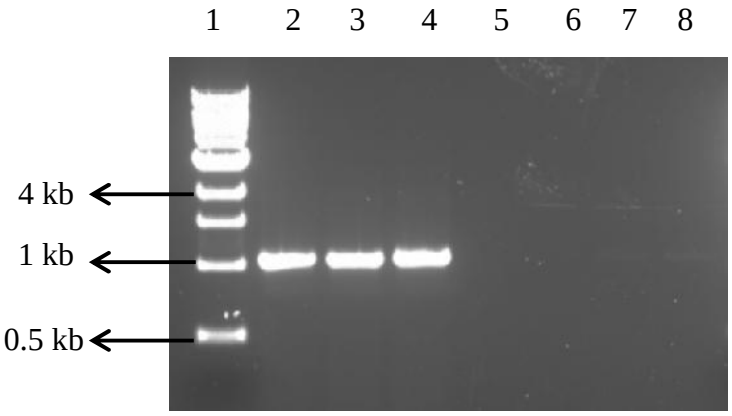


Figure 4.3.2.4: PCR amplification of the 1.07 kb *DEase*IV gene product from clone LD4 using specific primer pairs. Lane 1) 1 kb DNA molecular mass marker (NEB). Lane 2 - 4) Positive reactions. Lane 5) Negative control. Lane 6) Reverse primer control. Lane 7) Forward primer control. Lane 8) Control reaction using *E. coli* genomic DNA.

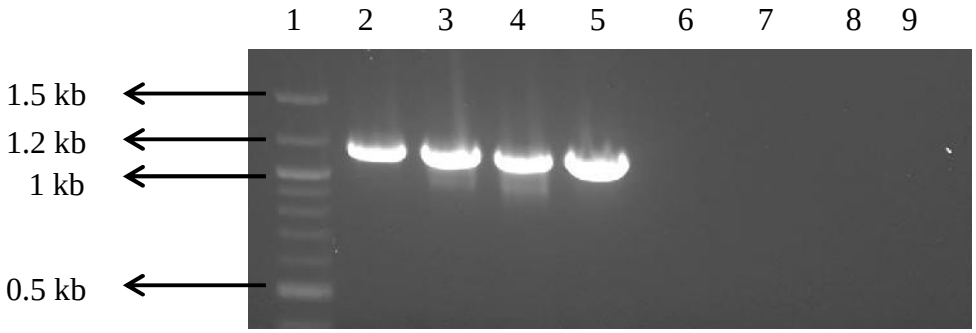


Figure 4.3.2.5: PCR amplification of the 1.15 kb *DEase*V gene product from clone LD13 using specific primer pairs. Lane 1) GeneRuler 1 kb plus™ DNA molecular mass marker (Fermentas). Lane 2 - 5) Amplicons of *DEase*V. Lane 6) Negative control. Lane 7) Reverse primer control. Lane 8) Forward primer control. Lane 9) Control reaction using *E. coli* genomic DNA.

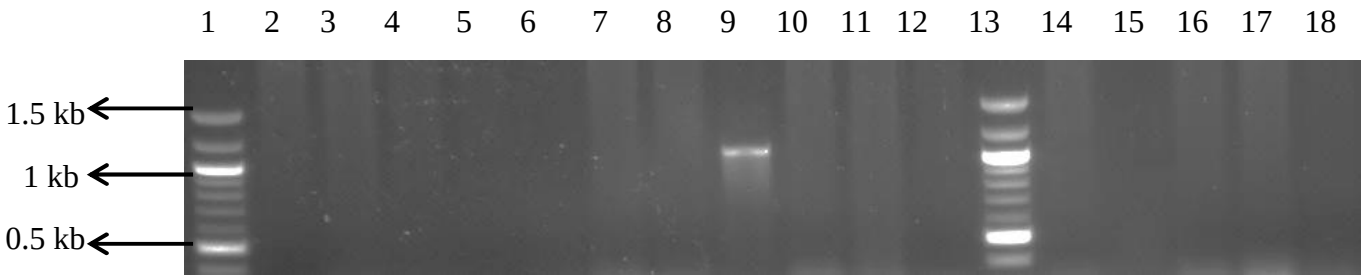


Figure 4.3.2.6. Colony PCR of randomly selected *DEase*IV-pET28a clones and pET28a parental vector containing clones. Lanes 1 and 13) GeneRuler 1 kb plus™ DNA molecular mass marker (Fermentas). Lane 2 – 12) Randomly selected *DEase*IV-pET28a clones with amplified gene product of interest from clone 8. Lane 14 and 15) pET28a parental vector containing clone. Lane 16) Negative control. Lane 17) Reverse primer control. Lane 18) Forward primer control.

Putative N-terminal signal peptide scores were only significant in *DEaseI*, an enzyme which most likely originated from a Gram negative organism. The lower SignalP score for *DEaseII* indicated that a signal sequence was unlikely to occur on the N-terminus of this gene. However, *DEaseII* was cloned into the same vector system as *DEaseI* (pET21a) in order to prevent possible difficulties in purification due to an N-terminal cleavage event. No signal peptides were detected in genes *DEase*IV or *DEase*V, and these were cloned into the pET28a and pCold vectors systems, respectively. In the case of *DEaseI*, *DEaseII* and *DEase*V, successful ligation into the vector systems was observed by subsequent transformation of GeneHog *E. coli* strains and hydrolytic activity of transformants on tributyrin agar. Successful cloning of *DEase*IV was confirmed by alkaline lysis plasmid extraction of random transformants followed by restriction digest analysis. The inability of insert containing clones to hydrolyse tributyrin further confirmed the transposon mutagenesis studies and eliminated a possible novel function for this enzyme. Both gene orientation and sequence were confirmed by sequencing plasmids extracted from the positive clones. In addition, the retention of

hydrolytic activity in subclones also indicated that the ORF predictions were accurate in terms of gene size, orientation and sequence (Figure 4.3.2.7). Based on the rare codon analysis, plasmids extracted from the verified clones (excluding *DEase*IV) were used separately to transform the Rosetta (DE3) pLysS *E. coli* expression host.

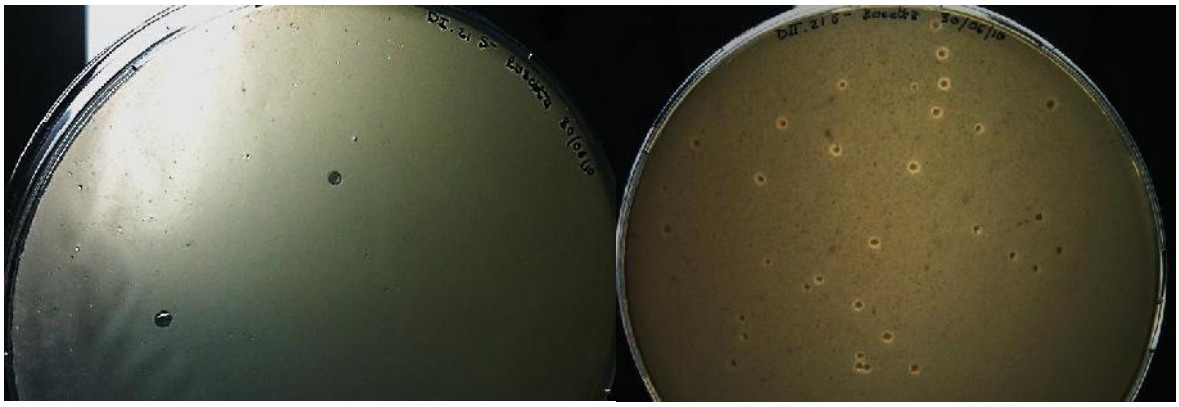


Figure 4.3.2.7 Zones of hydrolysis surrounding colonies following the screening of transformants on tributyrin agar. *DEase*I.21S$^-$ is shown on the left while *DEase*II.21S$^-$ is on the right.

4.3.3 Protein expression and purification

Two enzymes were the focus of further study; *DEase*I and *DEase*II. Although *DEase*V showed tributyrin hydrolysing activity, both *DEase*I and *DEase*II showed greater sequence novelty. Furthermore, the two genes occurred on different fosmid clones and clustered in the same HSL lipolytic family, but appeared to be distinct from each other in terms of structure and host organism. *DEase*I most likely originates from a Gram + psychrotrophic microorganism (*Actinobacterium*) while the likely origin of *DEase*II is a Gram – microbe considered to be a true psychrophile (*Psychrobacter*) [See Chapter 3].

*DEase*I.21S$^-$, *DEase*II.21S$^-$ and the pET21a parental vector were transformed into the Rosetta (DE3) pLysS expression host. Small scale expression studies performed at 16 °C and analysed by SDS-PAGE showed that *DEase*I was overexpressed in the soluble fraction after

induction with 0.4 mM IPTG (Figure 4.3.3.1). Attempts to optimise expression of *DEase*II using a variety of IPTG concentrations (0.4 mM, 0.8 mM, 1.0 mM, 1.2 mM and 2 mM) did not result in overexpression of the expected 52 kDa band of interest, prompting the use of a different approach. During functional screening, *DEase*II transformants formed zones of clearance around the colonies after a three day incubation at 4 °C in comparison to the overnight incubation of *DEase*I at this temperature. If *DEase*II was indeed a truly psychrophilic enzyme, the temperature and time of incubation may play a critical role in protein overexpression. To test this, *DEase*II cultures were induced with 0.8 mM IPTG and grown at 16 °C for two days followed by further 2 day incubation at 4 °C. Analysis of the resultant protein fractions revealed the presence of a slightly overexpressed protein band of the expected size in the soluble fraction of clarified preparations. This experimental procedure was repeated and the parental vector control was included. SDS-PAGE analysis again revealed the protein band of interest (Figure 4.3.3.2).

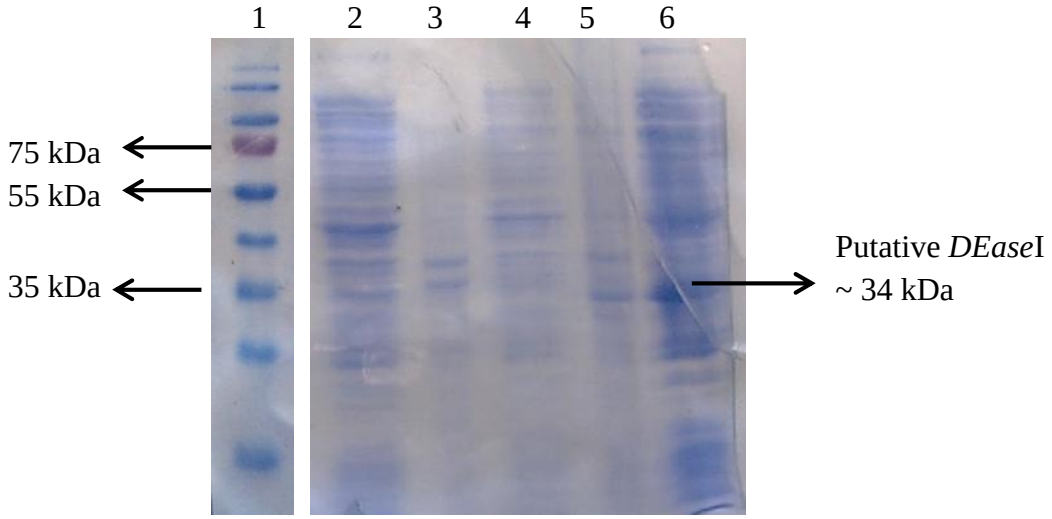


Figure 4.3.3.1: SDS-PAGE analysis of the protein expression profile of *DEase*I-pET21a vs. pET21a parental vector, both in Rosetta (DE3) pLysS. An overexpressed band in the expected size range calculated for *DEase*I is indicated. Lane 1) PageRuler™ Prestained protein marker. Lane 2) Induced insoluble fraction (0.4 mM IPTG) of pET21a vector

control. Lane 3) Induced soluble fraction (0.4 mM IPTG) of pET21a vector control. Lane 4) Extracellular TCA precipitated fraction from *DEase*I. Lane 5) Induced insoluble fraction (0.4 mM IPTG) of *DEase*I. Lane 6) Induced soluble fraction (0.4 mM IPTG) of *DEase*I.

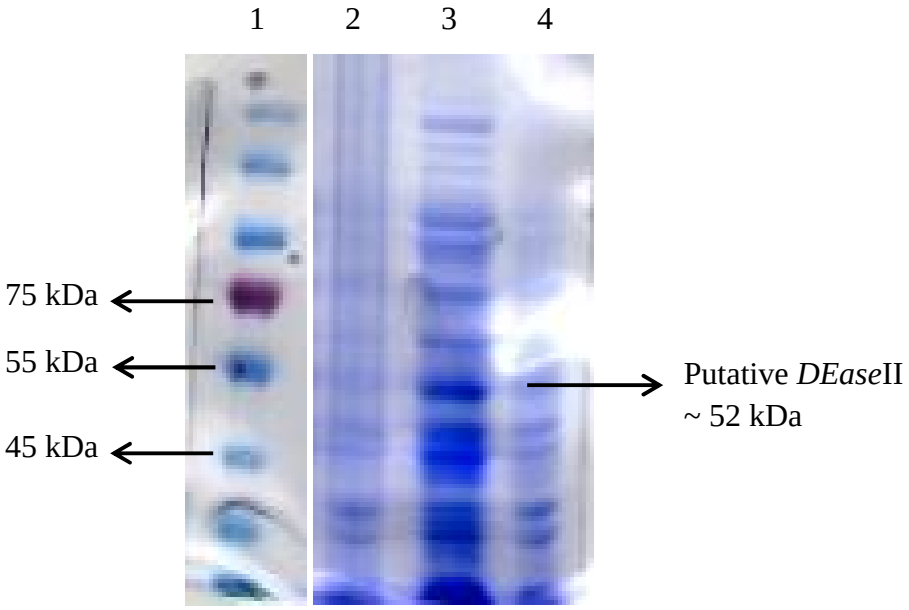


Figure 4.3.3.2: SDS-PAGE analysis of the protein expression profile of *DEase*II-pET21a vs. pET21a parental vector, both in Rosetta (DE3) pLysS. An overexpressed band in the expected size range calculated for *DEase*II is indicated. Lane 1) PageRuler™ prestained protein marker. Lane 2) Extracellular TCA precipitated fraction from *DEase*II. Lane 3) Induced soluble fraction (0.8 mM IPTG) of *DEase*II. Lane 4) Induced insoluble fraction (0.8 mM IPTG) of *DEase*II.

Purification strategies for lipolytic proteins are generally multi-step processes. No standard protocol exists and strategies must be designed for each individual protein target. Additionally, attempts to purify lipases from cold adapted microorganisms tend to be unsuccessful due to a strong association of enzymes with lipopolysaccharides produced by

the microorganisms. Some strategies have involved pre-purification steps whereby proteins are precipitated using ammonium sulphate or ethanol followed by further purification using chromatography techniques. Ion exchange chromatography is the most commonly used method and has been successfully employed for purification of the lipases from *Pyrobaculum calidifontis*, *Pichia burtoni* and *Proteus vulgaris* (Sharma *et al*., 2002). Histidine affinity purification is also utilised for the one-step purification of lipolytic enzymes such as Lipo1 (Roh and Villatte, 2008), EstMY (JunGang *et al*., 2010), EstPS2 (Bunterngsock *et al*., 2010), EstA (Soror *et al.*, 2009) and EstE1 (Rhee *et al*., 2005).

The prepared soluble extracts from *DEase*I and *DEase*II were subjected to metal affinity purification. In the case of *DEase*I, construction of the appropriate His-tagged fusion protein allowed for one-step purification of the enzyme following protein overexpression. *DEase*I was eluted with 250 mM imidazole and analysis by SDS-PAGE revealed a single band of ~34 kDa, the expected size of the enzyme monomer (Figure 4.3.3.3). Experiments were repeated in triplicate with successful recovery of the protein in all cases. Dialysed fractions of *DEase*I were quantified and the protein concentrations for the first and second expression and purification experiments were 0.17 mg/ml and 0.07 mg/ml respectively.

Attempts to purify *DEase*II, however, were not as simple. Although imidazole concentrations of the various buffers were changed, *DEase*II continuously appeared in the wash fractions of His-affinity purification. This indicated that *DEase*II bound to the column with very low affinity. Production of lipases is highly influenced by physicochemical factors such as pH, N– and C– sources, temperature and nutritional factors (Joseph *et al*., 2008; Gupta *et al*, 2004). As previously mentioned, the homology model of *DEase*II (Figure 4.3.1.9) revealed a unique β-sheet extension originating from residues 289-315, approximately 10 residues from the catalytic serine (towards the C-terminal). Domain insertions may shape the substrate binding sites and may even be large enough to seal off the catalytic cavity (Nardini and

Dijkstra, 1999). Although the hexahistidine residues on the C-terminus of the protein do not appear to be buried within the molecule, possible movement of the central extension may mimic a movable lid and perhaps limit exposure of the tag to the column. If the use of lipid based carbon sources could induce lipase production, perhaps something similar could be utilised in this study to induce a conformational change in *DEase*II structure, thereby creating an 'open' conformation, greater access of the tag and therefore higher binding capacity to the affinity column. For example, co-crystallisation of the *P. aeruginosa* PAO1 lipase (PAL) with the covalently bound substrate analog showed the cap domain (residues 109 – 163, forming 4 α-helices) in an open conformation providing substrate and solvent access to the active site (Nardini, 2000). In order to test this hypothesis, a liquid emulsion of tributyrin and gum arabic was prepared and added to the soluble extracts obtained for *DEase*II. Following a half hour incubation at 4 °C, the extract was loaded onto the affinity column. Protein was eluted in buffer containing 500 mM imidazole and SDS-PAGE analysis revealed a single protein band of ~52 kDa, which corresponds to the size expected for *DEase*II (Figure 4.3.3.4). The dialysed fraction was quantified and protein concentration determined to be 0.4 mg/ml. Proteins were stored at 4 °C until further analysis.

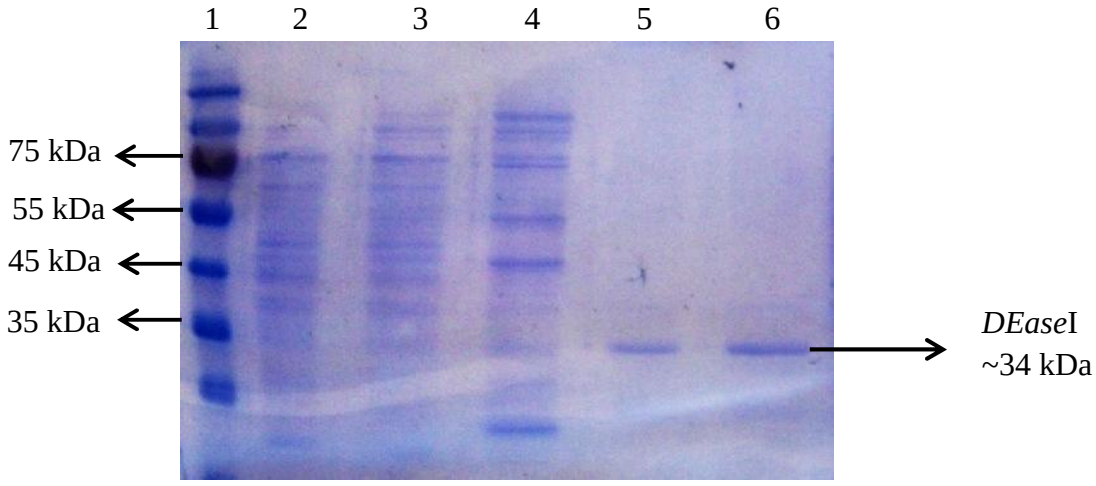


Figure 4.3.3.3: SDS-PAGE analysis of affinity purification for *DEase*I. Lane 1) PageRuler™ Prestained protein marker (Fermentas). Lane 2) *DEase*I flow through fraction. Lane 3) *DEase*I binding fraction. Lane 4) *DEase*I wash fraction. Lane 5 and 6) *DEase*I elute fractions.

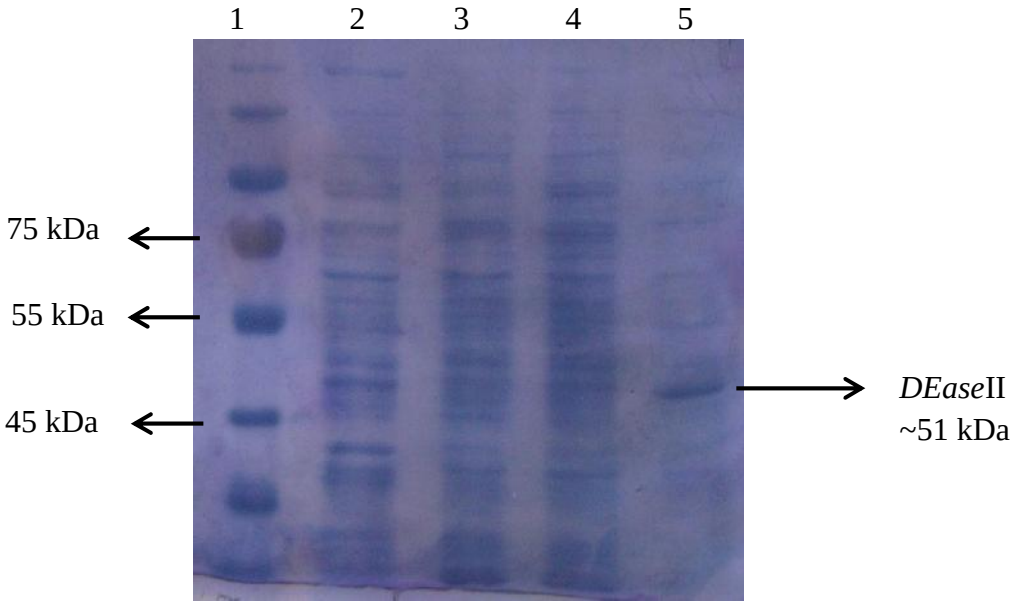


Figure 4.3.3.4: SDS-PAGE analysis of affinity purification for *DEase*II. Lane 1) PageRuler™ Prestained protein marker (Fermentas). Lane 2) *DEase*II flow through fraction. Lane 3) *DEase*II binding fraction. Lane 4) *DEase*II wash fraction. Lane 5) *DEase*II elute fraction.

4.3.4 Characterisation using *p*-Nitrophenyl Ester substrates

The most commonly used enzymatic method for the characterisation of lipolytic enzymes involves monitoring the hydrolysis of *para*-Nitrophenyl ester substrates of varying chain lengths at 405 nm (Winkler and Stuckman, 1979; Gupta *et al*, 2002). Esterases tend to follow Michaelis Menton kinetics, where activity is a function of the substrate concentration and the maximal rate is achieved at substrate saturation.

In the case of *DEase*II, kinetic characterisation using these substrates could not be achieved as the reaction rates were too slow for accurate analysis. Visual comparisons using *p*-NP decanoate, to control reactions containing no enzyme could be made only after an incubation periods of 12 hours. Additionally, these observations had to be made with long chain fatty acid substrates which do not undergo spontaneous non-enzymatic hydrolysis. Although *DEase*II does show activity on the long chain molecules, lipase-like activity could not be confirmed. It is therefore apparent that the *p*-NP ester substrates are not the natural substrates for this enzyme and are not suitable for characterisation.

*DEase*I was successfully characterised with regards to substrate preference, temperature and pH optima, thermal liability and tolerance to sodium chloride, using the *p*-NP esters. *DEase*I showed a preference for short- to medium- chain fatty acid esters. No activity was detected on substrates with chain length longer than C10, clearly indicating that *DEase*I is a carboxyl esterase. The substrate preference is comparable to other members of the HSL family which show specificity towards C2, C4, C6 and C8 substrates (Virk *et al*., 2011; Bunterngsock *et al*., 2010; Roh and Villatte, 2008; JunGang *et al*., 2010; Kim *et al*., 2005).

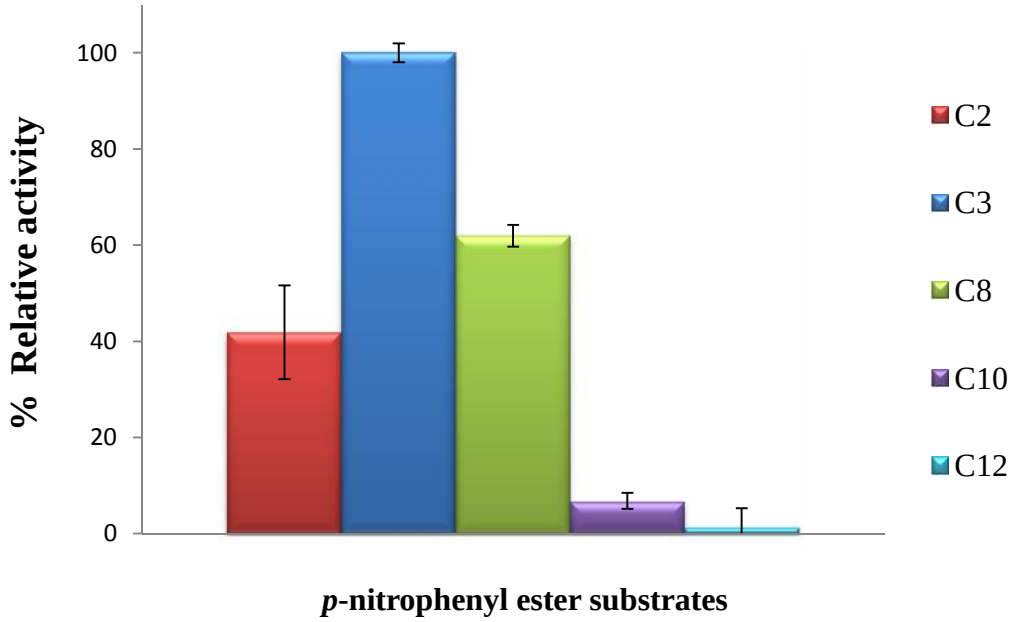
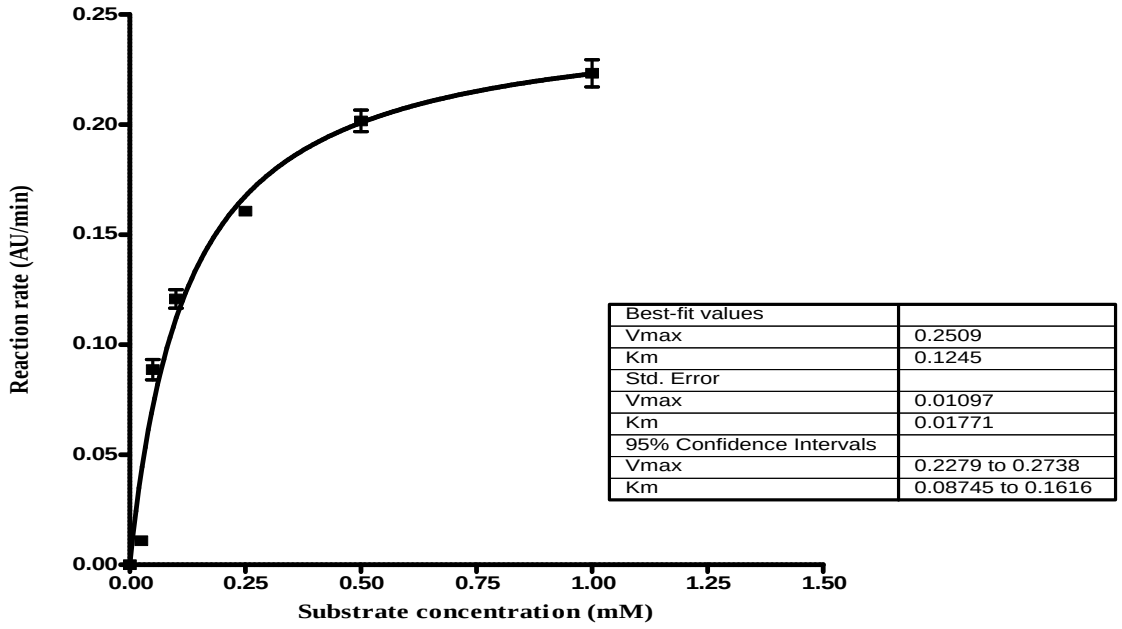


Figure 4.3.4.1: Activity of *DEase*I towards *p*-nitrophenyl esters of varying chain lengths. Activity against *p*NP- propionate was taken as 100 %. All substrates were tested at 25 °C in sodium phosphate buffer (pH7.5).

For all the substrates tested, the rate of formation of *p*-Nitrophenol was directly proportional to the amount of enzyme used in the assays. Additionally, there was a hyperbolic dependence of the rate on substrate concentration. Controls showed no appreciable activity towards any substrates tested. Data from the assays were analysed using the Graphpad Prism software and the K_M of *DEase*I was determined to be 0.1245 mM for *p*-NP propionate, while the specific activity was 85.3 µmol.min^{-1}.mg^{-1}. The k_{cat} of *DEase*I was 48 s^{-1}. Lipolytic enzymes from the HSL family demonstrate a wide range of substrate specificities, generally on short to medium chain fatty acid esters. The substrate preference of *DEase*I compares well to other enzymes of this family; EstPc showed the highest activity on C6 and was also able to hydrolyse *tert*-butyl acetate (Hotta *et al*, 2002), SshEStI catalysed C4 with greatest efficiency (Ejima *et al*., 2001), as did EstPS2 (Bunterngsock *et al*., 2010), Lipo1 (Roh and Villatte, 2008), the homotrimer

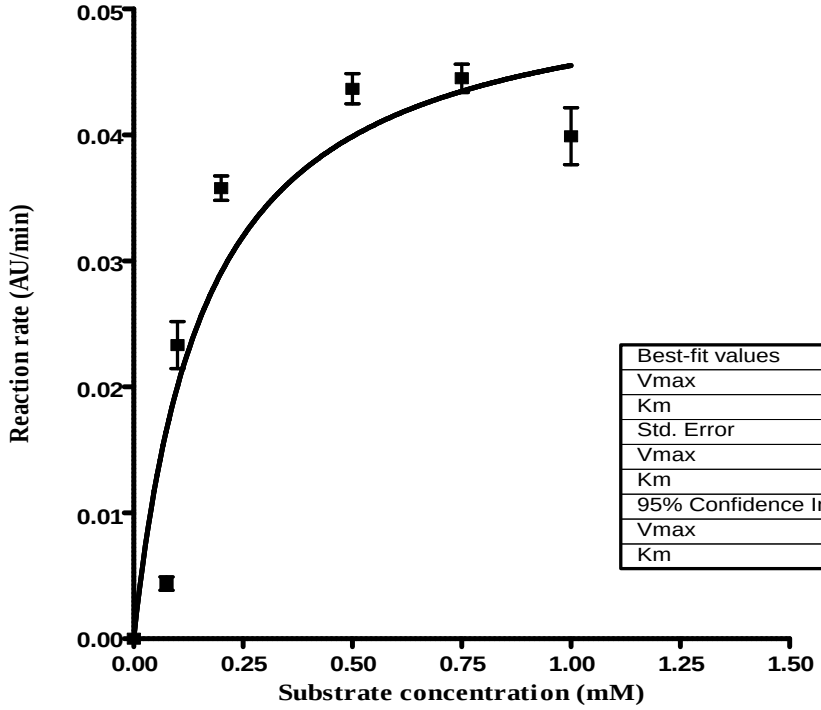
REst1 (Virk *et al*., 2011) and crude extracts of Lip2 from Moraxella spp TA144 (Choo *et al*., 1998). LipP, a cold adapted enzyme from the psychrophilic *Pseudomonas* catalysed C6 with high affinity. Other enzymes such as EstMY (JanGang *et al*., 2010) showed greatest activity on C8. Furthermore, the specific activity of *DEase*I also compares well with other family IV esterase and lipase members such as Est25 [63.7 U/mg] (Kim *et al*., 2005), Lipo1 [150 U/mg] (Roh and Villatte, 2008), EstPS2 [128 U/mg] (Bunterngsock *et al*., 2010) and EstAT11 [59.8 U/mg] (Jeon *et al*., 2009).

A



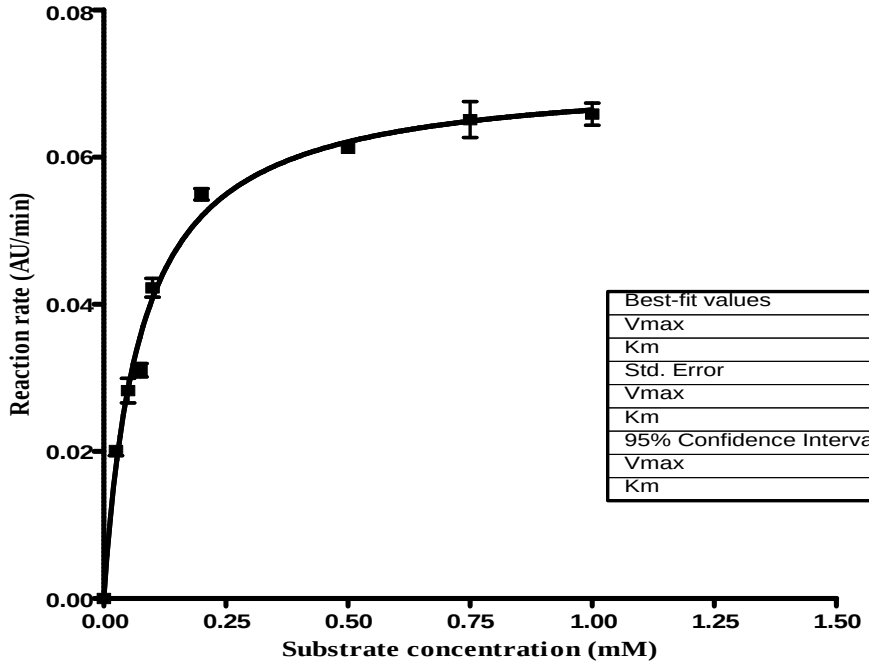
Best-fit values	
Vmax	0.2509
Km	0.1245
Std. Error	
Vmax	0.01097
Km	0.01771
95% Confidence Intervals	
Vmax	0.2279 to 0.2738
Km	0.08745 to 0.1616

B



Best-fit values	
Vmax	0.05301
Km	0.1648
Std. Error	
Vmax	0.004664
Km	0.04753
95% Confidence Intervals	
Vmax	0.04325 to 0.06278
Km	0.06534 to 0.2643

C



Best-fit values	
Vmax	0.07137
Km	0.07486
Std. Error	
Vmax	0.001123
Km	0.004338
95% Confidence Intervals	
Vmax	0.06905 to 0.07368
Km	0.06592 to 0.08379

Figure 4.3.4.2 : Michaelis-Menten non-linear regression curve analysis for *DEase*I on A] C3 (*p*-nitrophenyl propionate) B] C2 (*p*-nitrophenyl acetate) and C] C8 (*p*-nitrophenyl octanoate). Graphs were constructed using the Graphpad prism software.

Table 4.3.4.1: Kinetic parameters determined for *DEase*I on 0.5 mM substrate at 25 °C in sodium phosphate buffer (pH 7.5).

Substrate	K_m (mM)	V_{max}(U/mg)	k_{cat} (s^{-1})	k_{cat}/K_m($mM^{-1}s^{-1}$)
C2	0.168 ± 0.048	22.4	13	77.2
C3	0.125 ± 0.018	85.3	48	385.5
C8	0.075 ± 0.001	31.44	18	240

The pH optimum for *DEase*I was determined using *p*-nitrophenyl octanoate as substrate due to non-enzymatic hydrolysis of short chain ester substrates at acidic and alkaline pH values. A pH optima of 8.5 and a temperature optimum of 25 °C was observed for *DEase*I (Figures 4.3.4.3 and 4.3.4.4). Abiotic conditions experienced by microbes inhabiting the Dry Valley soils of Antarctica include high salt, low temperature and an alkaline pH (Balks and Campbell, 2001). As such, the results obtained for temperature optima, pH optima and the effect of varying salt concentrations on enzyme activity were expected. Thermal stability assays were conducted at 16, 25, 35 and 45 °C, and indicated a short half-life of 15 minutes for *DEase*I at 25 °C (Figure 4.3.4.5). No enzymatic activity was observed following incubation of the enzyme for 15 minutes at 45 °C, indicating that *DEase*I is sensitive to thermal inactivation. The results for *DEase*I are in agreement with those obtained for other cold-adapted HSL enzymes. Lipolytic enzymes such as H1Lip1 and PsyEst are also extremely temperature sensitive. The temperature optimum for H1Lip1 was 35 °C but the enzyme was unstable at 25 °C and rapidly inactivated at 5 °C above the optimum (Hårdeman and Sjöling, 2007). PsyEst from *Psychrobacter* ANT300 also had a temperature optimum of 35 °C and a half-life of 15 minutes at 5 °C above the optimum (Kulakovaa *et al*., 2004).

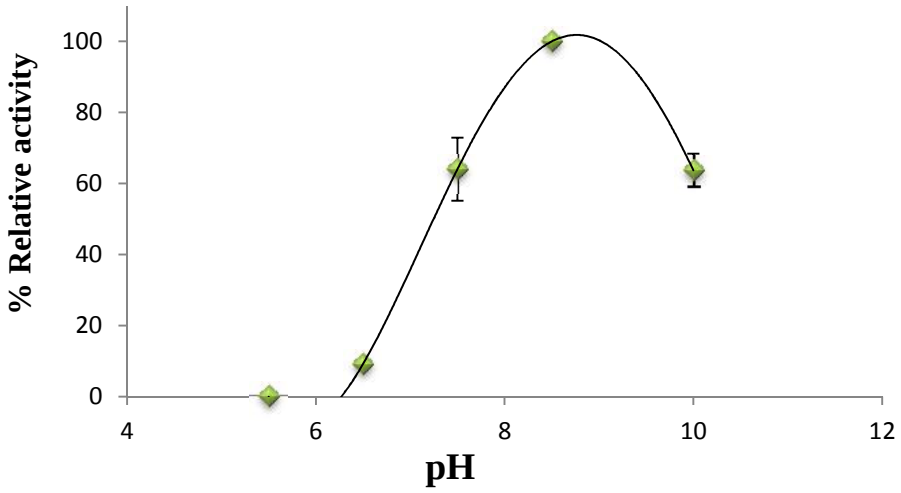


Figure 4.3.4.3: The pH optima for *DEase*I. The use of different buffers made no difference to the activity measured at an overlapping pH. Activity is expressed as a percentage relative to the standard assay conditions at 25 °C using *p*-NP propionate as substrate.

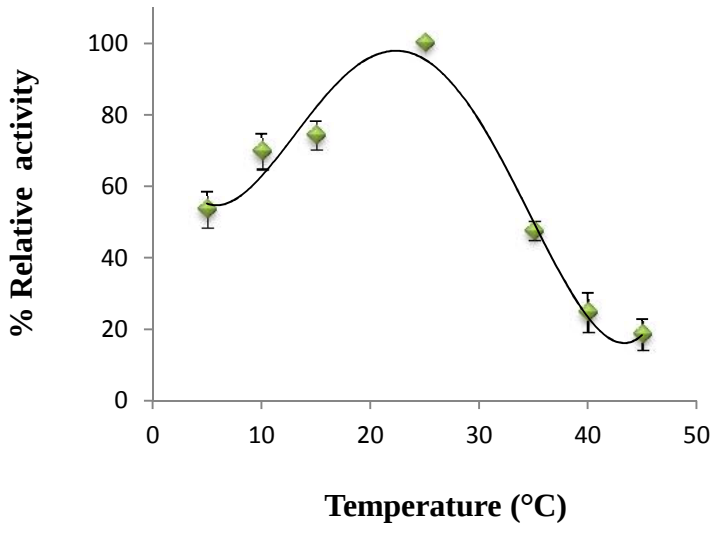


Figure 4.3.4.4: The temperature optima of *DEase*I. The activity is expressed as a percentage relative to the standard assay conditions in sodium phosphate buffer (pH7.5) using *p*-NP propionate as substrate.

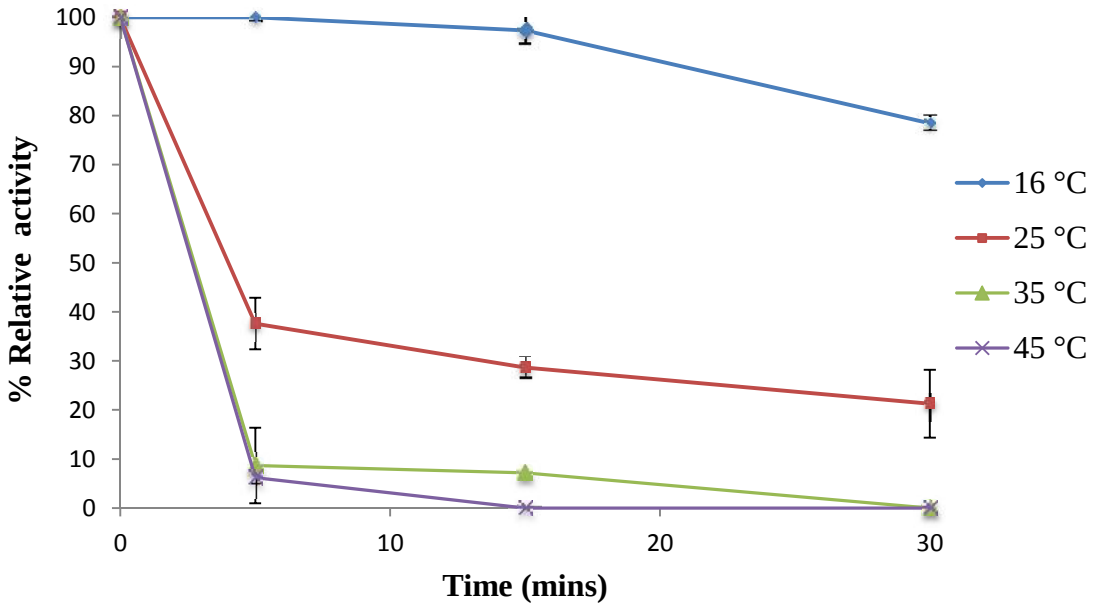


Figure 4.3.4.5: The thermal inactivation profile of *DEase*I at 16 °C (♦), 25 °C (■), 35 °C (▲) and 45 °C (X). Activity is expressed as a percentage relative to time zero in the standard assay at 25 °C, towards pNP-propionate.

In order to determine possible effects of high solute concentrations, varying concentrations of NaCl were added to buffers and the effect on enzyme activity was examined (Figure 4.3.4.6). *DEase*I retained 50 % of activity at 1 M NaCl with only 20 % residual activity at 4 M NaCl. To the author's knowledge, there are no reports which demonstrate the effect of sodium chloride on lipolytic enzyme activity. Table 4.3.5.1 summarises the experimental procedure followed for each enzyme in this study.

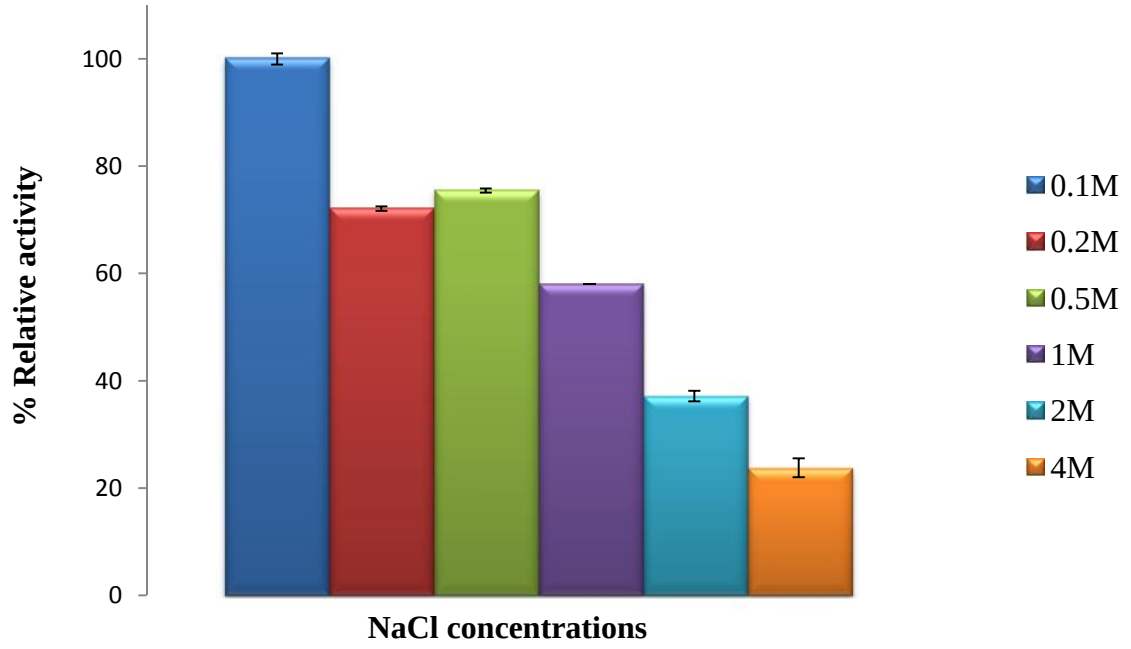


Figure 4.3.4.6: The effect of varying concentrations of NaCl in the used in the sodium phosphate assay buffer towards *DEase*I. The activity is expressed as a percentage relative to the standard assay conditions at 25 °C using *p*-NP propionate as substrate.

4.3.5 FPLC analysis

Following size exclusion chromatography of the protein standards, the retention time obtained for *DEase*I indicated that the enzyme formed a dimer in solution (Figure 4.3.5.1). There are several reports of multimeric proteins in the HSL family. Protein subunits of the lipolytic enzyme, AFEST, interact via salt bridges, hydrogen bonds and van der Waals forces to form dimers, but it is reported to be active as a monomer (De Simone *et al*., 2001; Manco *et al*., 2002). Similarly, SshEstI from *Sulfolobus shibatae* forms both dimers and trimers in solution (Ejima *et al*., 2001) and EstPc from *P. calidifontis* is active as a trimer (Hotta *et al*, 2002). It must, however, be noted that all the above mentioned enzymes are from hyperthermophilic sources. In the case of *DEase*I, it was interesting to observe possible

dimerization for a cold-adapted enzyme. However, dimerization may be not functionally significant if both active sites are fully accessible to substrate, as is the case with BFAE, the Brefeldin A esterase from *B. subtilis* (Wei *et al*., 1999). Dimer formation may therefore be expected to decrease the structural plasticity, and exposure to cold temperatures tends to give rise to dissociation of multimeric proteins (Gerday *et al*., 1997; Privalov and Makhatadze, 1990). However, salt bridges do not need to be completely absent in order for cold adaptation to occur in proteins. When the cold adapted lipase of *P. glumae* was compared to a mesophilic enzyme from *Xanthobacter autotrophicus*, it was observed that two of the four salt bridges were no longer present in the psychrophilic enzyme, thereby increasing the flexibility of the active site (Gerday *et al*., 1997). In addition, cold denaturation may not occur in some proteins; for instance, denaturation was not observed in the psychrophilic lactate dehydrogenase, as well as some other proteins (Privalov, 1990).

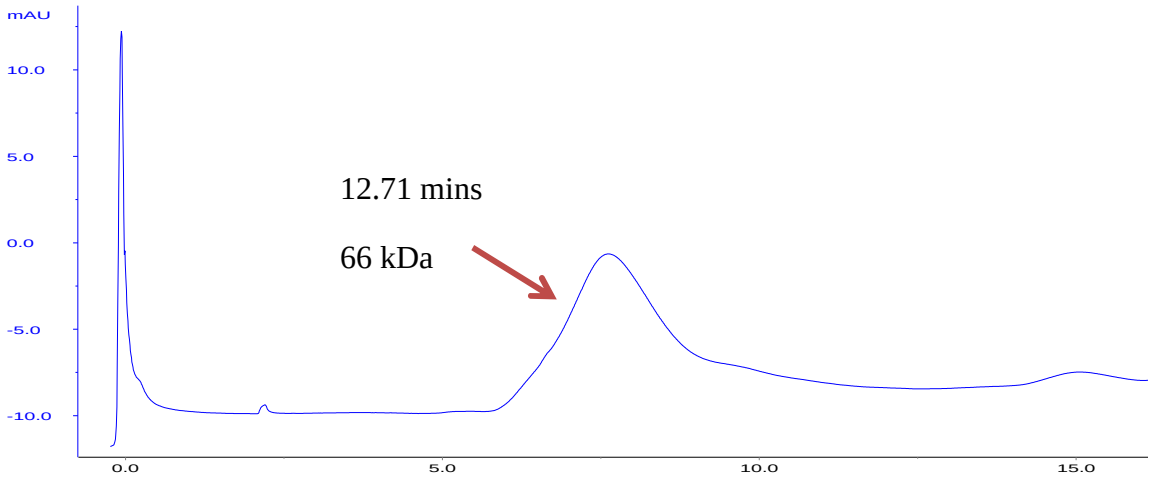


Figure 4.3.5.1: Graph of size exclusion FPLC of IMAC purified fractions of *DEase*I in 50 mM Tris-HCl (pH 7.5) and 400 mM NaCl. The blue line indicates the UV peak. X-Axis measurements are in ml (with fractions collected every 1 ml volume). The retention times and corresponding molecular mass estimates are indicated.

Table 4.3.5.1: Summary of experimental work performed on putative lipolytic enzymes in this study. Brackets highlight the main reason for discontinuing further study on a particular enzyme.

	***DEase*I**	***DEase*II**	***DEase*IV**	***DEase*V**	***DEase*VI**
Fosmid clone	LD7	LD4	LD4	LD13	LD4
ORF designation	7ORF28	4ORF5	4ORF3	13ORF9	4ORF4
Sub-cloning successful	YES	YES	YES	YES	NO
Tributyrin hydrolysis following sub-cloning	YES	YES	NO (MPE enzymes do not hydrolyse tributyrin)	YES	NO (MPE enzymes do not hydrolyse tributyrin)
Protein expression	YES	YES	NO	NO (No comparison to additional family VIII esterase in this study)	NO
Kinetic characterisation	YES	NO (*p*-NP esters not the natural substrate)	NO	NO	NO

4.4 Conclusion and future considerations

This study shows the value of metagenomics coupled to next generation sequencing technology for the identification of novel enzymes with functional characteristics. Bioinformatic analysis allowed for the accurate prediction of putative lipolytic genes in three sequenced fosmids. Predictions were verified by sub-cloning the respective genes and observing retention of functional lipolytic activity. In *DEase*II, homology modelling revealed an interesting structural modification consisting of two anti-parallel β- sheets. In the case of AFL (*Archeoglobus fulgidus* lipase), it was noted that the C-terminal β- barrel domain was essential for binding long chain fatty acid substrates. This domain interacted with the N-terminal lid and assisted in protein stability at high temperatures (90 °C) and alkaline pH (pH 10-11) as well as influencing the enzymes catalytic efficiency (Rusnak *et al*., 2005). As is often the case for an enzyme from a psychrophilic environment, heat inactivation occurs rapidly; perhaps the unusual extension of *DEase*II offers protection from such temperature fluctuations that take place in the summer months on the Antarctic continent. The inserted domain in *DEase*II may also be responsible for its substrate specificity, as is the case for PAF-AH's of Family III (Arpigny and Jaeger, 1999).

*DEase*I was purified to homogeneity using one-step affinity chromatography. The enzyme was successfully characterised using the *p*-nitrophenyl ester substrates. Enrichment of the esterase/lipase enzyme pool could possibly be an immediate source of genetically modified enzymes which may be optimised for industrial application by directed evolution studies. This enzyme could therefore also be accessed for possible industrial application, particularly for resolution of heat-labile racemic mixtures or in bioremediation schemes. For example; Jeon *et al* (2009) characterised two cold-adapted metagenome derived esterases and the ability to hydrolyse racemic ofloxacin esters was investigated. Both EstAt1 and EstAt11

belonged to the HSL family of lipases and esterases and EstAt11 hydrolysed the (*S*)-ofloxacin butyl ester with enantiomeric excess of 70.3 % (Jeon *et al*., 2009). Similarly, the family IV esterase Est25 was capable of hydrolysing racemic (*R, S*) - ketoprofen ethyl ester with a preference toward the (*R*) – ketoprofen enantiomer (Kim *et al*., 2005).

A solved crystal structure of *DEase*II could highlight interesting modifications for this group of proteins, particularly with respect to the unique β-sheet extension observed in this enzyme. Furthermore, both enzymes may be excellent candidates for structural comparison studies to corresponding mesophilic and thermophilic homologs.

In addition, future studies should include the expression, purification and characterisation of *DEase*IV and *DEase*V. One candidate artificial substrate to determine possible phosphodiesterase activity of *DEase*IV would be *bis*-nitrophenylphosphate. *DEase*V could exhibit promiscuous activity and should be analysed for esterase activity using *p*-nitrophenyl esters and β- lactamase activity using β-lactam drugs.

Chapter 5: From genetic potential to function- Water HYpersensitivity response gene

5.1 Introduction

Organisms living in natural environments are subject to multiple simultaneous abiotic stress conditions. Organisms experiencing the influence of abiotic stress on natural selection and fitness to the greatest degree live in environments where the conditions exceed the ranges required for normal growth and development (Roelofs *et al*., 2008). The Antarctic Dry Valley soils are considered to be the most extreme xerophytic environments on Earth and the microbial communities inhabiting these soils have adapted diverse mechanisms to cope with extremely low temperatures, low nutrient availability and elevated levels of oxidative stress. In these sandy gravel sediments, the water potential is decreased even further due to high levels of mineral salts, which result when the effects of evaporation exceeds that of leaching (Potts, 1994; Mahajan and Tuteja, 2005).

5.1.1 Desiccation stress

Water molecules are critical components of living systems. Accordingly, low water potential is considered to be the most life threatening abiotic stress as it negatively affects all biological functions (Kriško *et al*., 2010). Water potential is required to confer structural order to cells, stabilise proteins, lipids and nucleic acids, as well as for the establishment of a cellular microenvironment in which vital metabolic systems and chemical reactions are maintained (Billi and Potts, 2002; Potts, 1994). The scarcity of free water is, however, not the only danger associated with arid conditions. Hypo-osmotic shock, which is caused by sudden rehydration events, results in cell bursting or plasmolysis (Csonka, 1989). In addition, cold and desiccation are invariably linked. A droplet of water entering the gaseous phase from the liquid state is accompanied by a decrease in enthalpy, whereby bacterial cells that are exposed to air-drying conditions will also experience rapid cooling (Potts, 1994).

There are several changes associated with hyper-osmotic shock and the instant efflux of water out of cells, including cell shrinkage, changes in cell shape, colour and texture, an increase in intracellular salt concentration and cellular viscosity, and macromolecular crowding (Potts *et al*., 2005; Kriško *et al*., 2010; Csonka, 1989). An increase in the phase transition temperature occurs upon dehydration of phospholipid membranes due to increased van der Waals interactions between adjacent lipids. This results in decreased flexibility of cellular membranes, which is particularly unfavourable in cold environments where membrane plasticity is essential for growth (Potts, 1994). In addition, Maillard reactions occur in sugars and amino acid-containing molecules and therefore cause modification of proteins and nucleic acids. Haber-Weiss and Fenton reactions cause an accumulation of reactive oxygen species and free radicals, resulting in the oxidation and de-esterification of lipids (Mahajan and Tuteja, 2005). In addition, changes in enzyme conformation and/ or electron transport chains caused by protein dehydration also leads to the accumulation of reactive oxygen species. This in turn causes further peroxidation of lipids and proteins as well as DNA damage (Potts, 1994). Extensive DNA mutations by chemical modification such as acylation and depurination, and protein denaturation eventually lead to cell death (Potts *et al*., 2005).

Desiccation can influence the structure of proteins, and consequently affects function. For example; proteins subjected to low water concentrations undergo re-ordering of disulphides and side chains, thereby preventing changes in the protein backbone which ultimately leads to rigid conformations (Potts, 1994). The effects of desiccation are clearly as detrimental to cellular functions as extreme low temperatures. In combination, low relative humidity and low temperature are possibly the most unfavourable conditions for life in any given environment.

5.1.2 Desiccation survival strategies

Drought tolerance is an important element of the survival of organisms in environments with low water availability while maintaining high internal water concentrations (Berjak, 2006). A drought tolerant organism must therefore also be desiccation tolerant in order to survive the loss of intracellular water. In addition, desiccation tolerance is not simply the ability to survive extreme water loss, but implies continued survival for an extended period of time with the capacity for rehydration, i.e. dormancy (Berjak, 2006). This property allows for increased fitness and the ability to proliferate in ecological niches, and has evolved in a diverse number of taxa including nematodes, yeast cells and fungi, higher and lower plants and tardigrades, bdelloid rotifers, bacteria and archaea (Kriško *et al*., 2010). Interestingly, emerging evidence shows similar mechanisms of desiccation tolerance irrespective of taxonomic classification (Berjak, 2006).

Under low water conditions desiccation tolerant organisms may enter a dormant state known as anhydrobiosis. They do not grow, as residual water is so low (approximately 0.02 grams of water per gram of cell weight) that a monolayer around macromolecules cannot be maintained. However, anhydrobiotic cells may rapidly regain metabolic activity as soon as water becomes available (Billi and Potts, 2002; Potts *et al*., 2005). Generally, water deficiency experienced by anydrobiotic cells is considerably higher than that imposed on vegetative cells, even for extreme halophiles (Billi and Potts, 2002).

In prokaryotes, desiccation tolerance is generally attributed to the ability of the microorganisms to efficiently repair DNA damage, scavenge free radicals and accumulate high levels of compatible solutes (Kriško *et al*., 2010). There are two types of proteins which may be upregulated in response to drought stress, including those involved in a direct pathway for combatting stress, such as Heat Shock Proteins, transporters and

osmoprotectants, and those involved in regulatory processes, such as transcription factors, signalling proteins and kinases (Roelofs *et al*., 2008).

5.1.2.1 Compatible solutes

In general, cells maintain a higher internal solute level than that found in the surrounding environment (Kempf and Bremer, 1998). Maintaining positive turgor pressure across the semi-permeable membrane is a vital mechanism for coping with fluctuating osmolarity and water availability in the external milieu. In microorganisms, turgor pressure is adjusted accordingly by controlling the amount of osmotically active solutes in the cytoplasm (Kempf and Bremer, 1998). Another mechanism employed by some bacteria to tolerate hypo- and hyper- osmotic shock is the production of various extracellular polysaccharides (Potts, 1994; Billi and Potts, 2002). For example; in *Nostoc commune*, glycan is produced in large quantities. This is believed to provide a matrix which may maintain proteins related to water stress in an active state. *Chroococcidiopsis*, an Antarctic cyanobacterium, has also been shown to produce thick, multi-layered cell envelopes, and in *Deinococcus radiodurans* genes encoding surface structure proteins are overexpressed in response to desiccation (Billi and Potts, 2002).

Compatible solute accumulation decreases the water potential within cells, thereby preventing water loss into the environment, which would otherwise disrupt membrane structure, protein activity and nucleic acid stability (Mahajan and Tuteja, 2005; Potts, 1994). For example, potassium ions (K^+) serve as a major intracellular osmolyte for the maintenance of turgor and accumulate to high levels in halophilic microorganisms in order to cope with extreme extracellular salt concentration and osmotic pressure (Billi and Potts, 2002; Csonka, 1989). Similarly, intracellular levels of Mn (II) were shown to increase during radiation and desiccation stress in *D. radiodurans* and, while insufficient to provide complete radiation and desiccation tolerance, clearly contribute to the organisms survival by scavenging toxic

oxygen species (Potts *et al*., 2005). Accumulation of kosmotropes is a common adaptive strategy to maintain osmotic balance, and is evident in a number of eukaryotes, as well as prokaryotes (Billi and Potts, 2002; Kempf and Bremer, 1998). Further evidence for the existence of common adaptive strategies across multiple taxa is the nature of the osmoprotectants.

Osmoprotectants are water structure builders which are either synthesised *de novo*, or imported into the cell from the environment (Kempf and Bremer, 1998; Potts, 1994). These organic solvents, or their precursor molecules, enhance growth in hypertonic conditions by counteracting the flow of water out of cells. These solutes may be sugars (trehalose and sucrose), free amino acids (proline and glutamine), polyols (glycerol), quaternary amines (Glycine betaine) and sulphate esters, and are generally highly soluble in water (Kempf and Bremer, 1998; Potts, 1994; Csonka, 1989). Most do not carry a net charge at neutral pH and can therefore accumulate to high levels as metabolically inactive compounds which do not have a negative impact on cellular functions such as DNA replication and enzymatic metabolism (Kempf and Bremer, 1998, Csonka, 1989). These solutes accumulate to molar concentrations and lower osmotic potential, thereby restoring and maintaining cell turgor. In addition, disaccharides, such as sucrose and trehalose, are believed to 'hydrogen-bond' membrane phospholipids and proteins which prevent changes in transition phases (Billi and Potts, 2002).

The supply of compatible solutes in natural environments tends to be both variable and low. These molecules cannot diffuse passively across cellular membranes; transporters are required which exhibit high affinity for the substrates (Kempf and Bremer, 1998). In order to take advantage of the range of solutes in any given environment, micro-organisms may utilise several osmoprotectant transport systems, and the uptake and/ or synthesis of these substances is a tightly regulated and controlled process. For instance; the presence of

osmoprotectant molecules in growth medium will reduce the transcription of genes for the *de novo* synthesis of the particular compound. Some examples of bacterial solute transporters are provided in Table 5.1.3.1. The physiological and genetic responses of bacteria to desiccation in terms of osmoprotectants and compatible solutes are extensively reviewed in Csonka (1989) and Kempf and Bremer (1998).

Extensive studies relating to desiccation stress tolerance and / or adaptation in plant systems have been conducted; however, literature on the bacterial mechanisms is sparse and focuses mainly on the existence of compatible solutes in anhydrobiotic models (Potts, 1994; Roelofs *et al*., 2008). Studies of stress adaptation mechanisms in *D. radiodurans* have shown that mutations in genes for oxygen scavenging and DNA repair resulted in both radiation- and desiccation- sensitive phenotypes, suggesting an overlap in the cellular response to both ionising radiation and dehydration (Billi and Potts, 2002).

The recovery of vegetative prokaryotic cells from water deficit is clearly a multifunctional process, but very little is known regarding the molecular mechanisms employed. Due to the varied affects that water deficit has on cellular processes, it is reasonable to assume that no singular gene, or protein could offer complete protection from this stress, and that a number of synergistic mechanisms could be employed.

Table 5.1.3.1: Examples of transporters for compatible solutes in bacteria

Organism	Transporter	Specificity
Salmonella typhimurium and *E. coli*	ProP and ProU	High affinity for glycine betaine and proline betaine
	Porins OmpC (Hypertonic stress) and OmpF (Hypotonic stress)	Non-specific
Corynebacterium glutamicum	BetP	Glycine betaine
Bacillus subtilis	Opu ABC	Proline

5.1.2.2 Late embryogenesis abundant proteins

The accumulation of compatible solutes is not solely responsible for desiccation tolerance. Studies in *D. radiodurans* have shown that 33 of the 72 genes upregulated during radiation stress were also induced in cultures recovering from desiccation stress (Tanaka *et al*., 2004). One particular group of proteins which are consistently upregulated during salt- and osmotic-stress, and suggested to play a valuable role in desiccation tolerance in a variety of organisms, are the Late Embryogenesis Abundant (LEA) proteins (Kriško *et al*., 2011; Tolleter *et al*., 2010). LEA proteins were first described 30 years ago, associated with the late stages of cotton seed development (Hundertmark *et al*., 2010) and are classified as Intrinsically Disordered Proteins (IDP's).

Intrinsically disordered proteins (IDP's) defy the classical paradigm of structure- function in proteins, as the functional state has no well-defined 3D structure. However, many IDP's are believed to alter secondary structure from a disordered state to a more structured one, upon binding to target molecules (Hara, 2010). For example; in some cases, particularly for group 3 and 6 LEA proteins, secondary structure prediction programs do calculate a high degree of folding, primarily into α-helices. Group 1 and 2 LEA proteins are predicted to contain a

significantly higher proportion of unstructured loop regions compared to α- helices or β-sheets (Wise and Tunnacliffe, 2004).

Disordered proteins are rich in polar and charged amino acid residues. They also frequently possess (30-40 amino acid) regions of low complexity (LC), regions with low Shannon entropy, general structural disorder and do not form globular structures (Livernois *et al*., 2009; Kriško *et al*., 2010). The advantages associated with structural disorder include alternate conformations which rapidly interconvert allowing for multi-functionality and high specificity with weak and reversible interaction (Tompa and Kovacs, 2010). IDP's have been implicated in a number of major biological roles including transcriptional regulation, fatty acid synthesis, signal transduction and some molecular functions include roles as chaperones, oxygen scavengers and effectors (Tompa and Kovacs, 2010; Livernois *et al*., 2009; Hara, 2010). These proteins have the unique ability to bind numerous target molecules such as DNA, RNA and globular proteins (Hundertmark *et al*., 2010). The proteins of sporulating bacteria, host associated microbes and halophillic bacteria and archaea, generally show an abundance of unstructured, hydrophilic, low complexity regions (Kriško *et al*., 2010). *Deinnococcus radiodurans* is the most radiation and desiccation tolerant microorganism currently known and is considered the model organism for the study of resistance mechanisms to these stresses. This microorganism contains a large proportion of proteins containing long, hydrophilic, low complexity regions and these LC proteins, whose specific function is known, are thought to be involved in desiccation tolerance (Kriško *et al*., 2010).

5.1.2.2.1 Sequence and classification of LEA proteins

LEA proteins are primarily found in seeds, pollens and anhydrobiotic plants, but are not exclusive to them (Wise, 2002; Singh *et al*., 2005). Homologous genes have been found in nematodes (*C. elegans*, *Aphelenchus avenae*) and bacteria (*E.coli, B. subtilis, D. radiodurans* and *H. influenzae*). Although many of these proteins have no known function, those that have

been characterised are generally associated with cellular recovery processes (Kriško *et al*., 2010). LEA proteins are characterised by unusual amino acid compositions. They are generally rich in glycine (> 6 %) and small and/ or charged amino acid residues, and exhibit a high overall hydrophobicity index (Garay-Arroyo *et al*., 2000; Wise and Tunnacliffe, 2004). Flexibility is promoted by these residues and, with the exception of atypical hydrophobic representatives, LEA proteins exist as random coils in solution (Olvera-Carillo *et al*., 2010). These proteins tend to have a high content of amino acid which interacts with water, allowing for scavenging of polar molecules, which in turn provides hydration potential in periods of desiccation.

LEA proteins generally contain three signature motifs (Livernois *et al*., 2009; Tunnacliffe and Wise, 2007);

1. The lysine rich K segment is approximately 15 amino acids in length and forms amphipathic α-helices. This segment is thought to bind proteins and membranes during stress.
2. The S-segment is serine rich and may possibly be a site for phosphorylation.
3. The Y-segment shares homology to plant and bacterial nucleic acid binding sites of chaperones.

The organisation of LEA proteins into groups is generally based on sequence similarity to the prototypical *Gossypium hirsutum* LEA (Wise, 2003). Two main criteria are used for grouping LEA proteins. The first is based on conserved motifs and amino acid similarities, while the second is based on the protein or oligonucleotide probability profile (POPP) which clusters proteins based on over- or under- represented amino acid residues in the sequences (Shih *et al*., 2008). In all LEA protein sequences, phenylalanine, tryptophan, cysteine, isoleucine, leucine and asparagine are under-represented according to POPP classification (Wise, 2003).

A number of different LEA clusters have been reported (Table 5.1.3.2), but a consensus only exists for three; group 1 (D19), group 2 (D11) and group 3 (D7). The system used by Bray *et al* (1993) also distinguishes three additional groups; group 4 (D113), group 5 (D29) and group 6 (D34). There are two groups that do not feature in the Bray scheme but are represented by Pfam families; Lea5 (D73, PF03242 LEA_3) and Lea14 (D95, PF03168 LEA_2).

Group 1

LEA proteins in group 1 are highly hydrophilic and contain proteins from superfamily 4 and 6. Those clustered in superfamily 4 contain an increased percentage of charged amino acids. POPP classification of Group 1 proteins shows an overrepresentation of arginine, glutamate and glycine residues (Wise, 2003).

Group 2

Group 2 LEA proteins are split into two subgroups; subgroup 2a members are not upregulated by cold and contain decreased helix content (Battaglia *et al*., 2008). Members of subgroup 2b contain an increased proportion of loop regions and are induced by low temperatures. Both subgroups are highly hydrophilic; containing a polyserine stutter and an overrepresentation of histidine residues (Wise, 2003; Battaglia *et al*., 2008). Group 2 proteins also contain both the Y-segment (DEYGNP) and the EEKK motif.

Group 3

LEA proteins in Group 3 are characterised by an overall under-representation of glycine residues and over-representation of lysine and glutamate residues, as well as containing an 11-mer repeat. These proteins have a high helical content and, like subgroup 2b, are upregulated by cold stress and also contain a polyserine stutter (Wise, 2003; Battaglia *et al*., 2008).

Table 5.1.3.2: Classifications of LEA proteins into groups using different schemes.

Wise, 2003	Super-family Cluster	Bies-Ethéve *et al.*, 2008	Battaglia *et al.*, 2008	Dure (1993)	Bray, 1994	Pfam (Number)	Keywords *
Group 1	4	1	1	D19	I	LEA_5 (PF00477)	Histone H4, nuclear binding protein
	6						RNA binding, gyrase
Group 2	1	2	2	D11	II	Dehydrin (PF00257)	ATP binding, repair, topoisomerase
	3						Coiled-coil, histone H1, chaperone
	8						Topoisomerase, coiled-coil
	9						Transcription inhibition, glycosyl hydrolase
	10						nuclear binding protein, DNA binding, chaperone
Group 3	2	3	3A	D7	III	LEA_4 (PF02987)	Chaperone, filament, phosphorylation
	5						Coiled-coil, Histone H1, Hsp70
Group 4 ?	1	4	4B	D113	IV	LEA_1 (PF03760)	ATP binding, repair, topoisomerase
	2						Chaperone, filament, phosphorylation
	9						Transcription inhibition, glycosyl hydrolase
Group 5 ?	2	3	3B	D29	V	LEA_4 (PF02987)	Chaperone, filament, phosphorylation
Group 6	7	5	5A	D34	VI	SMP (PF04927)	GroEL, histone H1, DNA binding

Lea5	299	6	5B	D73	-	LEA_3 (PF03242)	DNA binding, transcription regulation
Lea14	297	7	5C	D95	-	LEA_2 (PF03168)	Esterase, glycoprotein, chaperone
-	-	4	4A	-	-	LEA_1 (PF03760)	
-	-	8	6	-	-	LEA_6 (PF10714)	

∗Keywords and phrases associated with POPP classification system

ʔ Group 4 proteins are redistributed into groups 2 and 3, while group 5 proteins are placed within group 3 according to Wise (2003)

- Groups not identified

5.1.2.2.2 Atypical LEA proteins and the Water HYpersensitivity domain

Lea14, Lea5 and group 6 LEA proteins all have average hydrophobicity scores and a lower percentage of polar residues, and are therefore considered atypical. There are some differences between these three groups but insufficient representative examples exist for adequate classification rules to be established (Wise, 2003). The atypical LEA proteins are unlike any other LEA proteins, not only in sequence but also in their biochemical properties; in particular, they are more heat sensitive and appear to have a more defined tertiary structure than LEA proteins belonging to other groups (Shih *et al*., 2008).

According to the POPP classification, Lea14 proteins contain an over-representation of residues D, K, I and IP, whereas R, Q and F are under-represented. In another scheme, LEA proteins from group 6 and Lea14 are clustered together into group 5, while Lea5 is not represented in any group. Alternatively, group 4 proteins of the Bray system are re-classified as group 5 and Lea14 is then clustered in group 4 (Wise, 2003). Although very little is known about the atypical LEA proteins, their transcripts have been shown to accumulate during a variety of stress conditions (Shih *et al*., 2008).

The most familiar typical LEA protein, and also the first LEA protein to be crystallised, is At1g01470, a Lea14 protein [PF03168 (LEA_2)] from *Arabidopsis thaliana*. Interestingly, expression ratios for At1g01470 under drought, cold and high salinity stress were five-fold higher than under non-stress conditions (Shih *et al*., 2008). In addition, Dunaeva and Adamska (2001) demonstrated that At1g01470 mRNA levels in the leaves of *A. thaliana*, increased significantly in response to light stress. These separate research findings have served to connect desiccation, high salt, high light, plant wounding and low temperature stress adaptation to an increased expression of At1g01470.

In 2005, Ciccarelli and Bork reported a novel domain known as the Water HYpersensitity domain (WHy), which provides a link to Hin1 genes (induced in plants in response to bacterial infection and part of the general stress response pathway) to the plant Lea14 proteins (expressed during a number of stress conditions) and a number of uncharacterised bacterial and archaeal proteins (Ciccarelli and Bork, 2005). The presence of this domain in plant proteins expressed in response to external stresses and prokaryotic proteins with unknown function, may suggest a similar molecular pathway for abiotic stress responses, most likely acquired by horizontal transfer (Ciccarelli and Bork, 2005). This domain is approximately 100 amino acids in length with an NPN motif at the N-terminus and alternating hydrophilic and hydrophobic residues. The secondary structure is predicted to contain mostly β-strands with a single C-terminal α-helix (Ciccarelli and Bork, 2005).

Interestingly, the WHy domain is absent in fungi, insects and mammals, but widespread in plants. In addition, the Hin and Lea14 genes, which contain this domain, share very low sequence similarity with other members of the LEA family (Maitra and Cushman, 1994).

5.1.2.2.3 Functions of the LEA proteins

Several functions have been proposed for the dehydrin proteins, including anti-freeze activity, space-filling molecular shielding and chaperone activity (Livernois *et al*., 2009). LEA proteins may act as hydration buffers as they bind large amounts of water, effectively sequester ions, and offer direct protection to proteins and membranes (Wise and Tunnacliffe, 2004). Although the mechanism of action of LEA proteins is still unknown, studies have shown diverse functions. It cannot be disputed that LEA proteins participate in adaptive responses for a variety of stresses. In *E.coli*, for example, 5 genes encode hydrophilin homologs and the deletion of the RMF gene results in an osmo-sensitive phenotype (Garay-Arroyo *et al*., 2000).

Functional screening of a cDNA library of C.W80 in cyanobacterial cells (*Synechococcus* spp.) allowed for the successful isolation of a 108 amino acid polypeptide with sequence homology to a group 3 LEA anti-stress protein. By introducing anti-stress genes of the halotolerant marine algae, *Chlamydomonas* spp., into higher plants, salt and oxidative stress tolerance was enhanced (Tanaka *et al*., 2004). The soybean LEA proteins from group 1 and group 2 confer salt and cold stress tolerance in bacterial systems, whereas the *Arabidopsis* group 2 and 4 LEA proteins negatively affected *E. coli* under normal growth conditions. The hot pepper group 5 LEA protein has been shown to enhance dehydration and salt tolerance in transgenic tobacco. In addition, antioxidant properties have been described for a group 5 LEA protein from *Arabidopsis* (Mowla *et al*., 2006).

One functional property observed for a number of LEA proteins is their propensity for binding to phospholipid membranes, maintaining topology and the phase transition temperature during stress (Mouillon *et al*., 2008). The membrane is a physical barrier which separates cells from their external environment and is the first structure to be negatively affected by stress (Mahajan and Tuteja, 2005). Cellular responses are initiated by the interaction of membrane-bound proteins with external stimuli. Clearly, for any organism to elicit the appropriate response to any given condition, the membranes and imbedded molecular subunits must be functioning correctly.

The extraction of LEA proteins from membrane fractions clearly demonstrates binding events between the dehydrin and lipid containing membranes. The first direct evidence of this was provided by Koag *et al*., (2003) when maize dehydrin DHN1 was observed bound to vesicles containing anionic phospholipids. Similarly, *Arabidopsis* dehydrins ERD10 and ERD14 also bind acidic phospholipid vesicles, suggesting an interaction between specific membrane regions and LEA proteins (Hara, 2010; Kovacs *et al*, 2008).

Further evidence for membrane protection activity of LEA proteins was provided in a study by Tolleter *et al* (2010). The mitochondrial LEA protein of *Pisum sativum* was shown to interact with negatively charged phosphate groups in lipid bilayers in the dry state. The interaction is thought to increase the lipid spacing which results in increased fatty acid chain mobility and subsequent decreases in lipid melting temperatures (Tolleter *et al*., 2010). Liposomes used in this study exhibited increased stability after drying and re-hydration as well as freeze-thaw cycles. In addition, enhanced protective effects were observed in membranes containing the cardiolipin (CL) phospholipid, which is exclusively found in bacterial and mitochondrial membranes, thereby possibly indicating substrate specificity for these proteins, related to their sub-cellular location (Tolleter *et al*., 2010).

In another study, LEA18, an uncharacterised protein from the Pfam LEA_1 domain, was found to bind to negatively charged membranes, inducing partial folding of the protein (Hundertmark *et al*., 2010). Subsequent vesicle aggregation and cellular leakage occurred which indicated destabilisation of membranes and membrane proteins. Although it seems that this study is in disagreement with the classical cellular stabilisation activity observed with other LEA proteins (ERD10, ERD14 and AavLEA1), the authors state that liposomes used in the study were in a hydrated state, whereas expression of LEA18 occurs maximally during late seed maturation, under water deficit conditions (Hundertmark *et al*., 2010). It may be possible that the authors have established another function for LEA proteins. Under rehydration stress, anhydrobiotic cells which have accumulated high levels of compatible solutes must be able to efficiently remove them. By binding membranes under hydrated conditions and causing leakage of the soluble cellular content, these compounds can be removed from cells.

Dehydrins, typically acidic dehydrins such as ERD10, ERD14, COR47 and VcaB45, are also capable of binding small molecules, most notably calcium. This binding allows for phosphorylation events to occur in LEA proteins (Hara, 2010). A major phosphorylation site in dehydrins is the S-segment, which has a calcium binding site upstream. These sites may indicate a functional role of LEA proteins as sensory proteins, in cell signalling events which are produced or altered as a result of signal transduction from receptors to secondary messengers, and which may interact with partners and initiate phosphorylation cascades. These, in turn, target major stress response genes and/or transcription factors which control the genes (Mahajan and Tuteja, 2005). Furthermore, dehydrins may also bind a variety of divalent metal ions which could point to their involvement in metal buffering and/or metal sensing (Hara, 2010). For example, the citrus dehydrin, CuCOR15, exhibits an altered structure when bound to zinc ions which, in turn, promotes its binding to nucleic acids. The Y segment (DEYGNP), which occurs at the N-terminus of the protein, is believed to be a putative nucleotide binding domain. However, there are no reports of DNA binding of this segment in CuCOR15. Rather, zinc dependant nucleic acid binding in this protein is linked to H-rich or polyK containing segments (Hara, 2010).

It appears that LEA proteins may bind to molecules that are susceptible to the effects of stress and could alleviate the damage caused by specifically targeting these macromolecules (Hara, 2010). The K segment is believed to form amphipathic helix structures which allow for binding to various macromolecules (Hara, 2010). Cycles of freeze-thaw produce irreversible inactivation of a variety of enzymes, including alcohol dehydrogenase, malate dehydrogenase and lactate dehydrogenase. Providing protection to enzymes under these conditions may point to a cryoprotective function of LEA proteins. This was demonstrated in a study using the nematode LEA protein (AavLEA1), protective efficacy was greater than known cryoprotectants, BSA and sucrose (Goyal *et al*., 2005).

Dehydrins have been shown to function as chaperones in dehydration-rehydration and heat-induced aggregation assays but, unlike classical chaperones, are not dependent on ATP for function (Hara, 2010). Chaperones can be classified into two broad classes; those that bind to proteins and facilitate correct folding, or those that bind nucleic acids, which perform a similar protective function and prevent misfolding of RNA into complex secondary structures (Tompa and Kovacs, 2010). In a bioinformatic based survey of protein disorder in these two classes of chaperones, the authors observed a very high proportion of disordered regions in RNA chaperones (54.2 %) when compared to protein chaperones (36.7 %) (Tompa and Kovacs, 2010), indicating that these flexible regions allow for binding to the high number of variable secondary structures that can be formed by any number of RNA molecules.

Group 2 LEA proteins ERD10 and ERD14 from *Arabidopsis* confer desiccation tolerance to the plant and are overexpressed under conditions of low temperature, high salinity and increased light (Kovacs *et al*., 2008). ERD10 and ERD14 both protect various substrates from heat induced protein aggregation, which also exceeds the protection conferred by BSA under the same conditions. Plant, bacterial (*E.coli*) and yeast (*S. cereviseae*) dehydrins have been shown to protect enzymatic activity of malate dehydrogenase (MDH) and lactate dehydrogenase (LDH) from desiccation *in vitro* (Reyes *et al*., 2005). These hydrophillins were also tested for the ability to confer freeze-thaw protection to the enzymes (Reyes *et al*., 2008). The molar ratio of dehydrin to enzyme in many studies is 1:1 and protective activity is presumably not only due to the formation of a hydration shell, but also due to direct protein-protein interaction. Similarly, desiccation and freezing induced protein aggregation of citrate synthase and lactate dehydrogenase were prevented by the protective actions of the nematode group 3 LEA protein (AavLEA1) and the wheat group 1 LEA protein (Em). The protection conferred by AavLEA1 only occurred in a trehalose dependant manner which indicates that

this protein may function as a molecular shield, rather than a true chaperone (Tompa and Kovacs, 2010).

Aims and objectives

3. Perform bioinformatic analysis on 13ORF6 (dWHy1)
4. Sub-clone dWHy1 and utilise *in vivo* assays to assess possible desiccation tolerance
5. Express and purify dWHy1 for further *in vitro* assays to identify possible functions

5.2 Materials and methods

5.2.1 Bioinformatic analysis of dWHy1

The translated nucleotide sequence for 13ORF6 was used for BLASTp searches in the Uniprot and NCBI protein databases and subsequently designated the name dWHy1. SignalP was used to determine signal peptides and targeting signatures. Rare codon content was predicted using the Rare Codon Caltor. General protein characteristics such as amino acid content, isoelectric point (pI), molecular weight and protein Grand Average Hydropathy (GRAVY) were predicted using the ProtParam tool (Expasy). Regions of protein disorder were predicted by using the IUpred program and Kyte and Doolittle hydrophilicity plots were generated (http://iupred.enzim.hu/;www.vivo.colostate.edu/molkit). Phosphorylation sites were predicted using NetPhosBac 1.0 (Miller *et al*., 2008).

Multiple sequence alignments using ClustalW (Larkin *et al*., 2007) were used to determine conserved catalytic regions in the genes. InterproScan was used to find matches to the predicted protein based protein family domains (Finn *et al*., 2008). PSIPRED was used to predict the secondary structure of the protein (McGuffin *et al*., 2000). The translated nucleotide sequences were used for homology modeling using the Swiss-protein modeler program (Schwede *et al*., 2003). PDB-sum was used to assess the accuracy of the models by generating individual Ramachandran plots (Lovell *et al*., 2002). Models were superimposed into the defined templates using the PyMol program.

Additional amino acid sequences were downloaded from NCBI or UniProt databases and used to construct neighbour joining phylogenetic trees in Mega. The stability of the relationships was assessed by performing bootstrap analysis based on 1000 resamplings.

5.2.2 Sub-cloning of dWHy1

5.2.2.1 Sub-cloning into pET21a

The 500 bp nucleotide sequence was amplified from extracted fosmid DNA using the LEA-F21/LEA-R21 primer pair. Primers contained EcoRI and XhoI restriction sites respectively. PCR amplification was performed using DreamTaq® in a standard PCR cycle (Table 4.2.1⁂) with an annealing temperature of 59 °C. The purified products were ligated into the pGEM-T easy cloning system according to manufacturer's instructions. Resulting ligation mixtures were transformed into electrocompetent GeneHog *E. coli* cells and transformants were selected on LB agar supplemented with carbenicillin and X-Gal. Randomly selected white colonies were selected and screened for inserts by alkaline lysis plasmid preparation, restriction digestion with EcoRI and XhoI, followed by agarose gel electrophoresis. Several clones containing the 500 bp insert were sent for sequence analysis at the University of Stellenbosch sequencing facility.

Plasmids were extracted from clones containing sequence-verified inserts. DNA was digested with XhoI and, following gel electrophoresis, linearized DNA fragments were excised and purified using the Nucleospin kit. The resulting DNA was then digested with EcoRI, analysed by agarose gel electrophoresis and the 500 bp fragment was excised, purified and quantified.

The fragment was ligated into pET21a vector, previously digested with EcoRI and XhoI, and the resulting mixture was used to transform GeneHog *E. coli* cells. Random clones selected after growth on LB-agar plates supplemented with carbenicillin were screened for recombinant plasmids by restriction digestion and agarose gel electrophoresis. Resulting positive clones were sent for sequence analysis at the University of Stellenbosch sequencing facility. Plasmids from sequence-verified clones were used to transform both Rosetta (DE3) pLysS and BL21 (DE3) *E. coli* strains.

5.2.2.2 Sub-cloning into pET17b

The 500 bp nucleotide sequence was amplified from extracted fosmid DNA using the primer pair W17R/W17F. Primers contained NheI and EcoRI restriction sites respectively. PCR amplification was performed using PrimeStar Taq polymerase in a standard PCR cycle (Table 4.2.1⁂) with an annealing temperature of 61 °C. The purified products were ligated into the pET17b plasmid vector, previously digested with NheI and EcoRI. Resulting ligation mixtures were used to transform electrocompetent GeneHog *E. coli* cells and transformants were selected on LB agar supplemented with carbenicillin. Randomly selected colonies were screened for insert by alkaline lysis plasmid preparation, restriction digestion with NheI and EcoRI, followed by agarose gel electrophoresis. Several clones containing the 500 bp insert were sent for sequence analysis at the University of Stellenbosch sequencing facility. Plasmids from these sequence-verified clones were used to transform the Rosetta (DE3) pLysS *E. coli* strain.

5.2.3 *In vivo* assays for desiccation tolerance

Mannitol agar was prepared from LB agar with the addition of 22 % [w/v] D-mannitol. Salt agar consisted of 1 % [w/v] tryptone, 0.5 % [w/v] yeast extract, 1.3 % [w/v] agar and 3.5 % [w/v] NaCl. To test for an osmotolerant phenotype, cultures were grown in liquid salt-reduced LB broth (1 % [w/v] tryptone, 0.5 % [w/v] yeast extract and 0.5% [w/v] NaCl) supplemented with the appropriate antibiotic. Cultures were grown at 37 °C until an OD_{600} of approximately 0.5 was obtained. Cultures were then induced with 0.4 – 0.8 mM IPTG and grown overnight at 30 °C. The OD_{600} of each culture was measured spectrophotometrically and corrected with salt-reduced LB broth to 0.6. Finally, serial dilutions up to 10^{-6} were made in quarter strength Ringer's solution. For the initial tests, 5 µl of each dilution was spot plated onto each stress agar plate and incubated at 37 °C overnight. For survival rate tests, the same procedure was followed except that 100 µl of each dilution was spread-plated onto each

stress plate and a corresponding LB agar plate. CFU's were recorded following an overnight incubation at 37 °C and the percentage survival rate was calculated as follows;

$$\% \text{ Survival rate} = \frac{\text{CFU/ml on stress plate}}{\text{CFU/ml on LB agar plate}} \text{ X } 100$$

All assays were performed twice, in triplicate, and control cultures containing parental vector in the expression host were routinely included. Statistical significance was calculated using ANOVA analysis.

5.2.4 Protein expression and purification

Recombinant clones containing either parental vector or vector containing dWHy1 were grown in LB broth supplemented with the appropriate antibiotic(s) at 37 °C until an OD_{600} of 0.4-0.5 was obtained. Cultures were then induced with either 0.4 or 0.8 mM IPTG and then grown for 24 or 48 hours, at 30 °C. Cultures were centrifuged at 6000 × *g* for 10 minutes and the culture supernatant was precipitated with 20 % TCA at 4 °C for 24 hours. Pellet fractions were resuspended in sonication buffer (30 mM Tris-HCl [pH 8.5], 300 mM NaCl, and 10 % [v/v] glycerol) and sonicated for 6 cycles of 30 seconds each. The soluble and insoluble fractions were separated by another round of centrifugation at 6000 × *g* for 10 minutes. Aliquots of all fractions were mixed with an equal volume of 2 × SDS loading buffer and analysed by SDS-PAGE. Soluble cell free extract containing the protein of interest was subjected to purification via metal affinity chromatography. Samples were dialysed against 3 L of buffer (50 mM Tris-HCl [pH8.5]) for 2 days at 4 °C. Fifty millilitres of dialysis buffer was stored at 4 °C and used as a control in assays. Recovered fractions were purified further by FPLC. Following SDS-PAGE, protein spots excised from 12 % [w/v] polyacrylamide gels were sent to the Proteomics department at the University of the Western Cape for MALDI-

TOF analysis. Resulting peptide mass abundance profiles were searched against the Mascot database.

5.2.5 *In vitro* freeze-thaw assays

All proteins were quantified according to the Bradford assay using BSA as standard. Malate dehydrogenase was purchased from Sigma and a stock solution of 1 mg/ml was prepared in 25 mM Tris-HCl buffer (pH 7.5). The final concentration of enzyme used in each assay was 250 nM. Purified dWHy1 was diluted to the same concentration and the appropriate volume was added to the MDH preparations in a molar ratio (MDH: dWHy1) of 1:5. Subsequently, 100 µl aliquots of MDH or MDH-dWHy1 mixtures in thin-walled 0.5 ml eppendorf tubes were frozen at -80 °C for 20 minutes and then transferred to a 25 °C water bath for 20 minutes. This constituted one freeze-thaw cycle and was repeated up to five times with activity measurements taken after 1 cycle, 3 cycles and 5 cycles. MDH enzymatic activities were determined using 8 µl aliquots in a final volume of 600 µl of reaction buffer (150 mM potassium phosphate buffer [pH 7.5], 0.2 mM oxaloacetate, 0.2 mM NADH). The volume of sample added to the reaction buffer was adjusted accordingly to compensate for dilution of MDH when dWHy1 was added. MDH activity was determined by a decrease in absorbance at 340 nm for 1 minute at 25 °C, due to conversion of NADH to NAD^+. Untreated samples were kept on ice and the activity was measured before each treatment was determined. The rate obtained for the untreated samples was taken as 100 %. Enzymatic assays were repeated in triplicate with activity buffer as the blank measurement. Controls used included 500 nM BSA, 25 µM trehalose and 0.1 % [v/v] glycerol.

5.2.6 Construction of vector pRareMod7

The plasmid pRareLysS was extracted from Rosetta (DE3) pLysS cells by alkaline lysis. DNA was digested with either EcoRI or NcoI and analysed by agarose gel electrophoresis. Linearized plasmid DNA obtained by NcoI restriction digestion was excised from the agarose gel and purified using the Nucleospin kit. The plasmid pET28a was subjected to restriction enzyme digestion with both AlwNI and DraIII. The resulting 1.5 kb band, which corresponds to the kanamycin aminoglycoside phosphotransferase (KAT) gene, was excised from the gel, purified and quantified. Purified DNA fragments were blunt-end repaired using Klenow DNA polymerase and the pRareLysS fragment was dephosphorylated using FastAP ™, according to the manufacturer's instructions. Following this, the 1.5 kb fragment was ligated to the 7.4 kb pRareLysS and the mixture was used to transform GeneHog *E.coli* cells. Transformants were selected on LB agar supplemented with kanamycin and assessed for chloramphenicol sensitivity. Plasmids were extracted from kanamycin resistant-chloramphenicol sensitive clones and subjected to restriction digest analysis, followed by agarose gel electrophoresis in order to verify any size shifts and patterns for known sequence regions. Once a positive recombinant plasmid was obtained, the DNA was transferred into BL21 (DE3) *E. coli* cells and transformants were selected by overnight incubation at 37 °C on LB agar supplemented with kanamycin.

5.2.7 *In vitro* transcription and translation

Cell-free protein expression was performed using the PURExpress® *in vitro* transcription and translation kit (New England Biolabs) according to the manufacturer's instructions. Briefly, extracted pET17b-dWHy1 DNA was purified by phenol: chloroform extraction and quantified by fluorimetry. Ten microliters of solution A was added to a sterile 0.5 ml eppendorf tube followed by 7.5 µl of solution B. DNA template was added to the reaction, and incubated at 37 °C for 4 hours. Three microliters of each sample was analysed by SDS-

PAGE. Samples were centrifuged using the Amicon®Ultra- 0.5 filter device (Millipore) and reverse His-tag purified using charged nickel ion resin (Invitrogen). All fractions were analysed by SDS-PAGE.

5.3 Results and discussion

5.3.1 Bioinformatic analysis of dWHy1

The 498 bp gene encoded by 13ORF6 on clone LD13 translated to a 165 amino acid protein. The translated protein is predicted to have a molecular mass of 18.6 kDa and a theoretical pI of 8.99. Interestingly, low molecular weight proteins of less than 25 kDa are generally involved in major biological and biochemical processes such as ribosome functioning, transcriptional regulation and stress responses and/or adaptation (Müller *et al.*, 2010). BLASTp analysis of the amino acid sequence revealed that 13ORF6 showed homology to a Water Hypersensitivity protein from *Pseudomonas mendocina* (51 %, E-value 3×10^{-44}) or a putative uncharacterised protein from *Azotobacter vinelandii* (55 %, E-value 6×10^{-40}). It seems plausible to find a desiccation tolerance protein in *Pseudomonas* spp. as these non-spore forming bacteria have been shown to represent up to 7 % of the bacterial biomass of viable microorganisms isolated from ice cores of sediments at Vostok station (Potts, 1994). There is a clear indication, therefore, that these microbes have the potential to exhibit unique molecular mechanisms for survival in extreme conditions.

InterproScan analysis showed that 13ORF6 exhibits classical signatures of the Water Hypersensitivity domain and the atypical LEA- 14 protein family [PF03168]; Figure 5.3.1.1). While the sequence of 13ORF6 did not indicate the presence of a signal peptide at the N-terminus, the LipoP 1.0 server (Junker *et al*., 2003) showed a possible lipoprotein signal peptide (signal peptidase II) which would cleave between amino acid residues 25 and 26 (Figure 5.3.1.2). Bacterial lipoproteins are involved in a number of cellular functions such as cell wall biogenesis and maintenance, as well as substrate transport. In addition, functionally

important lipoproteins in *E. coli*, such as LolB and LptE, are periplasmic chaperones (Okuda and Tokuda, 2011). Membrane specificity of lipoproteins has been investigated and the amino acid residues located at +2 and +3 at the N-terminus of the protein determine whether the protein is retained in the outer- or inner membrane (Okuda and Tokuda, 2011). Accordingly, 13ORF6 would be translocated across the inner membrane and would be retained in the outer-membrane.

A total of 5 % rare codons were detected in the 13ORF6 sequence, as well as the presence of one cluster of repeated rare codons (Table 5.3.1.1). This prompted the use of the Rosetta (DE3) pLysS expression strain, which is designed to supply rare codons.

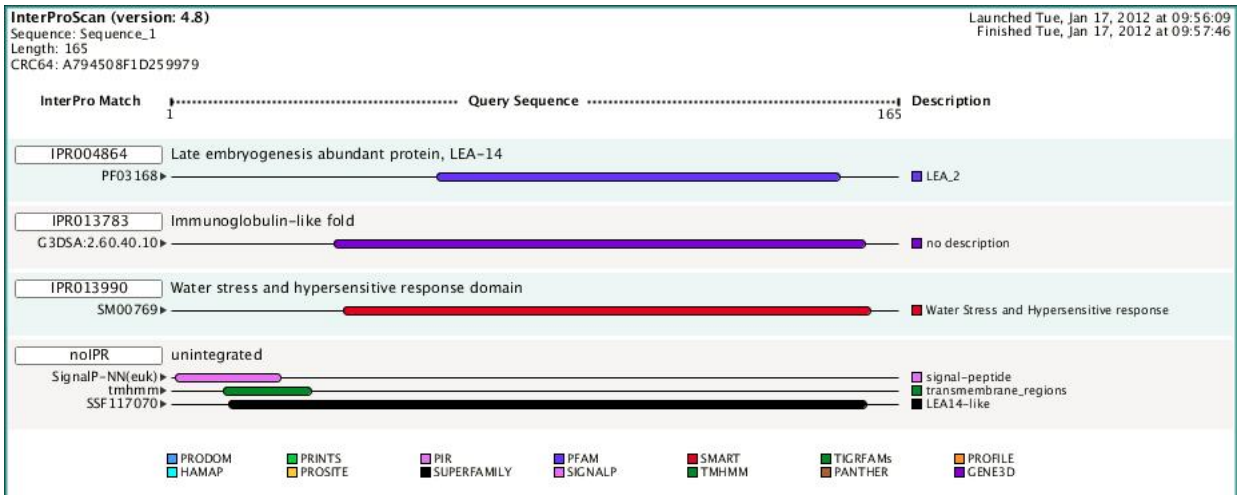


Figure 5.3.1.1: InterproScan results for 13ORF6 indicating the protein signature matches.

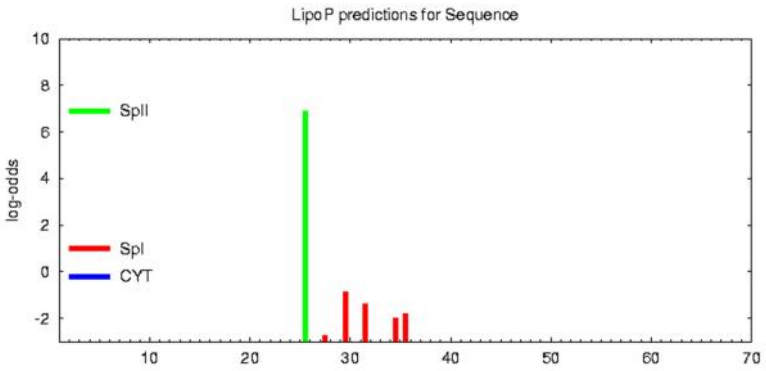


Figure 5.3.1.2: LipoP 1.0 server results for 13ORF6 indicating the predicted signal peptidase II cleavage site. Signal peptidase I cleavage sites (SpI) were predicted, but with very low scores, indicating that these sites are less likely to be cleaved.

Table 5.3.1.1: Rare codons occurring in the sequence of 13ORF6 and their frequency of occurrence.

Amino acid	Rare codon	Frequency
Arginine	CGA	0
	CGG	1
	AGG	0
	AGA	1
Glycine	GGA	0
	GGG	0
Isoleucine	AUA	3
Leucine	CUA	2
Proline	CCC	2
Threonine	ACG	0
Repeated and/ or consecutive rare codons		AUA-CCC

Using the NetPhosBac 1.0 server (Miller *et al*., 2008), putative phosphorylation sites were predicted, with high probability scores, in the 13ORF6 sequence (Figure 5.3.1.3). Phosphorylation is a reversible, post-translational modification important for regulating cellular function with phosphoproteins being involved in various aspects of cellular metabolism. Bacterial pathogens extensively use phosphorylation cascades to overcome host-defences. Phosphoproteins include enzymes required for protein and DNA metabolism as

well as stress response and, in *B. subtilis*, many of the proteins with phosphorylation potential have unknown functions (Soufi *et al*., 2008). A large number of bacterial proteins are phosphorylated on serine/ threonine/ tyrosine residues (Macek *et al*., 2007).

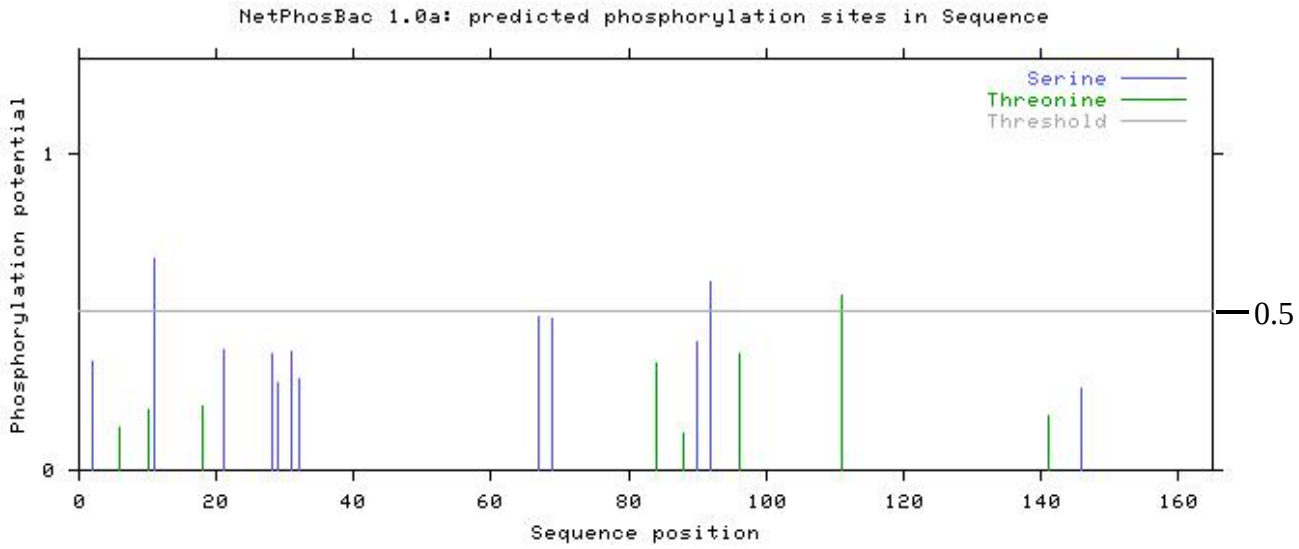


Figure 5.3.1.3: Prediction of phosphorylation sites in the protein sequence of 13ORF6. Residues with scores above 0.5 indicate a predicted phosphorylation site.

The GRAVY value for a peptide or protein is calculated as the sum of hydropathy values of the entire amino acid residues, divided by the number of residues in the sequence (Kyte and Doolittle, 1982). A positive value indicates that the protein has a hydrophobic nature, which has also been correlated to membrane associated proteins (Kyte and Doolittle, 1982), while negative values indicate a favourable interaction with water. The GRAVY value for 13ORF6 was predicted to be -0.087, indicating that this protein is slightly hydrophilic and may interact with polar solvents. As shown in Figure 5.3.1.4, the majority of amino acid residues in 13ORF6 are hydrophilic, indicating that these residues most likely interact with, and bind to, water molecules. Furthermore, both GRAVY and Kyte-Doolittle plots are in disagreement

with the BLASTp analysis, which indicates homology of 13ORF6 to the LEA_2 family. Atypical LEA proteins represented by PF03168 are known to exhibit overall hydrophobicity, whereas 13ORF6 is clearly hydrophilic.

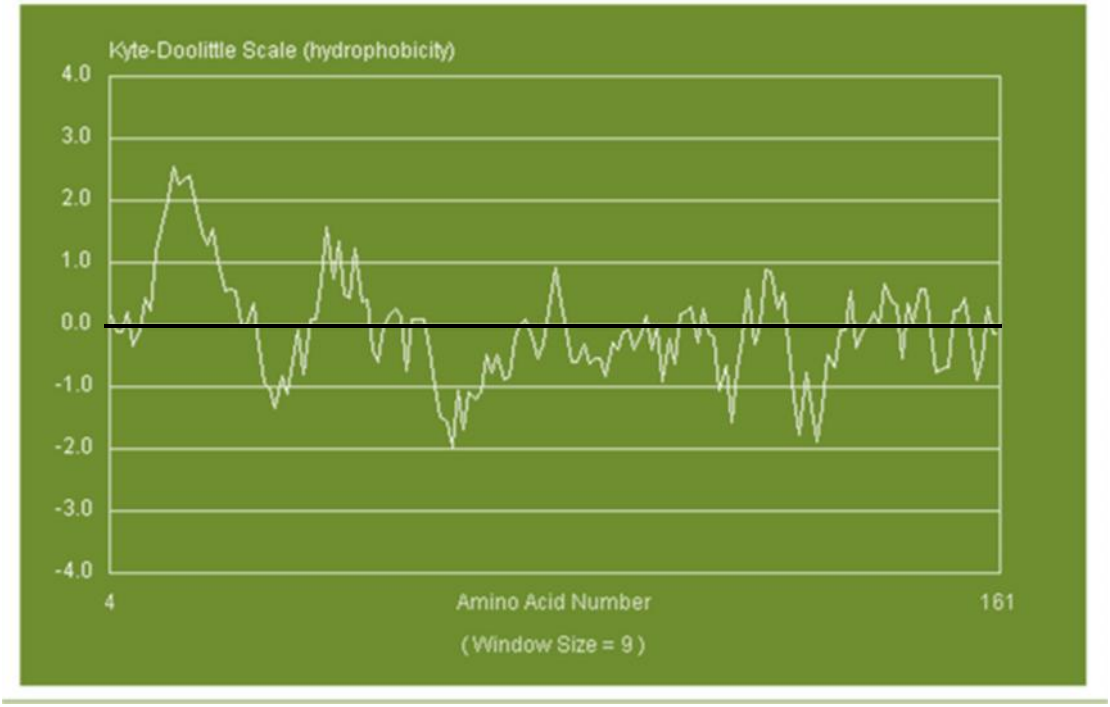


Figure 5.3.1.4: **The Kyte-Doolittle scale was used to assess the hydrophobic character of 13ORF6 with a window size of 9 for finding hydrophilic regions across the entire sequence. In this plot, regions with values above 0 are hydrophobic in character.**

IUpred was used to determine the overall short regions of disorder in dWHy1. Figure 5.3.1.5 indicates that only the extreme N- and C-terminal regions of 13ORF6 are disordered. Using these results, 13ORF6 does resemble atypical LEA proteins which also show lower levels of disorder and exhibit defined structures. In addition, the 32-160 residues of the protein could be modelled to the template 1xo8; the LEA14 protein from *Arabidopsis thaliana* (At1g01470) (20 % sequence identity). Model accuracy was assessed by generating Ramachandran plots which showed 85 % of residues in the favoured region and none in the

disallowed regions. The model of the protein in this study was therefore considered to be reasonably accurate (Figure 5.1.3.6 A and B). Another motivation for 13ORF6 classification as an atypical LEA protein is based on protein mobility in SDS-PAGE. Typical disordered proteins generally exhibit low electrophoretic mobility, most likely due to the hydrophilic nature of the proteins (Kovacs *et al.*, 2008; Hundertmark *et al*., 2010). For instance, the molecular mass of ERD10 is 29 kDa but migrates to an approximate molecular mass of 45 kDa (Kovacs *et al*., 2008). This was not the case for 13ORF6 and the protein migration was similar to the predicted molecular mass (Section 5.3.6). Clearly, 13ORF6 cannot be classified into a particular LEA group, based on the general characteristics used to define them. Based on the above analysis and multiple sequence alignments (Figure 5.3.1.7), 13ORF6 is rather classified as a bacterial Water Hypersensitivity protein and is designated as dWHy1 in this study. As increased numbers of 'unusual' sequences emerge, particularly from non-plant species, it may be beneficial to develop a different classification system for prokaryotic LEA-like proteins. However, due to the limited data available, any new system would require some link to current schemes in order to develop testable hypothesis to elucidate possible functional similarities.

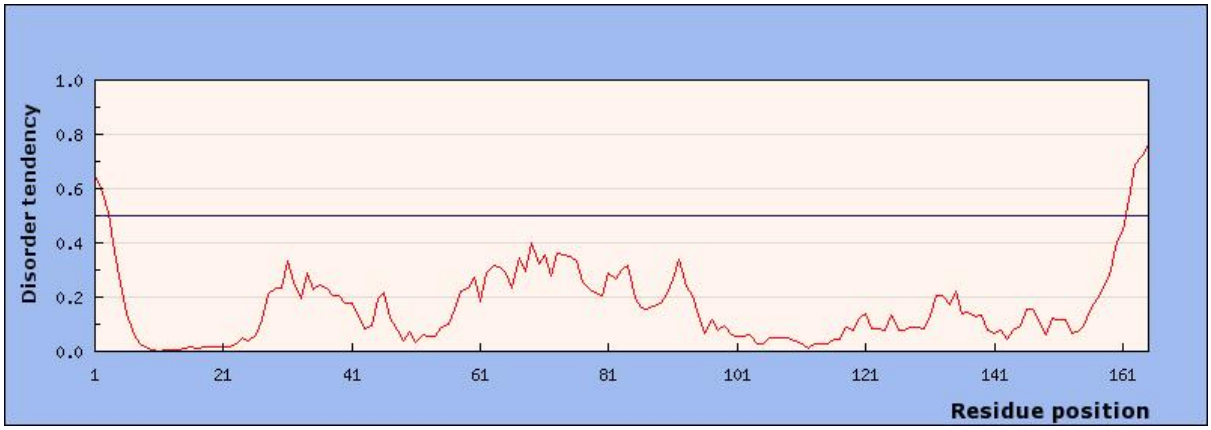
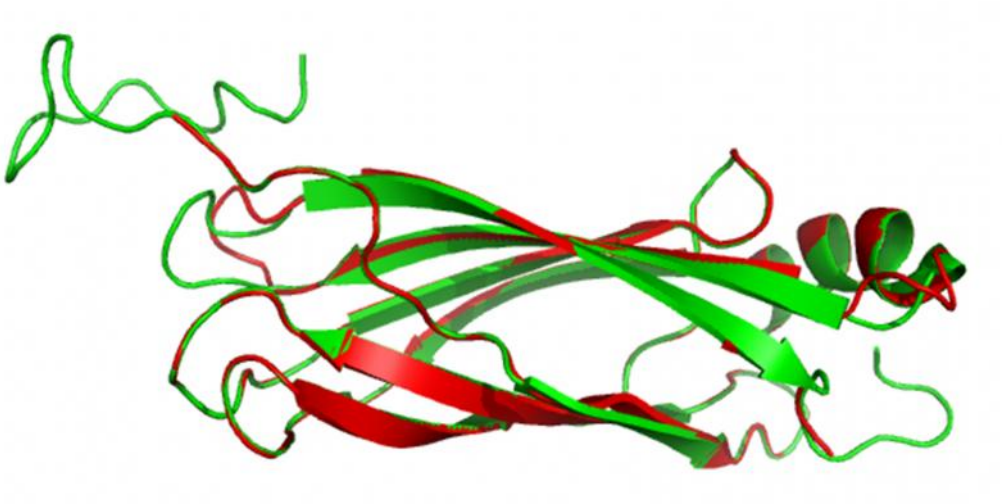
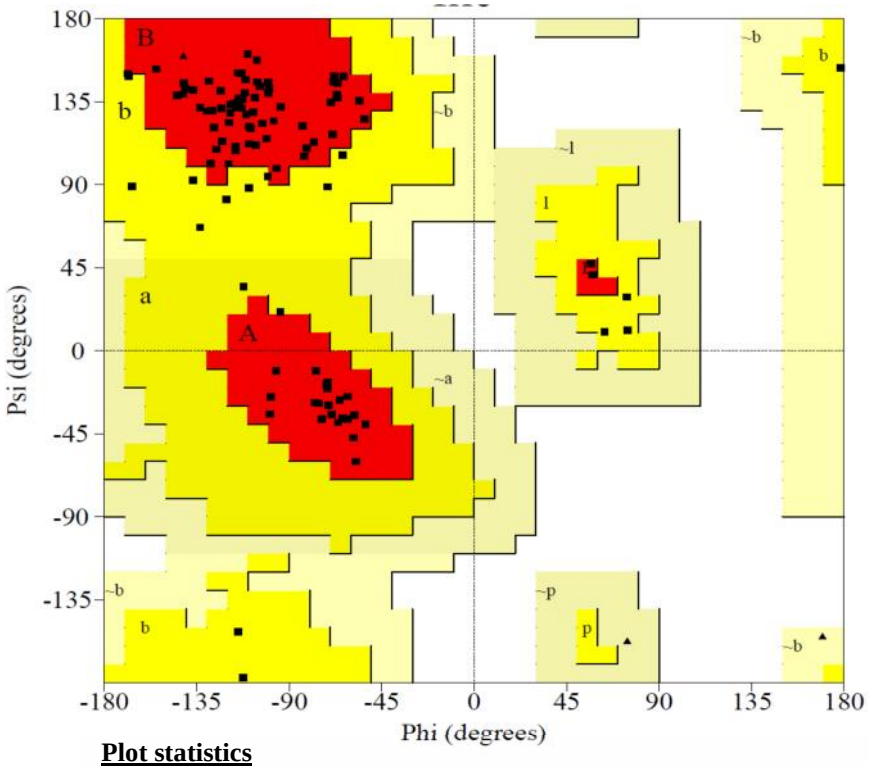


Figure 5.3.1.5: IUpred prediction of short regions of disorder. The line indicates the threshold at which residues (indicated on X-axis) are either disordered or ordered. Scores above 0.5 indicate disorder.

A



B



Plot statistics

Residues in most favoured region (A,B,L)	85 %
Residues in additional allowed regions (a, b, l, p)	15 %
Residues in disallowed regions	0.0 %

Figure 5.3.1.6: A] Constructed model of dWHy1 (Red) superimposed onto the template, 1xo8, the LEA14 protein from *Arabidopsis thaliana* (Green). B] Ramachandran plot analysis of the accuracy of the model.

As previously indicated, Ciccarelli and Bork (2005) reported a novel domain known as the Water HYpersensitity domain (WHy), and provided a link between Hin1 genes, plant Lea14 proteins and a number of uncharacterised bacterial and archaeal proteins, believed to have been acquired by horizontal transfer (Ciccarelli and Bork, 2005). This domain is approximately 100 amino acids in length with an NPN motif at the N-terminus and alternating hydrophilic and hydrophobic residues. The secondary structure is predicted to contain mostly β-strands with a single C-terminal α-helix (Ciccarelli and Bork, 2005), and its presence in the sequence of dWHy1 may suggest an abiotic stress response similar to pathways observed in plants (Figure 5.3.1.8).

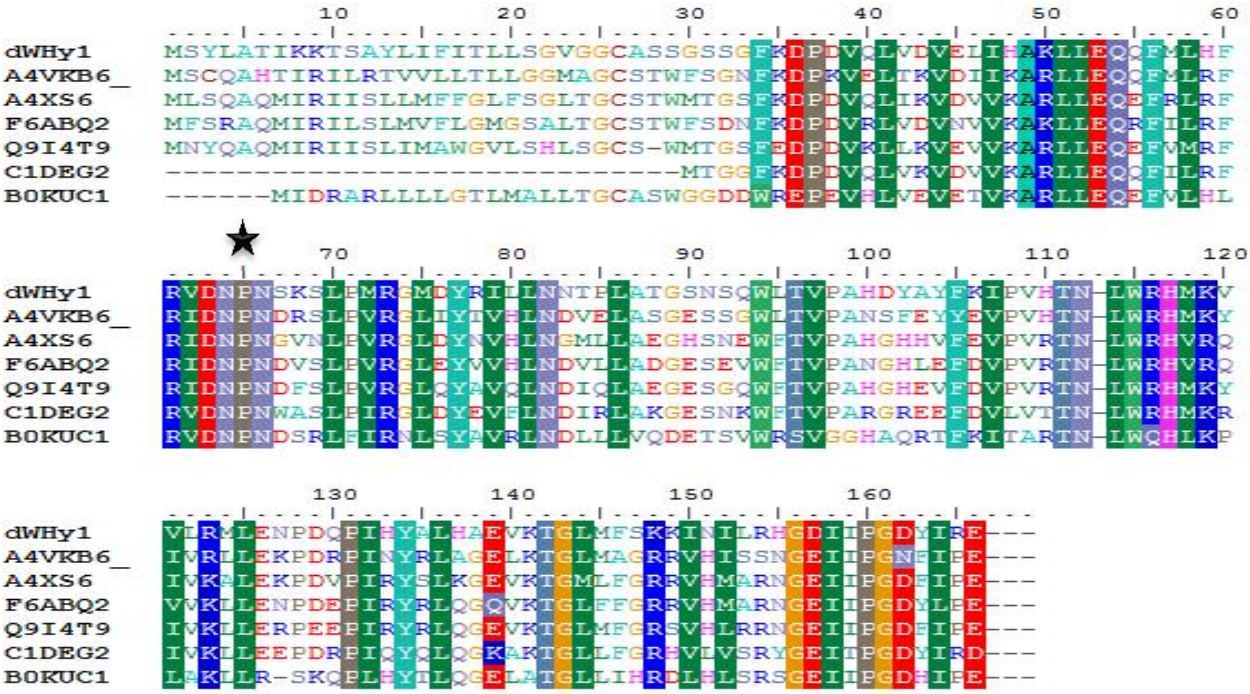


Figure 5.3.1.7: Multiple sequence alignment of dWHy1 with the top BLASTp hits from the UniProt database. Shading threshold of 90 % is used. Conserved WHy domain is indicated by ★. Accession numbers in the figure denote the following; A4VKB6: Putative lipoprotein [*P. stutzeri*], A4XS6: WHy protein [*P. mendocina*], F6ABQ2: WHy protein [*P. fulva*], Q9I4T9: Putative uncharacterised protein [P. aeruginosa], C1DEG2: Putative uncharacterised protein [*A. vinelandii*], B0KUC1: WHy protein [*P. putida*].

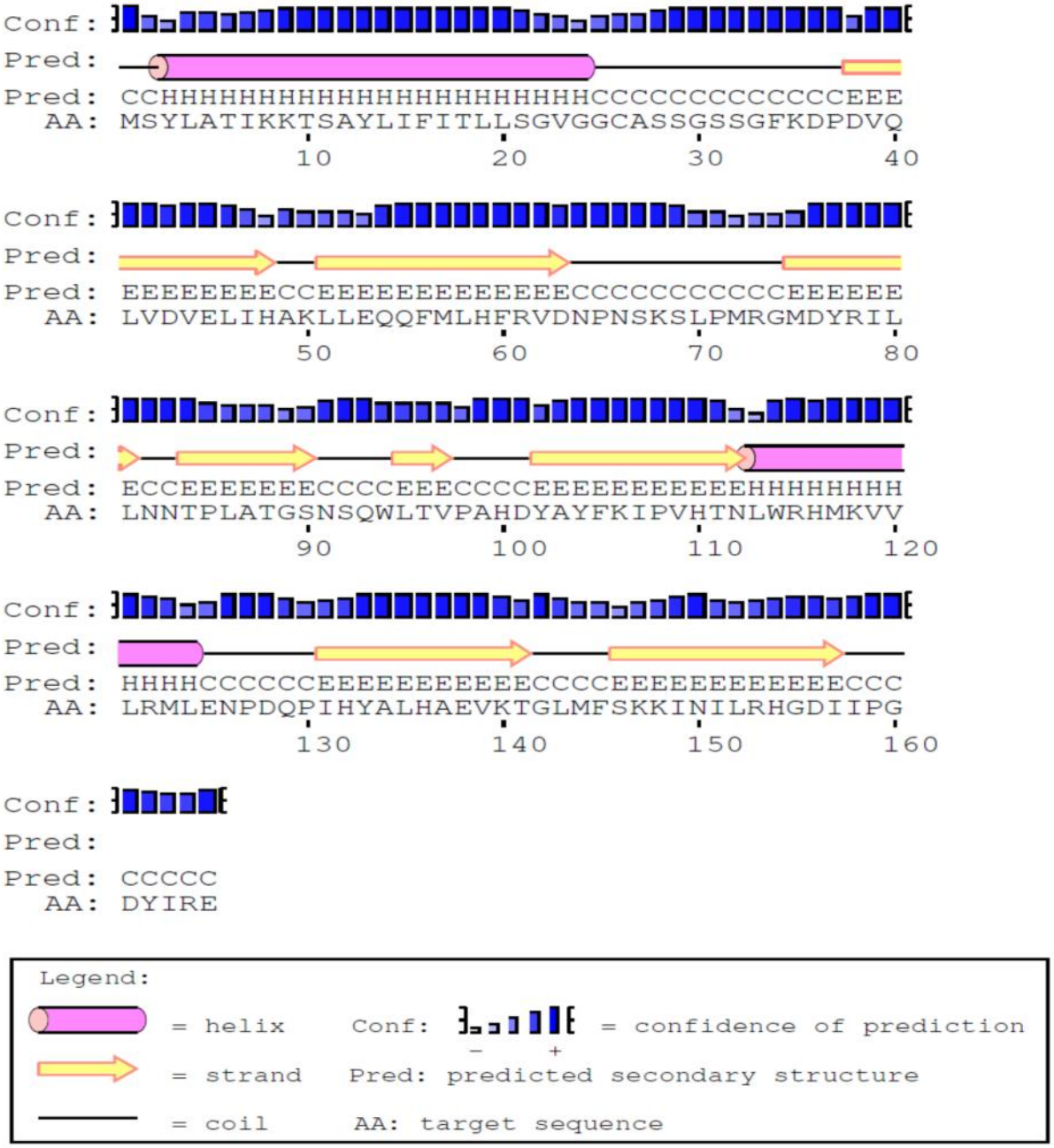


Figure 5.3.1.8: Secondary structure prediction of dWHy1. According to Ciccarelli and Bork (2005), the secondary structure of WHy domain proteins is composed mainly of β-strands and a C-terminal α-helix. Secondary structure predicted for dWHy1 shows predominant β- strands and C- and N-terminal α-helix structures.

One mechanism employed by bacterial pathogens to overcome host defences is the acquisition of genes responsible for defence mechanisms. This supports the hypothesis that proteins containing the WHy module were transferred from plants to bacteria (Ciccarelli and Bork, 2005). In addition, the Hin1 homologue is found in the ancient green algae, *Chlamydomonas reinhardtii*, and the WHy domain distribution seems to be restricted to plant pathogens or plant symbionts (Ciccarelli and Bork, 2005). If this is indeed the case, then the Antarctic organism from which dWHy1 originates, must have had some interaction with a plant host, and it is well known that the Dry Valleys have been devoid of plant life for a long period of time. In addition, water deficit and rehydration are among the predominant forces which influence the distribution and activity of microbial communities. Desiccation is therefore of great ecological significance and was most likely imposed on prokaryotes at an early evolutionary stage (Potts, 1994). It is therefore not unreasonable to assume that dWHy1 may be an ancient gene, yet it is difficult to explain why this gene is not more widespread, given the time for horizontal gene transfer events and the clear competitive advantage afforded to organisms that acquire it. In an effort to gain some insight into the phylogenetic relationship of dWHy1 to other WHy domain- containing sequences from plants, archaea and bacteria, a phylogenetic tree was constructed using the minimum evolution algorithm. As shown in Figure 5.3.1.9, sequences group into distinct clusters and can be grouped by taxa, i.e. plant, bacterial or archaeal. dWHy1 clearly clusters with the bacterial proteins, all of which contain the WHy domain motif, therefore indicating that dWHy1 is a bacterial water hypersensitivity response protein.

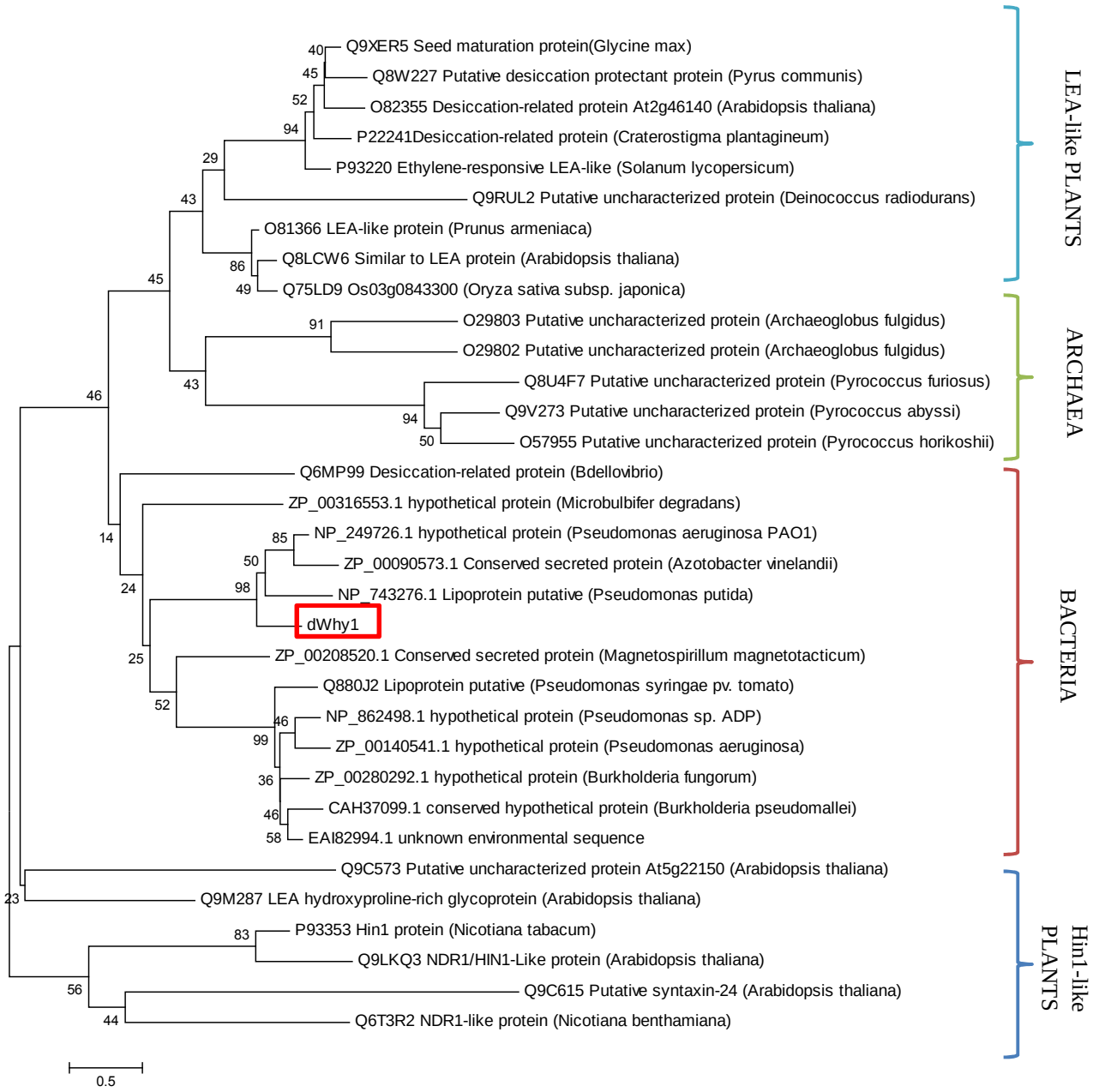


Figure 5.3.1.9: Phylogenetic analysis of dWHy1 to plant and bacterial sequences containing the WHy domain. Sequences included in the Ciccarelli and Bork (2005) analysis were obtained from the UniProt and NCBI databases and aligned using ClustalW. The Minimum Evolution tree was constructed in Mega4 (Kumar, *et al*., 2008), using the JTT substitution model with uniform rates among sites. The bootstrap test was set to 1000 replicates and the initial tree was calculated with the Neighbor-Joining algorithm (Jones *et al*., 1992).

5.3.2 Sub-cloning of dWHy1

PCR amplification of the gene encoding dWHy1 was successful and, although no classical signal peptides were detected on the N-terminus of the gene, bioinformatic analysis revealed the presence of a lipo-protein signal peptide. The pET21a vector system contains a C-terminal signal peptide and was therefore the most suitable choice for cloning dWHy1. Cloning the gene into pET17b for *in vitro* protein synthesis was also successful. Sequencing was routinely used due to the frequent observation of concatamers, which led to variable restriction profiles.

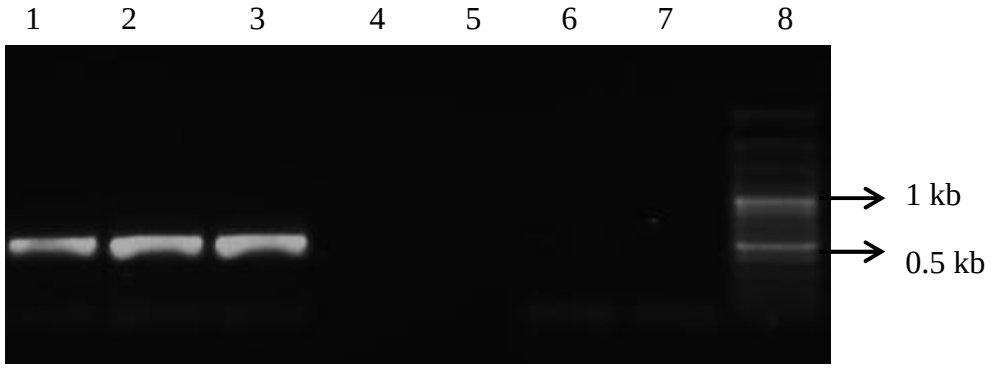


Figure 5.3.2.1: PCR amplification of dWHy from fosmid clone LD13 using DreamTaq (Fermentas). Lane 1 - 3) Amplicons of dWHy1. Lane 4) Forward primer control. Lane 5) Reverse primer control. Lane 6) Negative control. Lane 7) Control reaction using *E. coli* genomic DNA. Lane 8) GeneRuler 1 kb™ plus DNA molecular mass marker (Fermentas).

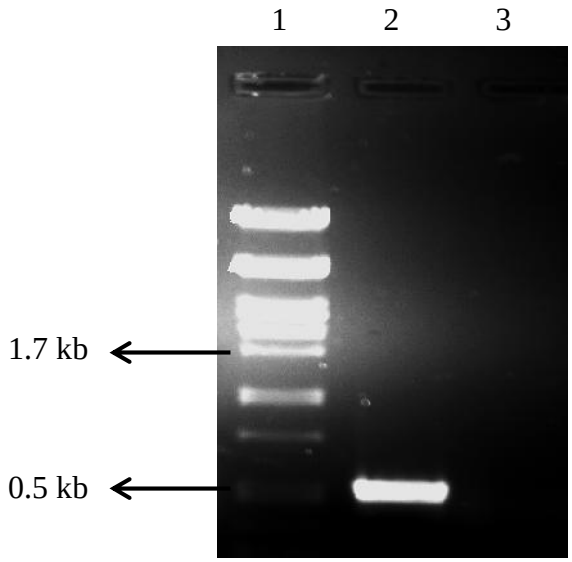


Figure 5.3.2.2: PCR amplification of dWHy from fosmid clone LD13 using PrimeStar polymerase (Takara). Lane 1) Phage lambda-*Pst*I DNA molecular mass marker. Lane 2) Amplicon of dWHy1. Lane 3) Negative control.

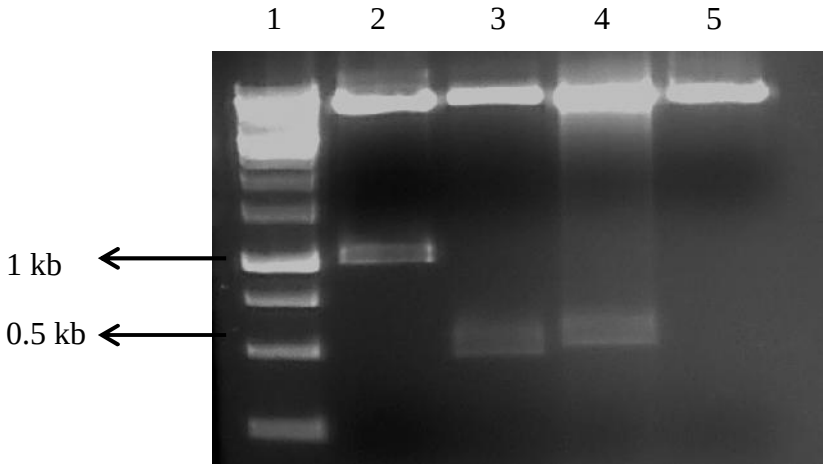


Figure 5.3.2.3: Restriction enzyme digestion of dWHy1-pET21a clones clearly indicating the formation of chimera sequences, with or without intact restriction sites. Lane 1) GeneRuler 1 kb™ DNA molecular mass marker (Fermentas). Lane 2) dWHy1-pET21a clone 15, chimera without second restriction site. Lane 3) dWHy1-pET21a clone 11 single gene sequence. Lane 4) dWHy1-pET21a clone 6, chimera with second restriction site. Lane 5) pET21a parental vector control.

5.3.3 *In vivo* phenotype assays

Following cloning of dWHy1, protein overexpression required verification before *in vivo* assays could be performed, as the expression of dWHy1 in a heterologous host was uncertain. Following small scale protein expression, a 22 kDa overexpressed band was observed by SDS-PAGE. Following purification by metal-ion and size-exclusion chromatography, the protein band was sent for MALDI-TOF analysis. Mascot database results indicated that this band was most likely Chloramphenicol acyltransferase (CAT), the selection marker on the pRareLysS plasmid in Rosetta (DE3) pLysS cells. This was a completely unexpected result as SDS-PAGE analysis revealed a highly overexpressed band which was smaller than 26 kDa, the molecular mass of CAT. In order to verify that dWHy1 was indeed being expressed in the host strain, one of the chimeric clones (dWHy1 C15), with two copies of the gene ligated together, was used for expression analysis. This allowed for visualisation of dWHy1 protein overexpression as the predicted molecular mass of the protein was approximately 43 kDa. SDS-PAGE did allow visualisation of two overexpressed proteins bands, one approximately 22 kDa in size and the other approximately 43 kDa in size corresponding to CAT and dWHy1 C11 respectively (Figure 5.3.3.1). In order to confirm observations of a desiccation tolerant phenotype, this clone, along with any other modified recombinant clones were tested in the same manner using stress plate assays (Section 5.5.4).

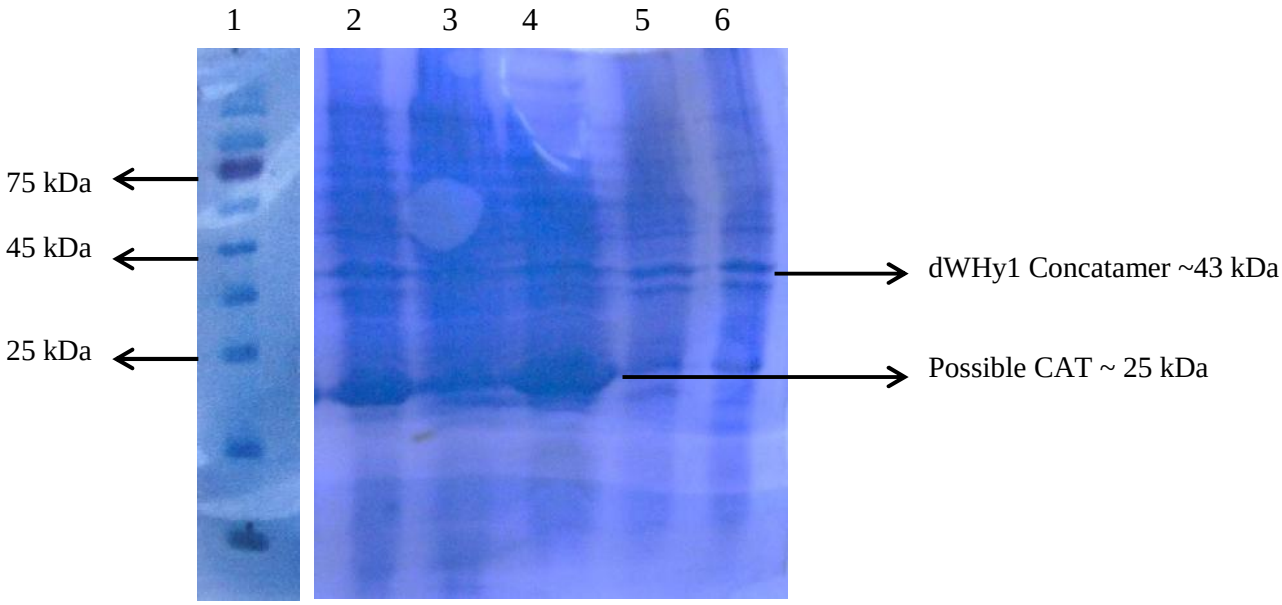


Figure 5.3.3.1: Protein overexpression analysis of dWHy1 C11 cultures. Lane 1) PageRuler™ Prestained protein marker (Fermentas). Lane 2) Soluble fraction of dWHy1 C11 induced with 0.4 mM IPTG. Lane 3) Soluble fraction of dWHy1 C11 uninduced. Lane 4) Soluble fraction of dWHy1 C11 induced with 0.8 mM IPTG. Lane 5) Insoluble fraction of dWHy1 C11 induced with 0.4 mM IPTG. Lane 6) Insoluble fraction of dWHy1 C11 induced with 0.8 mM IPTG.

Since homologs of dWHy1 are associated with desiccation stress tolerance, high concentrations of D-mannitol and NaCl were used to test for respective tolerance to osmotic and ionic stress. The concentration of NaCl was chosen based on reports that the salinity of sea water is ~3 % with 460 mM Na^+ (Mahajan and Tuteja, 2005). The common net effect of elevated levels of salinity and osmolarity is decreased water activity of cells, which results in disruption of ionic and osmotic equilibrium (Mahajan and Tuteja, 2005; Kriško *et al*., 2010). This tends to induce genes responsible for adaptation and/ or tolerance to adverse conditions. For example; Group 2 LEA proteins ERD10 and ERD14 from *Arabidopsis* are highly

overexpressed under conditions of low temperature, high salinity and increased light (Kovacs *et al*., 2008).

The percentage survival rates for both stress conditions were significantly higher for Rosetta (DE3) pLysS clones expressing dWHy1 than for clones containing parental vector, even when chimera recombinant plasmids were used. Colony growth on LB agar plates with no stress showed similar results for both dWHy1 containing cells and parental vector containing cells in all cases. Additionally, ANOVA analyses showed that stress survival rates were statistically significant (Table 5.3.3.1).

Table 5.3.3.1: Statistical significance determined by ANOVA analysis (single factor) for percentage survival rates of cells expressing dWHy1 on either D-Mannitol or Salt agar plates.

***E. coli* host**	**Stress condition**	**Concatamer (Y/N)**	**p-value**
Rosetta (DE3) pLysS	NaCl	Yes, two copies of the gene	< 0.005
Rosetta (DE3) pLysS	D-Mannitol	Yes, two copies of the gene	< 0.005
Rosetta (DE3) pLysS	NaCl	No	< 0.05
Rosetta (DE3) pLysS	D-Mannitol	No	< 0.05
BL21 (DE3)	NaCl	No	Not significant
BL21 (DE3)	D-Mannitol	No	Not significant
BL21 (DE3) pRareMod7	NaCl	No	< 0.005
BL21 (DE3) pRareMod7	D-Mannitol	No	< 0.005

These results are strongly suggestive that the *in vivo* function of dWHy1 is related to desiccation stress tolerance. However, exceptions to this were observed. When dWHy1-pET21 recombinant plasmids were transformed into the BL21 (DE3) *E. coli* expression host and GeneHog *E. coli* and tested for the *in vivo* phenotype, dWHy1- containing cells showed only marginally improved survival rates to the parental vector control. This could be explained for the GeneHog culture as this strain is a cloning host and may not express dWHy1 protein to required levels. If expression levels were indeed the case for this observation, then BL21 (DE3) cultures should have shown similar phenotype survival rates to the Rosetta (DE3) pLysS cultures. The rational explanation for this observation, which may apply to both GeneHog and BL21 hosts, is that the rare codons supplied by the Rosetta strain are vital for dWHy1 expression and subsequent function *in vivo*. While the requirement for rare codons in heterologous gene expression is well-documented, the general consensus for the use of an expression strain with such capabilities is determined bioinformatically, with a rare codon cut-off value of approximately 8 % total codon content. In addition, codon bias problems are increased when transcripts contain rare AGA and AGG codons and/or rare codon clusters (Sørensen and Mortensen, 2005, Kane, 1995). While dWHy1 only contains 5 % rare codons, none of which are the arginine codons AGA or AGG, it does contain one doublet cluster of Isoleucine (AUA) followed directly by Proline (CCC). It is not unprecedented that a single rare codon may negatively affect efficient translation of a heterologous gene (Kane *et al*., 1993). As seen in Figures 5.3.3.2- 5.3.3.5, the expression host was able to grow on the stress plates, albeit to a lesser extent. This is due to the inherent capabilities of *E. coli* to tolerate moderate desiccation conditions by mechanisms such as the synthesis of compatible solutes.

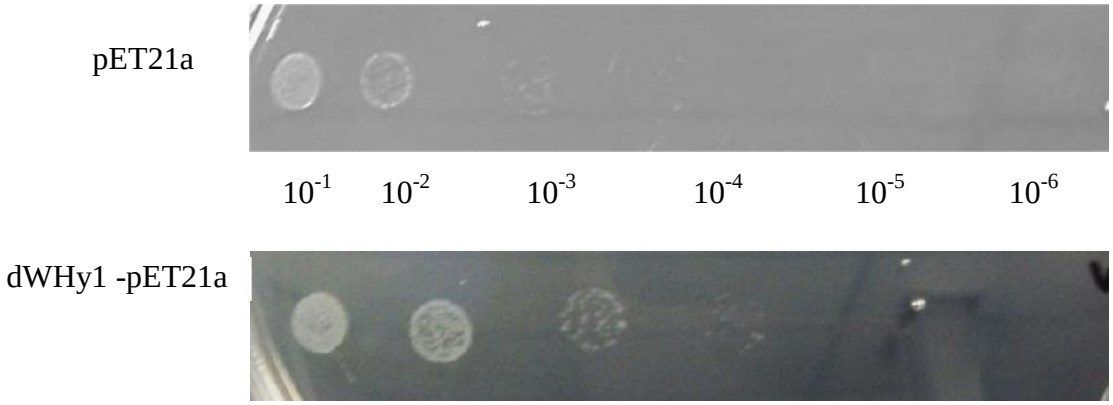


Figure 5.3.3.2: Growth of BL21 (DE3) *E. coli* on mannitol agar plates. Cultures were prepared as described and 5 µl of each dilution spot-plated onto LB agar containing 22 % D-mannitol.

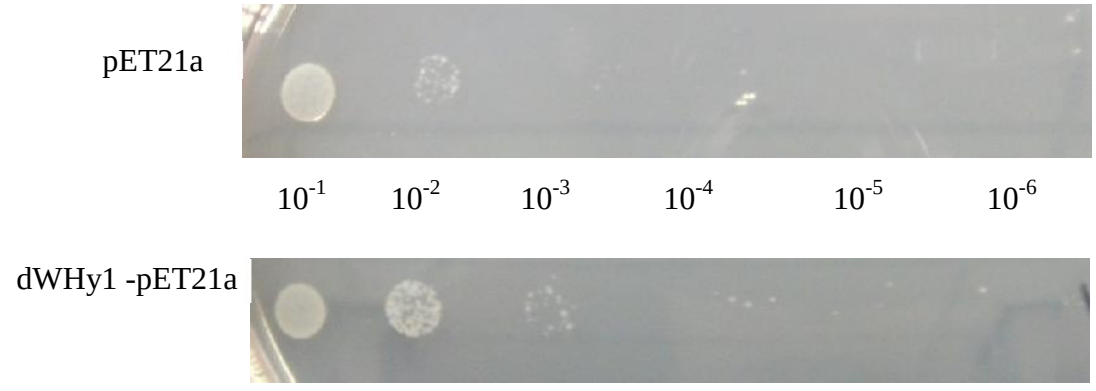


Figure 5.3.3.3: Growth of BL21 (DE3) *E. coli* on NaCl agar plates. Cultures were prepared as described and 5 µl of each dilution spot-plated onto LB agar containing 3.5 % NaCl.

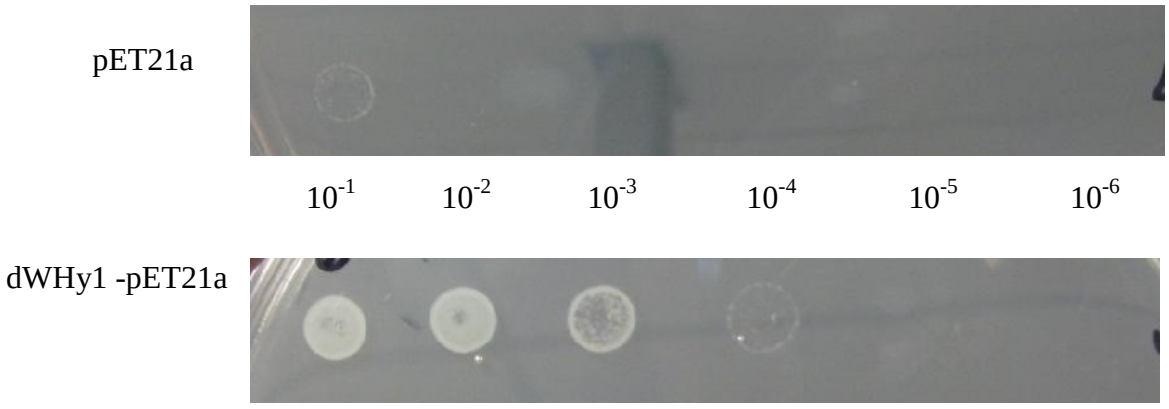


Figure 5.3.3.4: Growth of Rosetta (DE3) pLysS *E. coli* on mannitol agar plates. Cultures were prepared as described and 5 µl of each dilution spot-plated onto LB agar containing 22 % D-mannitol.

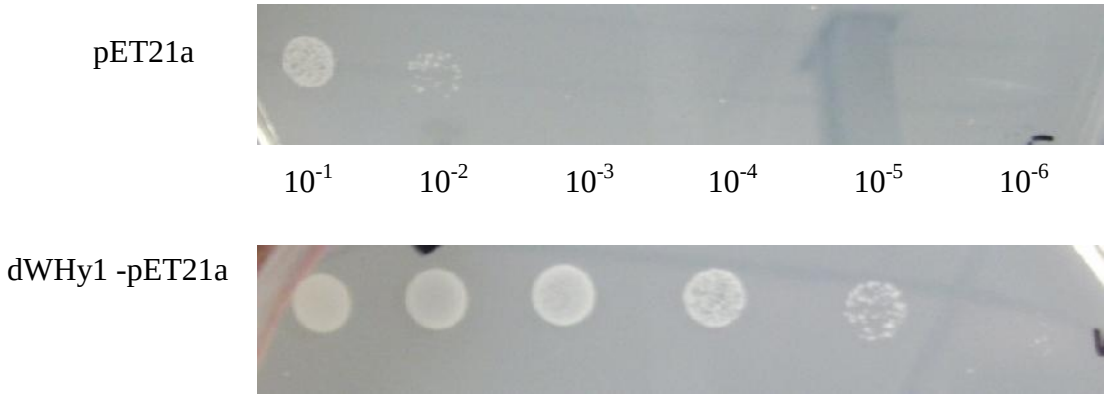


Figure 5.3.3.5: Growth of Rosetta (DE3) pLysS *E. coli* on NaCl agar plates. Cultures were prepared as described and 5 µl of each dilution spot-plated onto LB agar containing 3.5 % NaCl.

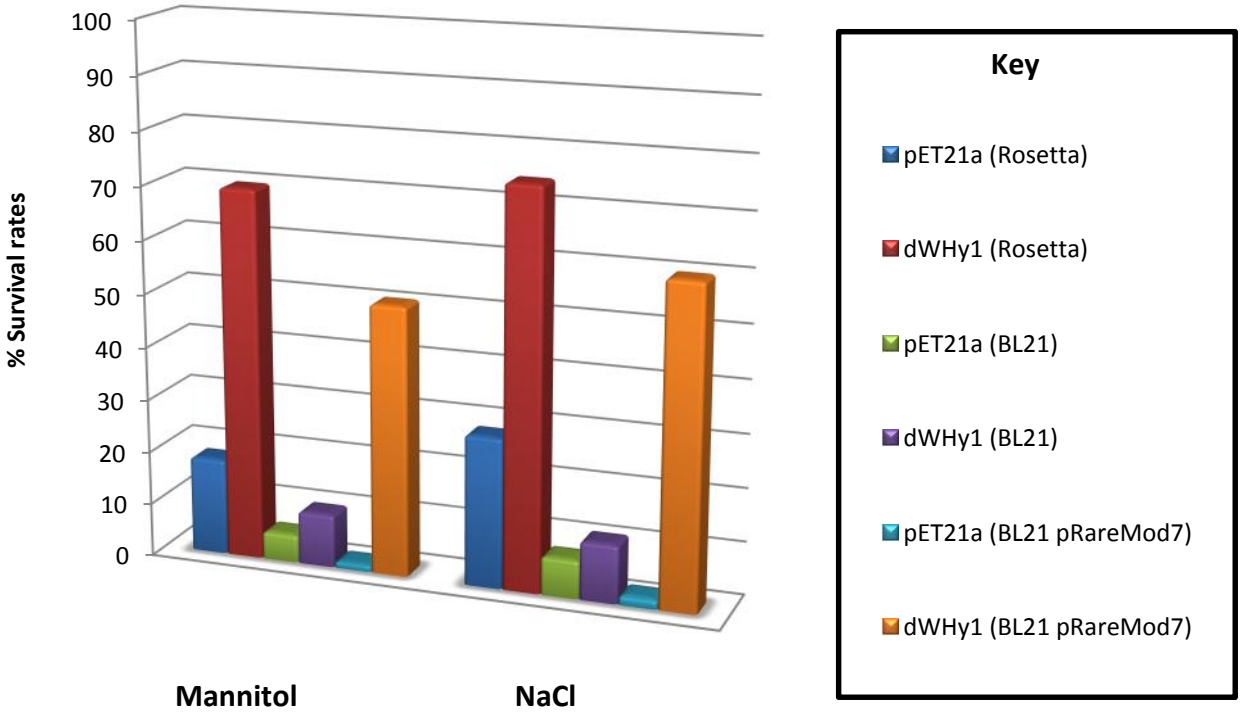


Figure 5.3.3.6: Graphical representation of the percentage survival rates calculated for desiccation tolerance conferred to *E. coli* strains expressing dWHy1. All cloned fragments were single copy genes.

5.3.4 Construction of pRareMod7

Results indicated that the presence of the chloramphenicol acyltransferase gene would be a major barrier to expression and purification of dWHy1. However, rare codons, which appeared to be important for dWHy1 function *in vivo*, are only supplied by *E. coli* strains which possess the pRare plasmid, all of which contain the CAT marker. pRareMod7 was developed to overcome limitations of rare codons and the presence of the CAT gene. pRareLysS is the native plasmid of the Rosetta (DE3) pLysS strain and can be extracted from

the host by alkaline lysis. A possible approach was to replace the CAT gene of this plasmid with the kanamycin aminoglycoside phosphotransferase (KAT) gene from pET28a, while maintaining rare codon supply. Unfortunately, the 7.4 kb sequence of pRareLysS is not available for public use. There was, however, sequence data for the CAT gene, which is derived from the vector pACYC. Two restriction sites, EcoRI and NcoI, would disrupt the CAT gene sequence, allowing for the insertion of KAT. The enzymes were initially used individually to digest pRareLys plasmid DNA, but when EcoRI was used, two fragments were visualised by gel electrophoresis (Figure 5.3.4.1 B), indicating a second EcoRI restriction site elsewhere in the vector. Digestion with NcoI liberated a single, linearized 7.4 kb fragment which was subsequently purified from the gel, blunt ends repaired and dephosphorylated to prevent self-ligation (Figure 5.3.4.1 A). The 1.3 kb KAT fragment (Figure 5.3.4.1 A) was successfully obtained from pET28a plasmid DNA and included the upstream promoter elements to ensure transcription in the new plasmid system. Recombinant clones were resistant to kanamycin and sensitive to chloramphenicol. However, this was not the only verification used to ensure pRareMod7 was functioning correctly. The plasmid DNA was restriction enzyme digested and expected profiles were compared to original pRareLysS digests. For instance; digestion of pRareLysS with EcoRI liberated two fragments. If construction of pRareMod7 was successful, digestion would again liberate two fragments, but one would increase in size by 1.3 kb (the KAT fragment contains no EcoRI restriction sites). As seen in Figure 5.3.4.1 B, expected digest patterns were observed and it was concluded that construction of pRareMod7 was successful.

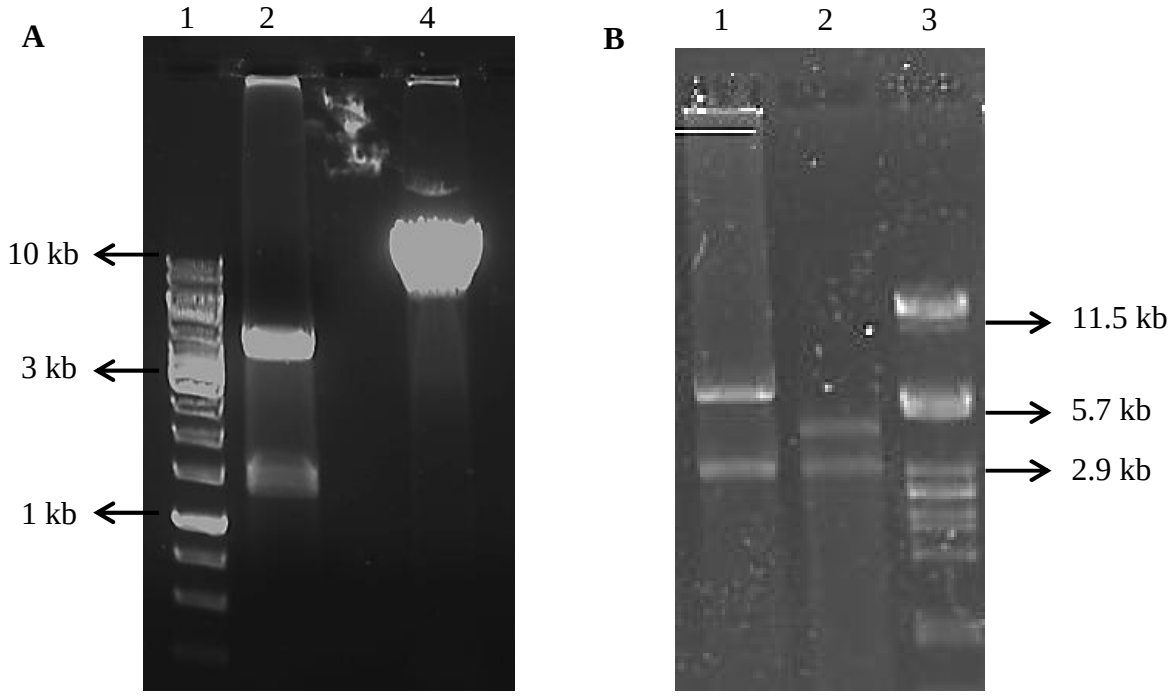


Figure 5.3.4.1: **A] Agarose gel electrophoresis of pET28a and pRareLysS restriction enzyme digest. Lane 1) GeneRuler™ 1 kb DNA marker (Fermentas). Lane 2) Double digest of pET28a with AlwNI and DraIII liberating 1.3 kb KAT gene. Lane 4) Single digest of pRareLysS with NcoI generating a linear 7.4 kb vector fragment. B] Agarose gel electrophoresis of pRareLysS and pRareMod7 digested with EcoRI. Lane 1) pRareMod7 digest showing a size shift of approximately 1.3 kb for one of the fragments. Lane 2) pRareLysS digest. Lane 3) Phage Lambda-PstI DNA marker.]**

In order to test whether pRareMod7 would still encode the rare codons, it was used to transform BL21 (DE3) cells. Rosetta cells could not be used as the original plasmid and the new construct have the same replication origin, and one would not be maintained in the system. Following verification procedures described previously, the dWHy1-pET21 construct, as well as pET21a parental vector, were transformed into the BL21 cells and *in vivo* phenotype assays were performed. Theoretically, a desiccation tolerant phenotype,

similar to that observed for the Rosetta (DE3) pLysS cells, would be observed. This was indeed the case (Section 5.3.3; Figure 5.3.3.5) and the newly constructed plasmid system was used for further expression studies. This strain is subsequently referred to as BL21pRM7.

5.3.5 Protein expression and purification

As previously mentioned, protein overexpression was required for *in vivo* phenotype assays. In the Rosetta (DE3) pLysS expression host transformed with dWHy1, different expression strategies consistently yielded an overexpressed band of 22 kDa (Figure 5.3.5.1 B). Expression studies under the same conditions did not show this protein band in cultures containing parental vector. In BL21pRM7 cultures, a faint protein band of the molecular mass expected for dWHy1 was observed and in order to produce enough protein for metal ion chromatography and subsequent size exclusion chromatography, culture volumes exceeding 2 L were used. One possible reason for the low protein expression is the metabolic burden imposed on the cells by the addition of two plasmid vector systems. Additionally, the pRareMod7 plasmid had increased in size from 7.4 kb to approximately 9 kb. Soluble fractions from both expression systems were subjected to metal ion chromatography but in the case of Rosetta (DE3) pLysS expression, the protein of interest consistently appeared in the wash fractions, irrespective of variations in salt and imidazole concentrations in the buffers (Figure 5.3.5.2 A). Analysis of BL21pRM expressions indicated the presence of two protein bands in fractions eluted with 500 mM imidazole, one approximately 22 kDa in size and corresponding to dWHy1, and the second of approximately 30 kDa (Figure 5.3.5.2 B). It was speculated that the second protein may be KAT, due to the high percentage of histidine residues in this protein, making it susceptible to co-purification. In addition, Bolanas-Garcia and Davies (2006) classify three groups of contaminating *E. coli* proteins which bind to IMAC columns. Class I proteins (Fur, ArgE, SlyD, YodA and others) and class II (CAT, YadF, G6-PD, GlgA) show strong binding affinity, requiring > 80 mM and 55- 80 mM

imidazole for elution, respectively. Aside from some exceptions, these proteins have pI values < 6 and more than 4 histidine residues. In addition, most of these proteins are stress-responsive enzymes which may be produced in high quantity due to cellular responses caused by heterologous gene expression (Bolanas-Garcia and Davies, 2006).

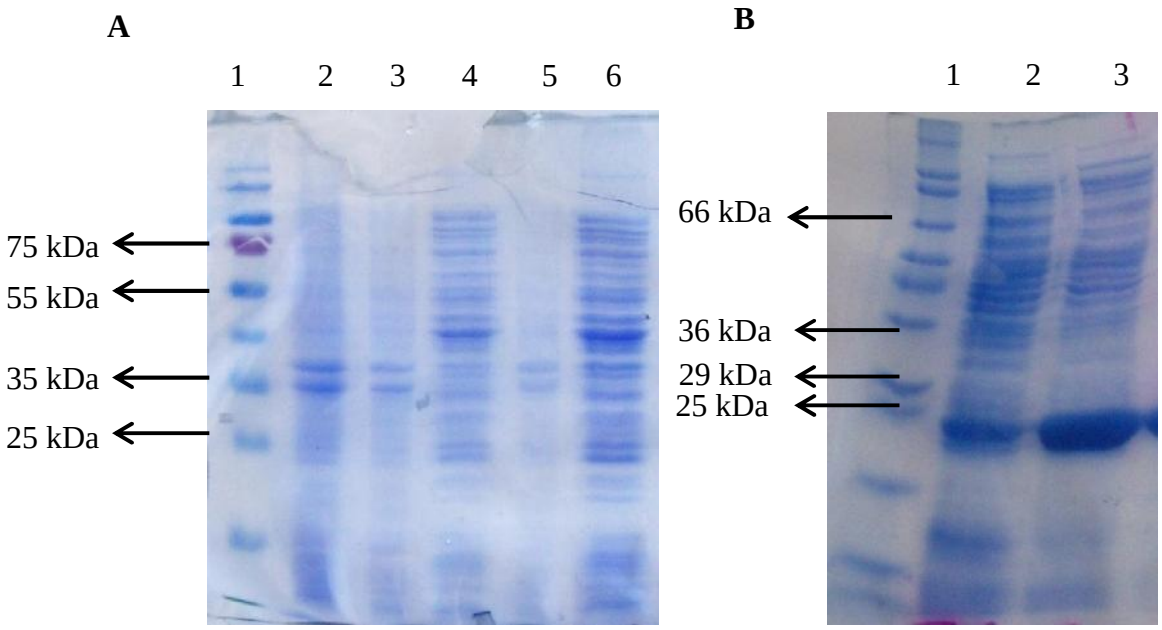


Figure 5.3.5.1: A] Protein expression of Rosetta (DE3) pLysS transformed with parental vector control pET21a. Lane 1) PageRuler™ Prestained protein marker (Fermentas). Lane 2 and 3) Uninduced insoluble fraction. Lane 4) Uninduced soluble fraction. Lane 5) Induced insoluble fraction (0.8 mM IPTG). Lane 6) Induced soluble fraction (0.8 mM IPTG) B] Protein expression of Rosetta (DE3) pLysS transformed with dWHy1-pET21a. Lane 1) WideRange™ protein marker (Sigma) Lane 2) Induced insoluble fraction (0.8 mM IPTG). Lane 3) Induced soluble fraction (0.8 mM IPTG).

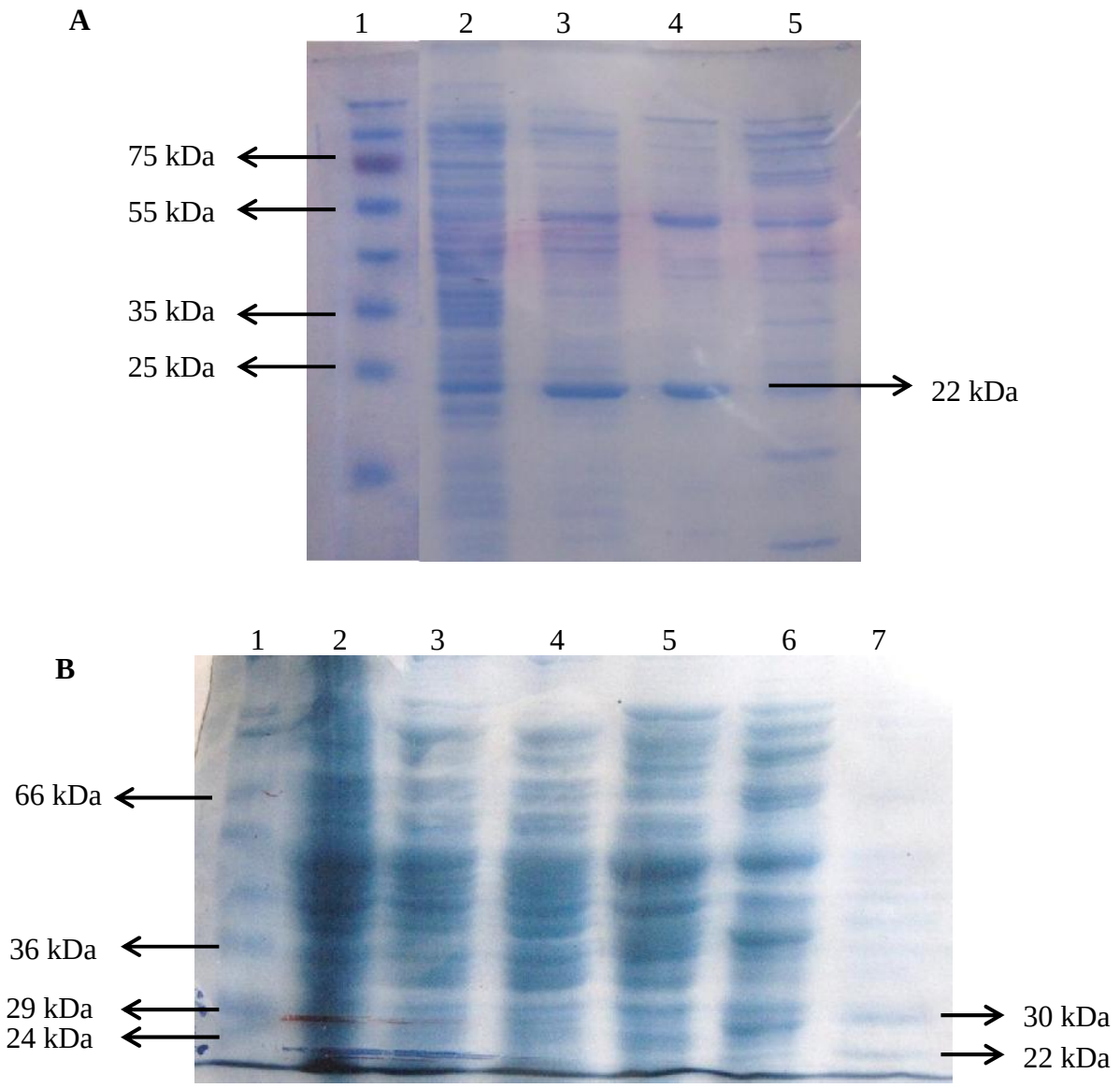
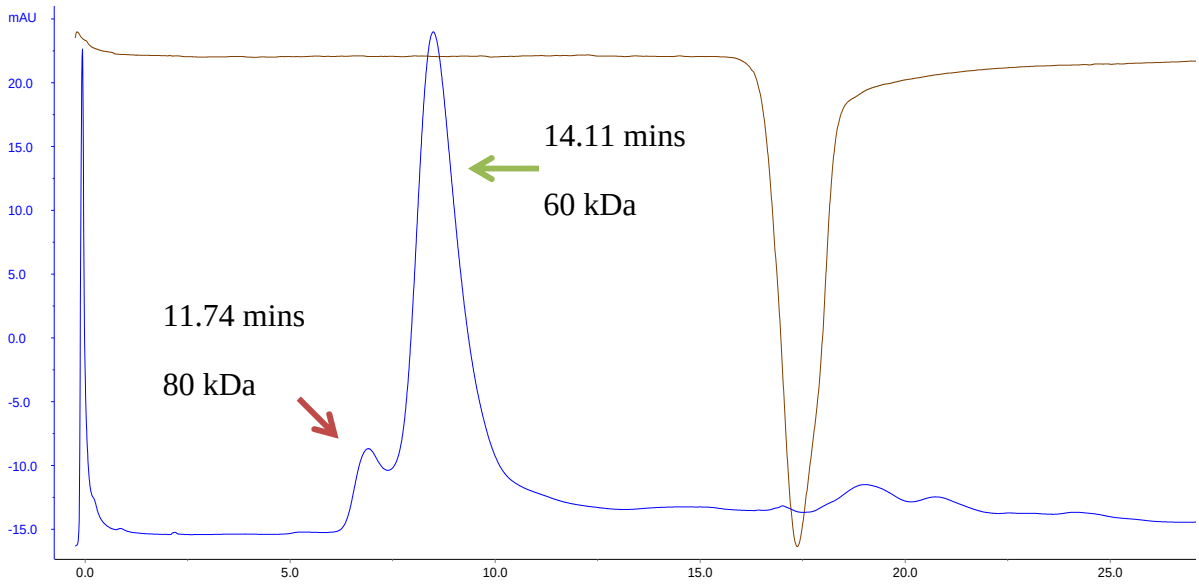


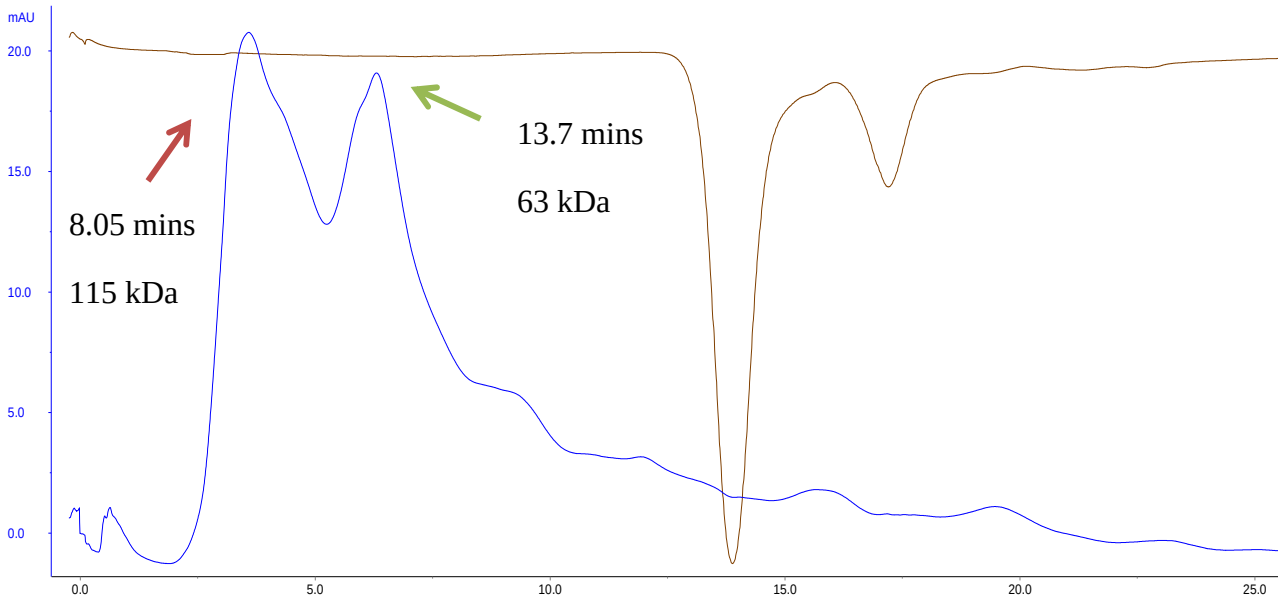
Figure 5.3.5.2: SDS-PAGE analysis of metal ion affinity chromatography. A] Purification of proteins from the Rosetta (DE3) pLysS expressions. Lane 1) PageRuler™ Prestained protein marker (Fermentas). Lane 2) Flow-through fraction. Lane3) Binding fraction. Lane 4) Wash Fraction. Lane 5) Elute fraction. B] Purification of proteins from BL21pRareMod7 expressions. Lane 1) WideRange™ protein marker (Sigma). Lane 2) induced insoluble fraction (0.8 mM IPTG). Lane3) Induced soluble fraction (0.8 mM IPTG). Lane 4) Flow-through fraction. Lane 5) Binding fraction. Lane 6) Wash Fraction. Lane 7) Elute fraction.

The wash and elute fractions from these experiments were dialysed and subjected to other chromatographic techniques. Size exclusion chromatography yielded two protein peaks for both BL21pRM7 and Rosetta expressions. In order to calculate the molecular mass of the proteins in the sample, retention times observed were related to those of the protein standards used. For Rosetta, the protein correlating to peak 1 was approximately 80 kDa in size and peak 2 was approximately 60 kDa in size. For BL21pRM7, the first peak correlated to 115 kDa and the second peak to 63 kDa. No additional peaks at the expected size range of 20 -25 kDa could be visualised (Figure 5.3.5.3 A and B). In these size exclusion chromatography studies, the observation of two common peaks from both expression strains which show similar retention times, and therefore a similar molecular weight, indicates the likely presence of dWHy1, although perhaps co-purified with a contaminating protein. It was hypothesised that dWHy1 may form a trimer in solution. For At1g01470, a crystallised homolog of dWHy1, protein structures which are related are known to form dimers and tetramers in solution (Shih *et al*., 2008) and it is therefore conceivable that dWHy1 could form multimers. In addition, fractions analysed by SDS-PAGE showed dissociation into monomer units of the expected 22 kDa size (Figure 5.3.5.3 C). Protein spots excised from the acrylamide gels were sent for Maldi-tof MS. Fractions were collected and stored at 4 °C and used in a subsequent MDH freeze-thaw cryoprotection assay (Section 5.3.6).

A


mAU
20.0
15.0
10.0
5.0
0.0
-5.0
-10.0
-15.0
0.0
5.0
10.0
15.0
20.0
25.0
14.11 mins
60 kDa
11.74 mins
80 kDa

B


mAU
20.0
15.0
10.0
5.0
0.0
0.0
5.0
10.0
15.0
20.0
25.0
13.7 mins
63 kDa
8.05 mins
115 kDa

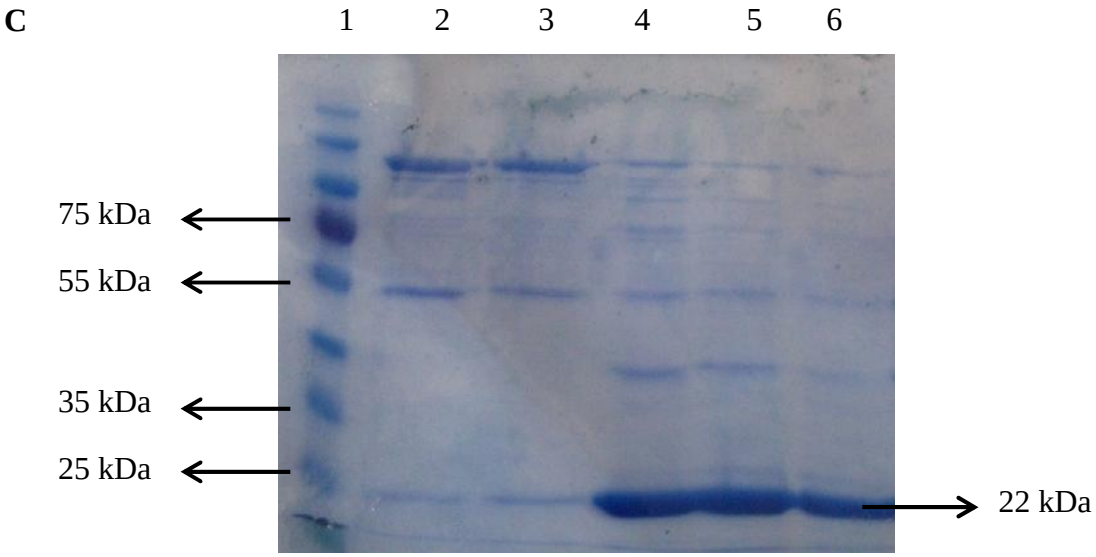
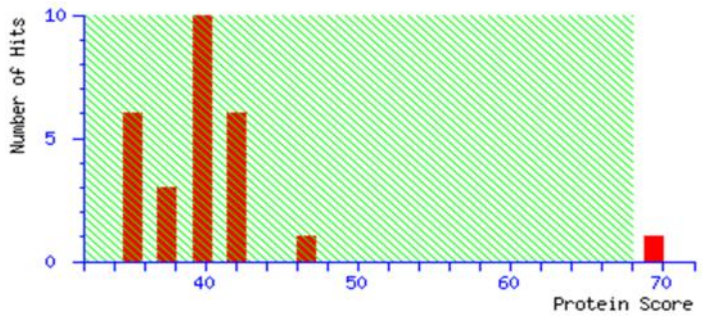


Figure 5.3.5.3: Graph of size exclusion FPLC of IMAC purified fractions from Rosetta [A] and BL21 [B] in 50 mM Tris-HCl (pH 7.5) and 400 mM NaCl. The blue line indicates the UV peak and the brown line shows conductance. X-Axis measurements are in ml (with fractions collected every 1 ml volume). The retention times and corresponding molecular mass estimates are indicated. C] SDS-PAGE of individual fractions collected from FPLC analysis (Rosetta). Lane 1) PageRuler™ Prestained protein marker. Lane 2 and 3) Fractions 5 -6 (Peak 1). Lane 4- 6) Fractions 7-9 (Peak 2).

As indicated previously, MALDI-TOF analysis revealed a completely unexpected result. Protein spots from the Rosetta (DE3) pLysS expressions clearly contained a mixture of proteins, with a major fraction being chloramphenicol acyltransferase (E.C. 2.3.1.28); Figure 5.3.5.4). This gene is the antibiotic marker of the pRareLysS plasmid in Rosetta. It translates to a 26 kDa protein with a pI of 5.4 and forms trimers in solution. Mascot results for the second peak protein spot from the BL21pRM7 experiments showed a likely match to a hypothetical protein homolog, but the match was not significant enough to provide complete verification. The other contaminating protein which appeared could possibly have been the KAT, similar to CAT, has a pI of 5.2 and also forms multimers in solution, explaining the

high molecular mass estimate from FPLC retention time analysis. Although it is known that contaminating proteins from an expression host may co-purify when IMAC columns are used, the levels of contaminant are generally low (Bolanos-Garcia and Davies, 2006). There are three main reasons why CAT contamination was unexpected in this case:

1. In expression trials with parental vector as control, an overexpressed 25 kDa band was not observed (Figure 5.3.4.1 A). In addition, this protein band was not observed in expressions with recombinant vector containing DEaseI or DEaseII (Chapter 4).
2. CAT is a 26 kDa protein and migration on polyacrylamide gels consistently showed a smaller molecular mass, approximately 22 kDa (Figure 5.3.4.1 B).
3. Although size exclusion profiles did show an 80 kDa protein peak, which would correlate to the CAT trimer, an additional peak at 60 kDa was observed. While dissociation of CAT into smaller subunits was possible, there is no evidence of the monomer unit which would ultimately be the largest peak, if this had occurred (Figure 5.3.5.2 A and B).



1.	CAT_ACIAN	Mass: 25646	Score: 69	Expect: 0.037	Matches: 7	
	Chloramphenicol acetyltransferase OS=Acinetobacter calcoaceticus subsp. anitratus GN=cat PE=3 SV=1					

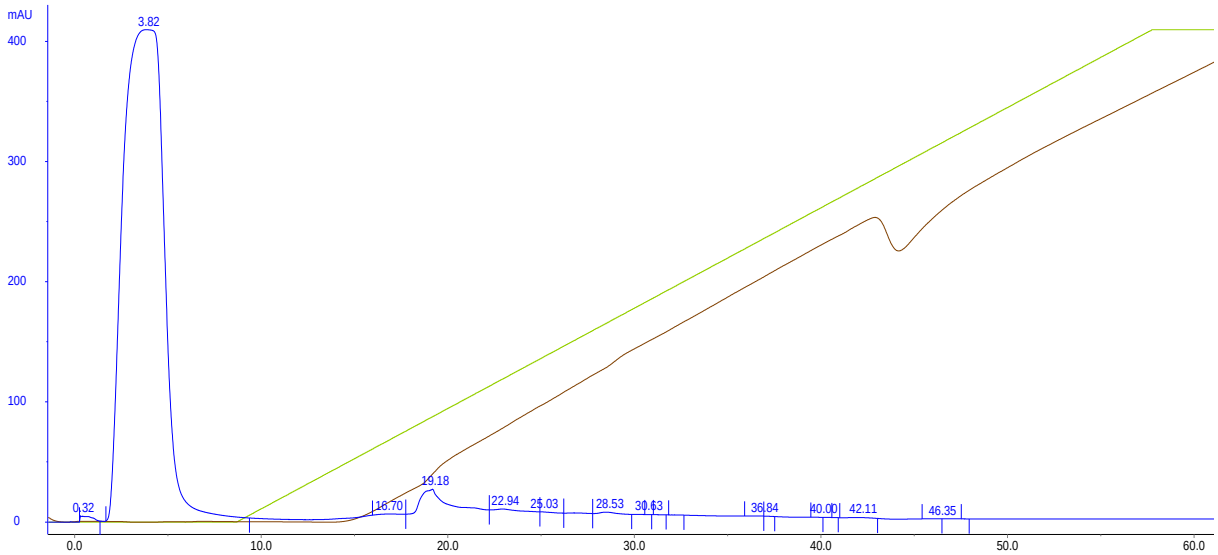
Figure 5.3.5.4: MS analysis of overexpressed protein band observed in the Rosetta (DE3) pLysS host. Mascot searches showed a top score of 69, which is statistically significant, to the CAT gene.

Despite the reasons given, it is clear from expressions with chimeric forms of dWHy1 in Rosetta (DE3) pLysS cells that overexpression of a contaminating protein does occur (Figure 5.3.3.1). However, it was confirmed in this study by testing strain BL21pRM7 that although a contaminating *E. coli* protein was present, it did not contribute to the desiccation tolerant phenotype observed.

For the further purification of dWHy1, an alternative method, not reliant on protein size alone, would be necessary. The theoretical pI of dWHy1 was estimated to be 8.99, while the pI of CAT is 5.4. Anion and cation exchange was therefore used in an attempt to purify dWHy1 from the co-purified contaminant. In both cases, the elution profiles of the IMAC wash fractions were typical of CAT, rather than dWHy1 (Figure 5.3.5.5 A and Figure 5.3.5.5 B).

A possible explanation for the continual co-purification of dWHy1 could be related to one of its proposed functions. Multiple studies on LEA proteins, relatives of the WHy domain proteins, indicate alternative 'moonlighting' capabilities. One of the reported activities of these proteins includes chaperone-like function, but unlike classical chaperones, they do not require energy generated from ATP hydrolysis (Hara, 2010; Kovacs *et al*., 2009). Interestingly, this ATP- independent chaperone function is also reported for some ribosomal proteins. These proteins which exhibit chaperone activities are critical for normal cell functioning, ribosome assembly function, and ultimately transcription processes (Kovacs *et al*., 2009). If dWHy1 exhibits strong chaperone activity, it would remain associated with the contaminating protein (W. D. Schubert personal communication).

A



B

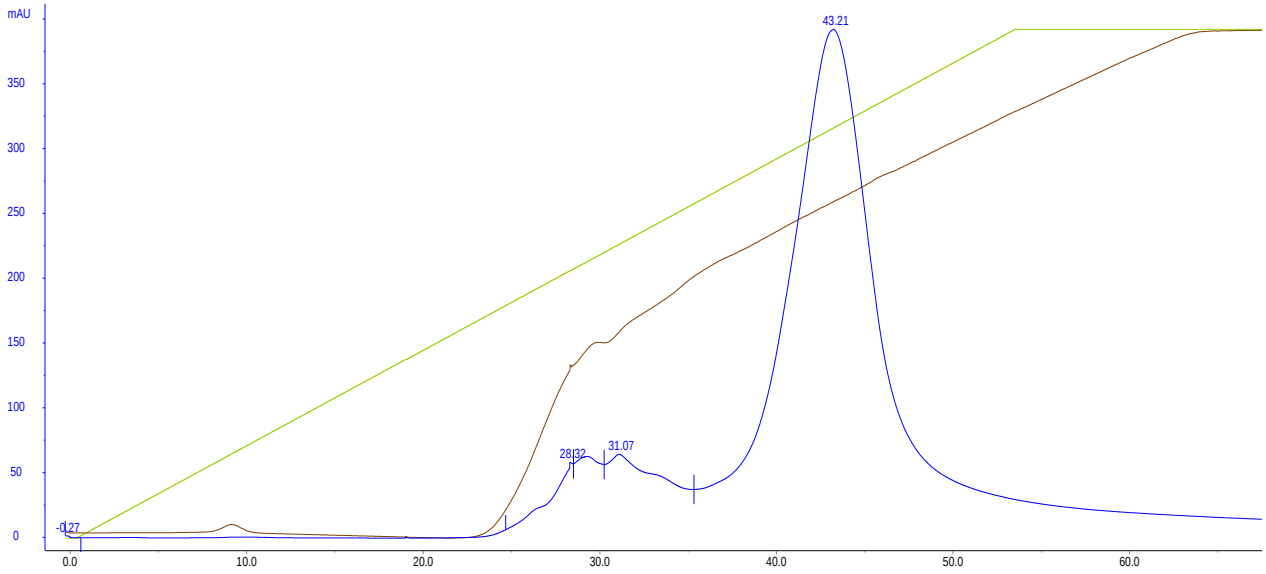


Figure 5.3.5.5: Graph of ion exchange FPLC of IMAC purified fractions from Rosetta [A] Cation exchange profile in 50 mM CAPS (pH 9.0) [B] Anion exchange profile in 50 mM MES (pH 5.5). The blue line indicates the UV peak, the brown line shows conductance while the green line indicates the NaCl gradient profile from 0 to 1 M. X-Axis measurements are in minutes.

A review of current literature indicated that a high level of CAT expression is not an uncommon phenomenon. Co-purification of CAT encoded on pACYC184 derived plasmids with the target recombinant protein has been reported (Hengen, 1996). Additionally, the 26 kDa protein contaminant generally cannot be removed either by addition of Triton X-100 to the elution buffer, an increased imidazole concentration in the binding buffer or increased NaCl concentrations in both the binding and elute buffers (Oswald and Rinas, 1996). Oswald and Rinas (1996) went so far as to recommend complete avoidance of this antibiotic marker in expression strains. Co-purification of CAT is most likely due to the histidine content of the protein, but one question which remains unanswered is why this protein is a problem in some recombinant protein expressions and not in others. For instance, for protein expression and purification of DEaseI and DEaseII from the Rosetta (DE3) pLysS strain (Chapter 4), CAT was neither overexpressed, nor co-purified. It is highly likely that co-purification of this protein becomes a greater issue when it is overexpressed alongside the target molecule.

Literature searches retrieved some experimental evidence which may explain the over-expression of CAT and subsequent purification issues. In addition, this may be applied to explain observations with dWHy1 in this study. Vemulapalli *et al* (2000) report overexpression of the CAT protein encoded on the plasmid pBBR1MCS. In this case, the authors were attempting to overexpress the recombinant protein Cu/Zn superoxide dismutase [Cu/Zn SOD] (E.C. 1.15.1.1) from *Brucella abortus* (Vemulapalli *et al.,* 2000). This metalloenzyme protects cells from oxidative toxicity of superoxide molecules (Bolanas-Garcia and Davies, 2006). The authors were able purify the target protein by anion exchange (Vemulapalli *et al.,* 2000). Similarly, Haslam *et al* (2002) indicated that CAT overexpression and co-purification occurred while attempting to characterise an *S*-formylglutathione hydrolase from *Arabidopsis thaliana*. This enzyme is required for formaldehyde detoxification and subsequent glutathione recycling, leaving formate as a carbon source via

the C1 pathway (Haslam *et al*., 2002). In another study, Kim *et al* (2009) report that a 1.5 fold increase in CAT expression was observed when RelA together with mutant ribosomes pRNA122-U791 was overexpressed compared to pRNA122-U791 alone. RelA synthesises the alarmone ppGpp, which interacts with ribosomes and triggers a bacterial adaptation response known as the stringent response (Kim *et al*., 2009). Furthermore, the authors report that RelA most likely complimented mutated ribosomes by direct or indirect association of the protein with the mutant 790 loop region, thereby changing the non-functional structure to a functional one (Kim *et al*., 2009). While the parameters for RelA-ribosome binding are not well characterised, RelA has been shown to bind to the 50 S ribosomal subunit (Ramagopal and Davis, 1974), and the ribosomal protein L11 regulates RelA activity by acting like a molecular switch which regulates accessibility of RNA in the GTPase- associated site during protein elongation (Yang and Ishiguro, 2001; Wimberly *et al*., 1999; Kim *et al*., 2009). Interestingly, a study by Gill *et al* (2000) demonstrated that the levels of proteases can also be increased during the stringent response and, by increased levels of CAT expression (Gill *et al*., 2000). This is due to the rapid depletion in phenylalanine resulting in the stringent response, and the authors clearly show that the OmpT protease levels increased with the concomitant increase in CAT expression.

In the first three studies, CAT overexpression appears to be linked to the expression of another target protein and, as in this study, all targets were involved in a stress response mechanism. In the study by Gill *et al* (2000), CAT overexpression actually increased target protein levels. Again, the target protein is related to a stress response and it is known that both molecular chaperones and proteases are overexpressed in response to heat stress (Gill *et al*., 2000). It is also known that molecular responses to stress may overlap, due to the upregulation of genes in response to multiple independent stimuli (Gill *et al*., 2000). In addition, studies in plants have demonstrated a general response to injury which includes

changes in the expression levels of chaperones, proteinases and other detoxification proteins, LEA proteins appear to bind molecules that are susceptible to the effects of stress and may reduce the damage caused by specifically targeting those macromolecules (Hara, 2010). Stress tolerance therefore results from the control and repair of damage incurred by the stress (Mahajan and Tuteja, 2005).

Since dWHy1 is involved in the desiccation stress response, it may cause CAT overexpression by an as-yet-unknown mechanism. It can be speculated that dWHy1 may function in a similar manner to L11. Ribosomal proteins (RP) are partially disordered (30 %) and demonstrate promiscuous behaviour, with the capacity to bind both RNA and proteins (Tompa and Kovacs, 2010). Furthermore, extra-ribosomal functions relating to RNA, DNA and protein interactions have been demonstrated for RP chaperones and, even at low levels of protein, are able to perform functions which may be critical for normal cell functioning, ribosome assembly function and, ultimately, transcription (Kovacs *et al*., 2009).

One possible target protein, in the case of dWHy1, could be antibiotic resistance genes, CAT and KAT. If this is the case, then one question arises as to why dWHy1 does not associate with the pET21a marker, the β-lactamase protein. One explanation may be the particular antibiotic targets. When translation mechanisms are stopped abruptly under a given stress condition, available energy and resources can be directed to only those stress proteins that are deemed essential (Henderson and Pockley, 2010). Ampicillin inhibits bacterial cell wall synthesis, ultimately causing lysis, whereas kanamycin and chloramphenicol inhibit bacterial growth by damaging ribosomes and preventing translation. dWHy1 may be involved in the up-regulation of proteins to counteract the effects of unfavourable ribosome functioning, effectively acting like a molecular switch. Furthermore, dWHy1 may exhibit another function as a chaperone, in the protection of protein targets required for counteracting a particular stress from denaturation events, until complete ribosome function is restored and protein

synthesis is no longer affected. In addition, co-expression of interacting partners may be required for the correct folding of certain proteins and the target protein may be protected from proteolytic cleavage during co-expression (Sørensen and Mortensen, 2005).

Considering that bacterial proteins which perform multiple functions are generally stress related proteins (Henderson and Pockley, 2010**)**, and due to predicted phosphorylation sites, it can be hypothesised that dWHy1 is part of a general stress response cascade, leading to the induction of genes which repair stress damage. This would indicate that the protective function of dWHy1 function is not limited to desiccation stress tolerance but also covers toxic stress caused by antimicrobial compounds.

5.3.6 *In vitro* freeze thaw assays

In order to determine if dWHy1 had a cryoprotective function, freeze–thaw assays were performed *in vitro.* Osmotic stress and freeze-thaw stress both negatively affect systems in a similar manner. At sub-zero temperatures, ice formation alters solute concentration and causes fluctuations in intracellular pH (Potts, 1994). Both malate dehydrogenase and lactate dehydrogenase are sensitive to the effects of freeze-thaw and Reyes *et al* (2008) demonstrated that MDH appears more sensitive to this treatment. It is for this reason that MDH was used in this study. Fractions obtained by FPLC, which correlated to the 60 kDa protein from both Rosetta (DE3) pLysS and BL21pRM7, were used for these assays. Although the purity of these fractions was not known, the molar ratio of dWHy1 to enzyme was calculated based on the monomer molecular mass of dWHy1. The molar ratio of dWHy1 to MDH used in these assays was 5:1; in this case, if there was an equal mixture of two proteins with similar molecular mass, such as CAT or KAT, the ratio of dWHy1 would be close to that used by Reyes *et al* (2005).

As seen in Figure 5.3.6.1, dWHy1 protected MDH from cycles of freeze-thaw damage to a greater extent than either BSA or trehalose, two known cryoprotectants. In this study,

glycerol proved to be an excellent protectant of MDH up to 3 cycles of freeze-thaw. In the case of dWHy1 [BL21pRM7], enough protein could not be obtained to assay all three freeze-thaw cycles and it was therefore decided that triplicate measures of 3-, and 5- freeze-thaw cycles would provide comparative information. dWHy1 [BL21pRM7] protein fractions clearly provided greater protection than Rosetta (DE3) pLysS fractions used in this study. This may be due to the presence of a higher concentration of dWHy1 protein in the BL21pRM7 fractions, while those from the Rosetta strain may contain higher concentrations of CAT thereby reducing the concentration of dWHy1. Nonetheless, in this study we show that dWHy1 has a protective function against desiccation caused by freezing. These results were similar to *in vitro* lactate dehydrogenase- hydrophilin freeze-thaw assay data obtained by Reyes *et al* (2008) and Goyal *et al* (2005).

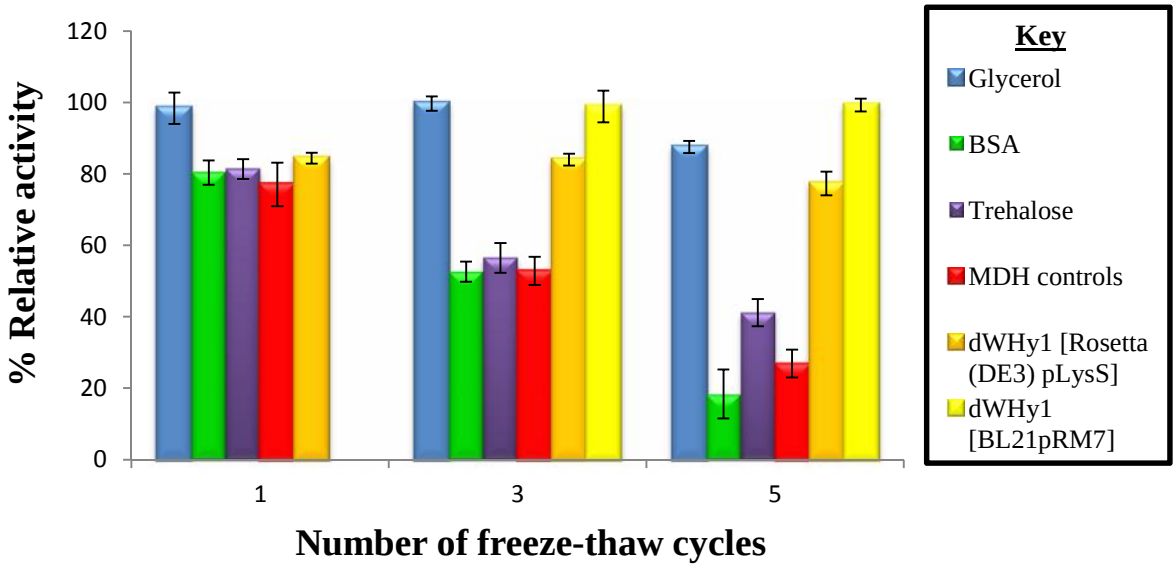


Figure 5.3.6.1: ***In vitro*** **freeze-thaw assay of MDH with and without protectant. Activity was assayed in 150 mM potassium phosphate buffer and measured by a decrease in absorbance at 340 nm for 1 minute at 25 °C, from the conversion of NADH to NAD^+. The rate obtained for the untreated MDH samples was taken as 100 %. All assays were performed in triplicate.**

How dWHy1 and similar proteins perform this cryoprotective function is still unknown. One hypothesis regarding the mechanism of action of dehydrins is based on the preferential exclusion hypothesis, which states

> *"…a stabilizing solute is excluded from the immediate vicinity of a protein. The protein is, in effect, preferentially hydrated as preferential exclusion is entropically unfavourable, it is a driving force to establish equilibrium between the native and unfolded state, towards protein stabilisation"* (Potts, 1994).

In this case, hydrophilins would act in a manner similar to osmoprotectant molecules such as trehalose. Experiments by Reyes *et al*, confirm that hydrophilins protect enzymes from adverse effects of water loss, which occur during freeze-thaw and dehydration events (Reyes *et al*., 2005; Reyes *et al*., 2008). However, the authors state that, due to massive differences in the molar concentrations required for trehalose-based protection versus dehydrin-based protection of LDH, that these two compounds probably confer protection by different mechanisms (Reyes *et al*., 2008).

LEA proteins are generally characterised by a high proportion of low complexity (LC), highly hydrophobic sequence features (Kriško *et al*., 2010). Regions of LC provide proteins with increased flexibility, allowing for promiscuous activity (also known as protein moonlighting). In eukaryotes, LC sequences mediate protein-protein, protein-DNA and protein-ligand interactions in various cellular processes (Kriško *et al*., 2010). It is believed that a number of rare prokaryotic proteins which exhibit similar characteristics may function in a similar manner. These LC regions give rise to increased solvent accessibility thereby providing multiple sites for interactions with polar solvents, such as water. It is believed that this particular property allows for the interaction of LC protein polar side chains with protein surfaces in the absence of water, effectively mimicking water molecules (Kriško *et al*., 2010).

In addition, antifreeze proteins are produced in cells and prevent damage by preventing the formation of ice crystals. In a cold and desiccated environment, such as the Antarctic Dry Valley soils, ice crystal formation requires a minimum of 200 molecules of water to associate in a 4 nm domain, at – 40 degrees Celsius. Therefore, the lower water availability becomes, the lower the temperature required for ice crystal formation (Potts, 1994).

It may be hypothesised that dWHy1 prevents enzyme inactivation from dehydration events by binding excess available water. In addition, dWHy1 may bind to a protein partner, effectively creating a hydration shell around the protein and, by lowering available water in the immediate vicinity, thus decreasing the probability of ice crystal formation.

5.3.7 *In vitro* protein synthesis

In a study by Livernois *et al* (2009), the authors exploited the heat stability property observed for LEA proteins to develop a purification strategy. Pure protein was obtained following boiling (to lyse cells and remove contaminating proteins) and a single reverse phase HPLC step. Unfortunately, this was not an option for dWHy1 purification as the protein is a member of the atypical LEA proteins, known for their heat instability. In order to obtain pure protein free of contaminating *E. coli* proteins, an *in vitro* transcription and translation kit was used. dWHy1 was successfully amplified and cloned into the pET17b vector system and used for *in vitro* protein synthesis. The expected size of dWHy1 protein was 18 kDa. For the PURExpress™ system, the target protein is purified from the reaction mixture by reverse metal ion chromatography, using the batch method. As seen in Figure 5.3.7.1, dWHy1 was successfully translated and increasing amounts of protein was obtained with an increase in plasmid DNA concentration. The optimum DNA concentration was 300 ng. Following expression, protein concentration columns were used to separate proteins using a molecular weight cut-off of 50 kDa. Smaller proteins were subjected to IMAC chromatography and, as shown in Figure 5.3.7.2, this method allowed for successful purification of dWHy1.

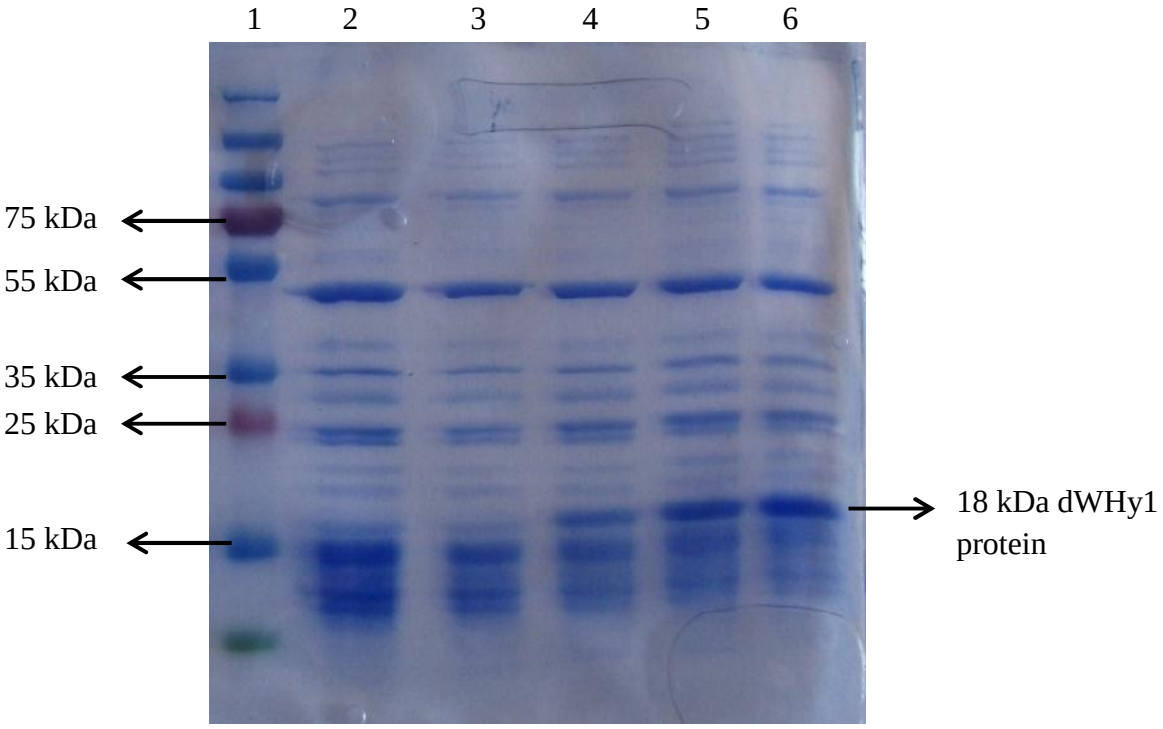


Figure 5.3.7.1: ***In vitro*** **protein synthesis of dWHy1-pET17b recombinant. The 18 kDa protein is clearly visible without any existance of it in the control reactions. Lane 1) PageRuler™ Prestained protein marker. Lane 2) Reaction containing no DNA. Lane3) Reaction containing 200 ng of pET17b parental vector DNA as control. Lane 4) Reaction containing 60 ng of dWHy1-pET17b recombinant DNA. Lane 5) reaction containing 120 ng of dWHy1-pET17b recombinant DNA. Lane 6) reaction containing 300 ng of dWHy1-pET17b recombinant DNA.**

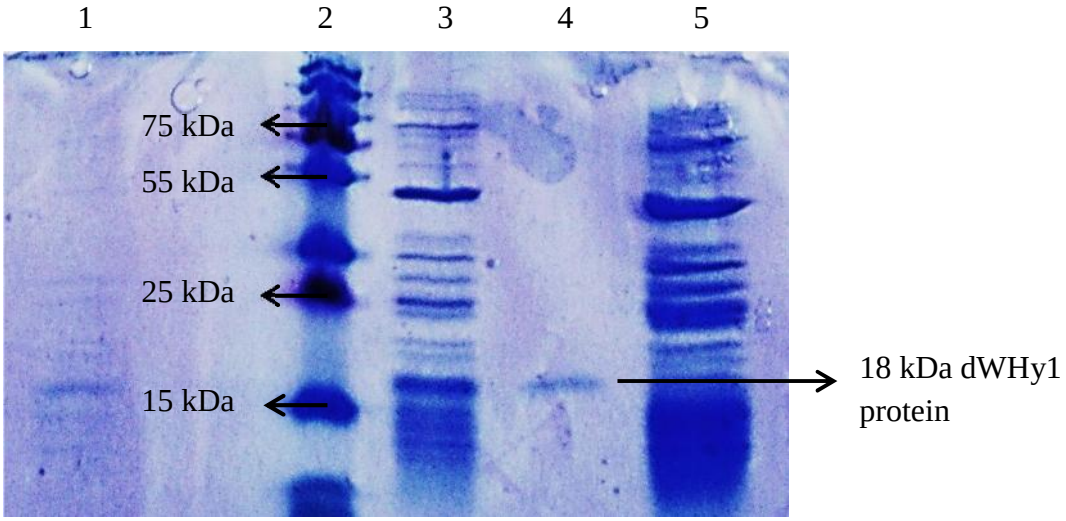


Figure 5.3.7.2: Purification of dWHy1 from *in vitro* transcription and translation. Lane 1) Flow-through fraction from centrifugation in Amicon® Ultra-0.5 filter devices (Millipore). Lane 2) PageRuler™ Prestained protein marker. Lane 3) Retentate from centrifugation in Amicon® Ultra-0.5 filter devices (Millipore). Lane 4) Flow-through fraction from metal ion chromatography. Lane 5) Protein synthesis reaction containing no DNA.

Due to the significant time investment in developing the purification of dWHy1, it was not possible to comprehensively investigate the functions of this unique protein using enzyme assays. However, some considerations are laid out in Section 5.4.

5.4 Conclusions and future considerations

In this study, a putative water hypersensitive response protein was successfully cloned, expressed and purified. In addition, this protein has been shown to be functional, and actively confers ionic and osmotic tolerance to the heterologous *E. coli* host *in vivo*. It was concluded that rare codons are essential to dWHy1 function *in vivo*, as this phenotype was not observed in hosts which do not encode rare codon tRNA's. A number of problems associated with the purification of dWHy1 led to an interesting observation; overexpression of dWHy1 leads to the overexpression of the CAT protein, transcribed and translated from the pRareLysS plasmid of the Rosetta (DE3) pLysS expression strain. Literature surveys indicated that studies reporting similar observations all generally involve a variety of stress related proteins. All such studies, and the current study, do not provide an explanation for this observation, but it is evident that this aspect of bacterial physiology should be investigated further. Complications in purification and expression also led to the development of a modified vector system, whereby CAT was replaced with KAT. While it appeared that this new vector system was effective in providing rare tRNA's, as demonstrated by successful *in vivo* and *in vitro* studies, it would be of great value to obtain the vector sequence of the pRareLysS vector in order to employ possible novel improvements, which would ultimately lead to fewer limitations in fundamental biology research.

In this study, it was shown that dWHy1 can protect enzymes from freeze-thaw damage *in vitro.* Furthermore, dWHy1 was successfully expressed and purified using a cell-free protein synthesis kit and the preparation appeared to be free of contaminating proteins. Using the system described, further research into the functions of dWHy1 can be performed. The ultimate goal of future work would include X-ray crystallography studies and elucidation of the mechanism of function of dWHy1. To the authors knowledge, this is the first report of a

WHy protein occurring in an Antarctic Dry Valley metagenome, and the first attempt at *in vivo* and *in vitro* characterisation of a bacterial WHy protein. This study allows for the development of testable hypothesis to elucidate mechanisms employed by bacteria in these Dry Valley soils to cope with the extreme desiccation conditions.

Future considerations

In this study, disordered proteins such as LEA were used as a general guide to develop further hypotheses and experimental procedures for testing. These proteins have been implicated in a number of functions and, while most do not exhibit all of the functions, it may be valuable to test dWHy1 for a variety of activities. Firstly, phosphorylation sites occur with high frequency in disordered regions of proteins and this post-translational modification is important for regulating cellular function (Mouillon *et al*., 2008). Considering that putative phosphorylation sites have been detected by bioinformatic analysis, it would be interesting to determine if the putative phosphorylation sites in dWHy1 can be phosphorylated. This may indicate that dWHy1 plays a key role as a regulatory factor for the interaction with protein targets. For example, experiments with α-synuclein demonstrated that the protein could be phosphorylated by casein kinase II which increased the disorder character of the protein and α-synuclein self-association (Sasakawa *et al*., 2007). In addition, the binding of bivalent metal ions may also indicate a possible role for dWHy1 in metal ion scavenging processes.

Secondly, dehydrin proteins ERD10 and ERD14 have been shown to protect various substrates from heat induced protein aggregation, exceeding the protection conferred by BSA under the same conditions (Kovacs *et al*., 2008). In addition, the disordered protein α-synuclein exhibits chaperone-like activity and has been evaluated for the ability to protect microbial esterases from high temperatures, low pH and organic solvents. Other disordered proteins such as α-casein and microtubule associated protein 2 (MAP2) have been used to

prevent heat and chemical induced aggregation of a variety of unrelated proteins and enzymes (Tompa and Kovacs, 2010). dWHy1 should therefore be tested for its ability to protect a variety of enzymes from potentially damaging conditions.

Thirdly, LEA proteins such as LEA18, ERD10 and ERD14 have been shown to bind acidic phospholipid vesicles, possibly indicating some function in maintaining bacterial or plant cell membranes under adverse conditions. dWHy1 should also be evaluated for this ability as it would implicate involvement in stress response systems at the onset of the condition. As with the group V LEA protein from *Arabidopsis*, dWHy1 should also be tested for involvement in other stress conditions, such as oxidative stress, where it may function directly as a free radical scavenger, or indirectly, by controlling transcription and translation of other proteins involved in the oxidative stress response, or by protecting enzymes from oxidative damage.

Finally, assessing dWHy1 for RNA, DNA and protein interactions would be valuable in elucidating the mechanism of action of this multi-functional desiccation tolerance protein:

> "*Understanding the mechanisms of desiccation tolerance holds promise not only because it may solve important problems in cell biology but also because it may find biotechnological applications in conferring desiccation tolerance on otherwise desiccation-sensitive microorganisms*" (Billi and Potts, 2002).

Chapter 6: General conclusions and considerations.

General conclusions and considerations

Metagenomics allows researchers to investigate the 'yet-to-be-cultured' microbial world, and the full potential of this approach has been realised by recent vast improvements in sequencing technology. Culture-independent techniques are used to examine the genetic complement of micro-organisms and any conclusions drawn, are therefore subject to the strengths, and weaknesses, of bioinformatic analysis of the metadata obtained. Furthermore, annotation of this data is affected by curations in extant databases, where uncorrected errors will be carried over to current and future studies of this nature. As such, it is important to experimentally verify the inferred function of predicted genes and to investigate possible new functions. While DNA-based techniques allow for the discovery of novel sequences, they do not permit the study of the organisms themselves and as such, the mechanism by which genetic elements, and their often complex regulatory systems, influence phenotype in a particular organism cannot be investigated. Nonetheless, the sequencing of large insert metagenomic clones can allow the linkage of genes to organisms (if phylogenetic markers are present in the sequence), and therefore the linking of organisms to function. Selection based on phylogenetic lineage provides evolutionarily relevant data while a function-based approach may provide data which is ecologically relevant.

In this study, several clones were selected from a large insert metagenomic library using a function-based approach, and sequenced using Illumina technology. The motivation for this was twofold; first, the generation of sequence data would provide a platform for *in silico* analysis leading to the development of hypotheses regarding microbial adaptation to the extreme Antarctic environment and, secondly, this data could be used to investigate the function of individual genes which may be valuable in industrial applications or which may contribute to a greater understanding of fundamental biology.

Bioinformatic analysis of the sequence data revealed diverse strategies for adaptation to the abiotic factors in the Antarctic environment. The observation of similar genes within three individual fosmid clones may reflect common strategies in a variety of microbes and could also indicate that these genes would be part of the pan-genome of organisms which inhabit cold environments. Although a number of interesting hypotheses can be drawn with respect to functional capacity, sequence data alone cannot be used to determine gene function and, as with many sequence-based projects, the presence of hypothetical proteins hinders progress toward elucidating complex genetic and metabolic interactions.

Aside from investigating the biocatalytic potential of novel lipolytic genes (Chapter 4), the complete fosmid sequences allowed for the identification of a Water HYpersensitivity stress response protein. While there are a number of reports on the existence of low complexity desiccation tolerance proteins, these proteins have low sequence homology to dWHy1, and all functionally characterised members are from higher eukaryotes. The first *in vivo* and *in vitro* characterisation of a bacterial WHy protein is presented in this study (Chapter 5). The analysis of the functional protein may point to a novel mechanism for desiccation survival employed by microbes inhabiting Antarctic desert soil environments. In addition, protein expression studies yielded an unexpected result; the overexpression of the CAT gene. This is not the first report of CAT overexpression and co-purification and in all cases where this phenomenon is observed, proteins under investigation were involved in some stress response.

This observation in itself warrants further investigation. Successful expression of this protein was obtained by use of a cell-free transcription and translation system, allowing for further research in an attempt to elucidate the mechanism of action of WHy proteins.

This study utilises a combination of techniques to evaluate a portion of the Antarctic metagenome for genes which may be used in industry. For example, the characterised

lipolytic enzymes may be used for the resolution of racemic mixtures or in bioremediation schemes. In addition, and with reference to *DEase*II, novel structural changes were observed which may have evolved to increase catalytic efficiency and substrate specificity of this protein. The WHy protein may be considered for use in the agricultural industry, as genetic engineering with this gene may impart desiccation tolerance in valuable food crops. Furthermore, examination of cloned sequences correlated well with current knowledge of the physiological adaptations of microbial consortia in cold environments and may provide clues as to the mechanisms by which they sustain essential biogeochemical cycles in the Antarctic Dry Valley soils. This thesis comprises a full portfolio of techniques required to carry out a culture-independent study on environmental samples from one of the harshest environments on the planet. From DNA extraction, to cloning, protein expression, sequencing, *in vitro* assays and bioinformatic analysis, this work demonstrates the synergy achieved when complementary techniques such as these are employed.

This study is the first sequence-based investigation of the Dry Valley soil metagenome and provides a platform to enable large-scale comparative ecogenomics. Therefore, future studies should include further sequencing of the Antarctic metagenome, which would most definitely reveal novel molecular pathways employed by microorganisms which have colonised this extreme environment. Inferring ecosystem function by interpretation of genomic data should, however, be performed with caution. In environments where temperatures are constantly low, degradation of bacterial DNA may occur at slower rates and as such, these genetic elements can persist (Willerslev *et al*., 2010). Therefore, to truly understand the physicochemical contributions of functional, complex microbial consortia, metatranscriptomic and metaproteomic methods should be used in conjunction with metagenomic studies. In addition, a global understanding of microbial metabolism may also provide the basis for development of new culture-based methodologies for the enrichment of culture collections.

References

References

1. Adams, B. J., R. D. Bardgett, *et al*. (2006). "Diversity and distribution of Victoria Land biota." *Soil Biology and Biochemistry* **38**: 3003-3018.
2. Adams, M. D., S. E. Celniker, *et al*. (2000). "The genome sequence of *Drosophila melanogaster*." *Science* **287**: 2185–2195.
3. Aislabie J.M., K. Chhour, *et al*. (2006). "Dominant bacteria in soils of Marble Point and Wright Valley, Victoria Land, Antarctica." *Soil Biology and Biochemistry* **38**: 3041- 3056.
4. Akoh, C., G. Lee, *et al*. (2004). "GDSL family of serine esterases/lipases." *Progress in Lipid Research* **43**: 534-552.
5. Allen, M. A., F. M. Lauro, *et al*. (2009). "The genome sequence of the psychrophilic archaeon, *Methanococcoides burtonii*: the role of genome evolution in cold adaptation." *ISME Journal* **3**: 1012–1035.
6. Altschul S. F., T. L. Madden, *et al*. (1997). "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs." *Nucleic Acids Research* **25**: 3389-3402
7. Amann, R. I., W. Ludwig and K. H. Schleifer (1995). "Phylogenetic identification and in situ detection of individual microbial cells without cultivation." *Microbiological Reviews* **59**: 143-169.
8. Anderson, D. E. (2008). "Gene discovery in Antarctic Dry Valley soils." *Thesis, (MSc).* University of the Western Cape.
9. Angelova, M., S. Kalajdziski and L. Kocarev (2010). "Computational methods for gene finding in prokaryotes." *ICT innovations*, Web Proceedings (Editor: M. Gusev).
10. Angkawidjaja, C. and S. Kanaya (2006). "Family I.3 lipase: bacterial lipases secreted by the type I secretion system." *Cellular and Molecular Life Sciences* **63**: 2804-2817.
11. Aravind, L. and E. V. Koonin (1998). "Phosphoesterase domains associated with DNA polymerases of diverse origin." *Nucleic Acids Research* **26**: 3746-3752.
12. Arpigny, J. L. and K. E. Jaeger (1999). "Bacterial lipolytic enzymes: Classification and properties." *Biochemical Journal* **343**: 177-183.
13. Atomi, H. (2004). "Recent progress towards the application of hyperthermophiles and their enzymes." *Current Opinion in Chemical Biology* **9**: 166-173.
14. Aurilia, V., A. Parracino and S. D'Auria (2008). "Microbial carbohydrate esterases in cold adapted environments." *Gene* **410**: 234-240.
15. Ayala-del-Rio, H. L., P. S. Chain, *et al*. (2010). "The genome sequence of *Psychrobacter arcticus* 273-4, a psychroactive Siberian permafrost bacterium, reveals mechanisms for adaptation to low temperature growth." *Applied and Environmental Microbiology* **76**: 2304-2312.
16. Bakermans, C., S. L. Tollaksen, *et al*. (2007). "Proteomic analysis of *Psychrobacter cryohalolentis* K5 during growth at subzero temperatures." *Extremophiles* **11**: 343-354.
17. Balks, M. and I. Campbell (2001). *Ross Sea region 2001: A state of the environment report for the Ross Sea region of Antarctic.* Christchurch, New Zealand Antarctic Institute.
18. Bansal, A. K. (2005). "Bioinformatics in microbial biotechnology- a mini review." *Microbial Cell Factories* **4**: 19.

19. Barwick, R. E. and R. W. Balham (1967). "Mummified seal carcasses in a deglaciated region of South Victoria land, Antarctica." *Tuatara* **15**: 165-180.
20. Battaglia, M., Y. Olvera-Carrillo, *et al*. (2008). "The enigmatic LEA proteins and other hydrophilins." *Plant Physiology* **148**: 6-24.
21. Bennett, S. T., C. Barnes, *et al*. (2005). "Toward the 1,000 dollars human genome." *Pharmacogenomics* **6**: 373–382.
22. Berjak, P. (2006). "Unifying perspectives of some mechanisms basic to desiccation tolerance across life forms." *Seed Science Research* **16**: 1-15.
23. Bertin, P. N., C. Médigue and P. Normand (2008). "Advances in environmental genomics: towards an integrated view of micro-organisms and ecosystems." *Microbiology* **154**: 347-359.
24. Besemer, J. and M. Borodovsky (2005). "GeneMark: web software for gene finding in prokaryotes, eukaryotes and viruses." *Nucleic Acids Research* **33**: W451-W454.
25. Bies-Etheve, N., P. Gauber-Comella, *et al*. (2008). "Inventory, evolution and expression profiling diversity of the LEA protein gene family in *Arabidopsis thaliana*." *Plant Molecular Biology* **67**: 107-124.
26. Billi, D and M. Potts (2002). "Life and death of dried prokaryotes." *Research in Microbiology* **153**: 7-12.
27. Blattner, F. R., G. Plunkett, *et al*. (1997). "The complete genome sequence of *Escherichia coli* K-12." *Science* **277**: 1453–1474.
28. Bolanos-Garcia, V. M. and O. R. Davies (2006). "Structural analysis and classification of native proteins from *E. coli* commonly co-purified by immobilised metal ion chromatography." *Biochimica et Biophysica Acta* **1760**: 1304-1313.
29. Bork, P. (2000). "Powers and pitfalls in sequence analysis: the 70 % hurdle." *Genome Research* **10**: 398-400.
30. Bornscheuer, U. T. (2002). "Microbial carboxyl esterases: classification, properties and application in biocatalysis." *FEMS Microbiology Reviews* **26**: 73-81.
31. Bradford, M. M. (1976). "A Rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding." *Analytical Biochemistry* **72**: 248-254.
32. Bray, E. A. (1993). "Molecular responses to water deficit." *Plat Physiology* **103**: 1035-1040.
33. Bult, C. J., O. White, *et al*. (1996). "Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii*." *Science* **273**: 1058-1073.
34. Bunterngsock, B., P. Kanokratana, *et al*. (2010). "Identification and characterisation of lipolytic enzymes from a peat-swamp forest soil metagenome." *Bioscience Biotechnology and Biochemistry* **74**: 1848-1854.
35. Cacace, G., M. F. Mazzeo, *et al*. (2010). "Proteomics for the elucidation of cold adaptation mechanisms in *Listeria monocytogenes*." *Journal of Proteomics:* **DOI 10.1016/jprot.2010.06.011**
36. Casanueva, A., M. Tuffin, *et al*. (2010). "Molecular adaptations to psychrophily: the impact of 'omic' technologies." *Trends in Microbiology* **18**: 374-381.
37. Cascales, E. and P. J. Christie (2003). "The versatile bacterial type IV secretion systems." *Nature Reviews Microbiology* **1**: 137–149.
38. Cavicchioli, R., K. Siddiqui, *et al*. (2002). "Low temperature extremophiles and their applications." *Current Opinion in Biotechnology* **13**: 253-261.

39. Cavicchioli, R. (2006). “Cold-adapted archaea.” Nature Reveiws Microbiology 4: 331-343.
40. Chattopadhyay, M. K. (2006). "Mechanism of bacterial adaptation to low temperature." *Journal of Biosciences* **31**: 157-165.
41. Chen, C. K., G. C. Lee, *et al*. (2009). “Structure of the alkalohyperthermophilic *Archaeoglobus fulgidus* lipase contains a unique C-terminal domain essential for long-chain substrate binding.” *Journal of Molecular Biology* **390**: 672-685.
42. Chintalapati, S., M. D. Kiran and S. Shivaji (2004). “Role of membrane lipid fatty acids in cold adaptation.” *Cellular and Molecular Biology* **50**: 631-642.
43. Chistoserdova, L. (2010). “Recent progress and new challenges in metagenomics for biotechnology.” *Biotechnology Letters* **32**: 1351-1359.
44. Choo, D., T. Kurihara, *et al*. (1998). "A cold-adapted lipase of an Alaskan psychrotroph, *Pseudomonas* sp. strain B11-1: gene cloning and enzyme purification and characterisation." *Applied and Environmental Microbiology* **64**: 486-491.
45. Churchill., G. A. and M. S. Waterman (1992). “The accuracy of DNA sequences: estimating sequence quality.” *Genomics* **14**: 89-98.
46. Ciccarelli, F. D. and P. Bork (2005). “The WHy domain mediates the response to desiccation in plants and bacteria.” *Bioinformation* **21**: 1304-1307.
47. Cowan, D. A., A. Arslanoglu, *et al*. (2004). "Metagenomics, gene discovery and the ideal biocatalyst." *Biochemical Society Transactions* **32**: 298-302.
48. Csonka, L. N. (1989). “Physiological and genetic responses of bacteria to osmotic stress.” *Microbiological Reviews* **53**: 121-147.
49. D'Amico, S., P. Claverie, *et al*. (2002). "Molecular basis of cold adaptation." *Philosophical Transaction of the Royal Society London B* **357**: 917-925.
50. D'Amico, S., T. Collins, *et al*. (2006). “Psychrophilic microorganisms: challenges for life.” *EMBO Reports* **7**: 385-389.
51. Daniel, R. (2005). "The metagenomics of soil." *Nature Reviews Microbiology* **3**: 470- 478.
52. Darwin, A. J. (2005). “The phage-shock-protein response.” *Molecular Microbiology* **57**: 621-628.
53. De Simone, G., V. Menchise, *et al*. (2001). “The crystal structure of a hyperthermophilic carboxylesterase from the archaeon *Archaeoglobus fulgidus*.” *Journal of Molecular Biology* **314**: 507-518.
54. De Vries, A. L., S. K. Komatsu, *et al*. (1970). "Chemical and physical properties of freezing point-depressing glycoproteins from Antarctic fishes." *Journal of Biological Chemistry* **245**: 2901-2908.
55. Deming. J. W. (2002). “Psychrophiles and polar regions.” *Current Opinion in Microbiology* **5**: 301-309.
56. Dohm, J. C., C. Lottaz, *et al*. (2008). “Substantial biases in ultra-short read data sets from high-throughput DNA sequencing.” *Nucleic Acids Research* **36**:e105.
57. Dort, W. J. (1982). "The mummified seals of Southern Victoria Land, Antarctica." *Antarctic Research Series* **30**: 123-154.
58. Dunaeva, M. and I. Adamska (2001). “Identity of genes expressed in response to light stress in leaves of *Arabidopsis thaliana* using RNA differential display.” *European Journal of Biochemistry* **268**: 5521-5529.
59. Duplantis, B. N., M. Osusky, *et al*. (2010). “Essential genes from Arctic bacteria used to construct stable, temperature-sensitive bacterial vaccines.” *PNAS*: **DOI 10.1073/pnas.1004119107.**

60. Dure, L. III (1993). “Structural motifs in LEA proteins. (In) Plant responses to cellular dehydration during environmental stress.” The American Society of Plant Physiologists, Rockville, MD, USA. (Editors: T. J. Close and E. A. Bray): 91-103.
61. Ejima, K. J. Liu, *et al*. (2004). “Molecular cloning and characterization of a thermostable carboxylesterase from an Archaeon, *Sulfolobus shibatae* DSM5389: non-linear kinetic behavior of a hormone-sensitive lipase family enzyme.” *Journal of Bioscience and Bioengineereing* **98**: 445-451.
62. Eland, C., C. Schmeisser, *et al.* (2006). “Isolation and biochemical characterisation of two novel metagenome-derived esterases.” *Applied and Environmental Microbiology* **72**: 3637-3645.
63. Emanuelsson, O., S. Brunak, *et al*. (2007). “Locating proteins in the cell using TargetP, SignalP, and related tools.” *Nature Protocols* **2**: 953-971.
64. Feller, G. (2003). “Molecular adaptations to cold in psychrophilic enzymes.” *Cellular and Molecular Life Sciences* **60:** 648-662.
65. Feller, G. and C. Gerday (2003). “Psychrophilic enzymes: hot topics in cold adaptation.” *Nature Reviews Microbiology* **1**: 200-208.
66. Feller, G. (2010). “Protein stability and enzyme activity at extreme biological temperatures.” *Journal of Physics. Condensed matter* **22**: 323101.
67. Ferrer, M., V. Olga, *et al.* (2005). "Novel hydrolase diversity retrieved from a metagenome library of bovine rumen microflora." *Environmental Microbiology* **7**: 1996-2010.
68. Finn, R. D., J. Tate, *et al*. (2008). "The Pfam protein families database." *Nucleic Acids Research* **36** (Database Issue): D281-D288.
69. Fleischmann, R. D., M. D. Adams, *et al*. (1995). “Whole-genome random sequencing and assembly of *Haemophilus influenza*” *Science* **269**: 496–512.
70. Fojan, P., P. H. Jonson, *et al*. (2000). "What distinguishes an esterase from a lipase: A novel structural approach." *Biochimie* **82**: 1033-1041.
71. Fraser, C. M., J. D. Gocayne, *et al*. (1995). “The minimal gene complement of *Mycoplasma genitalium*.” *Science* **270**: 397–403
72. Gandhi, N. (1997). "Applications of lipase." *Journal of the American Oil Chemists' Society* **74**: 621-634.
73. Garay-Arroya, A., J. M. Colmenero-Flores, *et al*. (2000). “Highly hydrophobic proteins in prokaryotes and eukaryotes are common during conditions of water deficit.” *Journal of Biological Chemistry* **275**: 5668-5674.
74. Georlette, D., V. Blaise, *et al*. (2004) “Some like it cold: biocatalysis at low temperatures.” *FEMS Microbiology Reviews* **28**: 25-42.
75. Gerday, C., M. Aittaleb, *et al*. (1997). "Psychrophilic enzymes: a thermodynamic challenge." *Biochimica et Biophysica Acta (BBA) - Protein Structure and Molecular Enzymology* **1342**: 119-131.
76. Gerday, C., M. Aittaleb, *et al*. (2000). "Cold-adapted enzymes: from fundamentals to biotechnology." *Trends in Biotechnology* **18**: 103-107.
77. Gianese, G., P. Argos and S. Pascarella (2001). “Structural adaptation of enzymes to low temperatures.” *Protein Engineering.* **14**: 141-148.
78. Gianese, G., F. Bossa and S. Pascarella (2002). “Comparative structural analysis of psychrophilic and meso- and thermophilic enzymes.” *Proteins: Structure, Function, and Bioinformatics* 47: 236–249.
79. Gilham, D. and R. Lehner (2005). "Techniques to measure lipase and esterase activity *in vitro*." *Methods* **36**: 139-147.

80. Gill, R. T., M. P. DeLisa, *et al*. (2000). "OmpT expression and activity increase in response to recombinant chloramphenicol acyltransferase overexpression and heat shock in *E. coli*." *Journal of Molecular Microbiology and Biotechnology* **2**: 283-289
81. Goldstein, R. A. (2007). "Amino-acid interactions in psychrophiles, mesophiles, thermophiles and hyperthermophiles: insights from the quasi-chemical approximation." *Protein science* **16**: 1887-1895.
82. Gomes, J. and W. Steiner (2004). "The biocatalytic potential of extremophiles and extremozymes." *Food technology and Biotechnology* **42**: 223-235.
83. Goodchild, A., M. Raftery, *et al*. (2005). "Cold adaptation of the Antarctic archaeon, *Methanococcoides burtonii* assessed by proteomics using ICAT." *Journal of Proteome Research* **4**: 473–480.
84. Goyal, K., L. J. Walton and A. Tunnacliffe (2005). "LEA poteins prevent protein aggregation due to water stress." *Biochemical Journal* **388**: 151-157.
85. Grzymski, J. J., B. J. Carter, *et al*. (2006). "Comparative genomics of DNA fragments from six Antarctic marine planktonic bacteria." *Applied and Environmental Microbiology* **72**: 1532-1541.
86. Gualerzi., C. O., A. M. Giuliodori and C. L. Pon (2003). "Transcriptional and post-transcriptional control of cold shock genes." *Journal of Molecular Biology* **331**: 527-539.
87. Guigó, R., P. Agarwal, *et al*. (2000). "An assessment of gene prediction accuracy in large DNA sequences." *Genome Research* **10**: 1631-1642.
88. Guina, T., W. Manhong, *et al* (2003). "Proteomic analysis of *Pseudomonas aeruginosa* grown under magnesium limitation**"** *Journal of American Society for Mass Spectrometry* **14**:742–751.
89. Gupta, N., P. Rathi and R. Gupta (2002). "Simplified *para*-nitrophenyl palmitate assay for lipases and esterases." *Analytical Biochemistry* **311**: 98-99.
90. Gupta, R., N. Gupta and P. Rathi (2004). "Bacterial lipases: an overview of production, purification and biochemical properties." *Applied Microbiology and Biotechnology* **64**: 763–781.
91. Handelsman, J. (2004). "Metagenomics: application of genomics to uncultured microorganisms." *Microbial Molecular Biology Reviews* **68**: 669-685.
92. Hara, M. (2010). "The multifunctionality of dehydrins." *Plant Signalling and Behaviour* **5**: 503-508.
93. Hårdeman, F. and S. Sjöling (2007). "Metagenomic approach for the isolation of a novel low-temperature-active lipase from uncultured bacteria of marine sediment." *FEMS Microbiology Ecology* **59**: 524-534.
94. Harrismendy, O. P. C. Ng, *et al*. (2009). "Evaluation of next generation sequencing platforms for population targeted sequencing studies." *Genome Biology* **10**: R32.
95. Hasan, F., A. A. Shah, *et al*. (2006). "Industrial applications of microbial lipases." *Enzyme and Microbial Technology* **39**: 235-251.
96. Haslam, R., S. Rust, *et al*. (2002). "Cloning and characterisation of S-formylglutathione hydrolase from *Arabidopsis thaliana*: a pathway for formaldehyde detoxification." *Plant Physiology and Biochemistry* **40**: 281-288.
97. Hazes, B. and L. Frost (2008). "Towards a systems biology approach to study type II/IV secretion systems." *Biochimica et Biophysica Acta* **1778**: 1839-1850.

98. Hebraud, M. and P. Poitier (1999). "Cold shock response and low temperature adaptation in psychrotrophic bacteria." *Journal of Molecular Microbiology and Biotechnology* **1**: 211-219.
99. Henderson, B. and A. G. Pockley (2010). "Molecular chaperones and protein folding catalysts as intercellular signalling regulators in immunity and inflammation." *Journal of Leukocyte Biology* **88**: 1-17.
100. Henderson, I. R., F. Navarro-Garcia, *et al*. (2004). "Type V Protein Secretion Pathway: the Autotransporter Story." *Microbiology and Molecular Biology Reviews* **68**: 692-744.
101. Hengen, P. N. (1996). "Expression profiling using messenger RNA assays." *TIBS* **21**: 492-493.
102. Henke, E., U. T. Bornscheuer, *et al*. (2003). "A molecular mechanism of enantiorecognition of tertiary alcohols by carboxylesterases." *ChemBioChem* **6**: 485-493.
103. Hernandez, D., P. Francois, *et al*. (2008). "De novo bacterial genome sequencing: millions of very short reads assembled on a desktop computer." *Genome Research*. **18**: 80209.
104. Hoff, K. J., M. Tech, *et al*. (2008). "Gene prediction in metagenomic fragments: large-scale machine learning approach." *BMC bioinformatics* **9**: 217.
105. Hogg, I. D., C. Cary, *et al*. (2006). "Biotic interactions in Antarctic terrestrial ecosystems: Are they a factor?" *Soil Biology and Biochemistry* **38**: 3035-3040.
106. Horn, G., R. Hofweber, *et al*. (2007). "Structure and function of bacterial cold shock proteins." *Cellular and Molecular Life Sciences* **64**: 1457-1470.
107. Hotta, Y. S. Ezaki, *et al* (2002). "Extremely stable and versatile carboxylesterase from a hyperthermophilic archaeon." *Applied and Environmental Microbiology* **68**: 3925-3931.
108. Huang, X. C., M. A. Quesada and R. A. Mathies (1992). "DNA sequencing using capillary array electrophoresis." *Analytical Chemistry* 64: 2149–2154.
109. Hundertmark, M., R. Dimova, *et al*. (2010). "The intrinsically disordered Late Embryogenesis Abundant protein, LEA18 from *Arabidopsis thaliana* modulates membrane stability through binding and folding." *Biochimica et Biophysica Acta*: **DOI 10.1016/j.bbamem.2010.09.010**.
110. Hunter-Cevera, J. (1998). "The value of microbial diversity." *Current Opinion in Microbiology* **1**: 278–285.
111. Huson, D. H., D. C. Richter, *et al*. (2009). "Methods for comparative metagenomics." *BCM Bioinformation* **10**:S12.
112. Hutchison III, C. A. (2007). "DNA sequencing: bench to bedside and beyond." *Nucleic Acids Research* **35**: 6227–6237.
113. Ikai, A. J. (1980). "Thermostability and aliphatic index of globular proteins." *Journal of Biochemistry* **88**: 1895-1898.
114. Inouye, S. and M. Inouye (1996). "Structure, function and evolution of bacterial reverse transcriptase." *Virus genes* **11**: 81-94.
115. Jacob-Dubuisson F., R. Fernandez, *et al*. (2004). "Protein secretion through autotransporter and two-partner pathways." *Biochimica et Biophysica Acta* **1694:** 235-57.
116. Jaeger, K. E., S. Ransac, *et al*. (1994). "Bacterial Lipases." *FEMS Microbiology Reviews* **15**: 29-63.
117. Jaeger, K. E., B. Schneidinger, *et al*. (1997). "Bacterial lipases for biotechnological applications." *Journal of Molecular Catalysis - B Enzymatic* **3**: 3-12.

118. Jaeger, K. E., B. W. Dijkstra, *et al*. (1999). Bacterial biocatalysts: Molecular biology, three-dimensional structures, and biotechnological applications of lipases. *Annual Review of Microbiology* **53:** 315-351.
119. Jaeger, K. E. and T. Eggert (2002). "Lipases for biotechnology." *Current Opinion in Biotechnology* **13**: 390-397.
120. Jahandideh, S., P. Abdolmaleki, *et al*. (2007). "Sequence and structural parameters enhancing adaptation of proteins to low temperatures." *Journal of Theoretical Biology* **246**: 159-166.
121. Jahandideh, M., S. M. Barkooie, *et al*. (2008). "Elucidating the protein cold-adaptation: Investigation of the parameters enhancing protein psychrophilicity." *Journal of Theoretical Biology* **255**: 113-118.
122. Jeon, J. H., J-T. Kim, *et al*. (2009). "Characterisation and its potential application of the esterases derived from the arctic sediment metagenome." *Mar-Biotechnology* **11**: 307-316.
123. Jones, D. T., W. R. Taylor and J. M. Thornton. (1992). "The rapid generation of mutation data matrices from protein sequences." *Computer Applications in the Biosciences* **8**: 275–282.
124. Joseph, B., P. W. Ramteke, *et al*. (2007). "Standard review cold-active lipases: a versitile tool for industrial applications." *Biotechnology and Molecular Biology Review* **2**: 39-48.
125. Joseph, B., P. W. Ramteke, *et al*. (2008). "Cold active microbial lipases: Some hot issues and recent developments." *Biotechnology Advances* **26**: 457-470.
126. JunGang, L., Z. KeGui and H. WenJun (2010). "Cloning and biochemical characterisation of a novel lipolytic gene from activated sludge metagenome, and its gene product." *Microbial Cell Factories* **9**: 83.
127. Junker, A. S., H. Willenbrock, *et al*. (2003). "Prediction of lipoprotein signal peptides in gram negative bacteria." *Protein Science* **12**: 1652-1662.
128. Kaczanowska, M. and M. Rydén-Aulin (2007) "Ribosome biogenesis and the translation process in Escherichia coli." *Microbiology and Molecular Biology Reviews* **71**: 477-794.
129. Kakugawa, S., S. Fushinobu, *et al*. (2007). "Characterisation of a thermostable carboxylesterase from the hyperthermophilic bacterium *Thermotoga maritime*." *Applied Microbiology and Biotechnology* **74**: 585-591.
130. Kane, F. J., B. N. Violand and D. F. Curran (1993). "Novel in-frame two codon translational hop during synthesis of bovine placental lactogen in recombinant strain of *E. coli*." *Nucleic Acids Research* **20**: 6707-6712.
131. Kane, J. F. (1995). "Effects of rare codon clusters on high-level expression of heterologous proteins in *Escherichia coli*." Current Opinions in Biotechnology **6**: 494–500.
132. Kang, C., Oh, K., *et al*. (2011). "A novel family VII esterase with industrial potential from compost metagenomic library." *Microbial Cell Factories* **10**: 41-49.
133. Karpinets, T. V., A. Y. Obraztsova, *et al*. (2009). "Conserved synteny at the protein family level reveals genes underlying *Shewanella* species' cold tolerance and predicts their novel phenotypes." *Functional Integrated Genomics*: **DOI 10.1007/s10142-009-0142-y.**
134. Karzai, A. W., M. M. Susskind and R. T. Sauer (1999). "SmpB, a unique RNA-binding protein essential for the peptide-tagging activity of SsrA (tmRNA)." *EMBO Journal* **18**: 3793-3799.

135. Kawamoto, J., T. Kurihara., *et al*. (2007). "Proteomic studies of an Antarctic cold-adapted bacterium, *Shewanella livingstonensis* Ac10, for global identification of cold-inducible proteins." *Extremophiles* 11: 819-826.
136. Kawaoka, J. and A. M. Pyle (2005). "Choosing between DNA and RNA: the polymer specificity of RNA helicase NPH-II." *Nucleic Acids Research* **33**: 644-649.
137. Kempf, B and E. Bremer (1998). "Uptake and synthesis of compatible solutes as microbial stress responses to high osmolality environments." *Microbiological Reviews* **70**: 319-330.
138. Kim, H., S. Ryou, *et al*. (2009). "Genetic analysis of the invariant residue G791 in *E. coli* 16S rRNA implicates RelA in ribosome function." *Journal of Bacteriology* **191**: 2042-2050.
139. Kim, J-N., M. J. Seo, *et al*. (2005). "Screening and characterisation of an esterase from a metagenomic library." *Journal of Microbial Biotechnology* **15**: 1067-1072.
140. Kim, K. K., H. K. Song, *et al*. (1997). "The crystal structure of a triacylglycerol lipase from *Pseudomonas cepacia* reveals a highly open conformation in the absence of a bound inhibitor." *Structure* **5**: 173-185.
141. Koag, M. C., R. D. Fenton, *et al*. (2003). "The binding of maize DHN1 to lipid vesicles: gain of structure and lipid specificity." *Plant Physiology* **131**: 309-316.
142. Konstantinidis, K. T., J. Braff., *et al*. (2009). "Comparative metagenomic analysis of a microbial community residing at a depth of 4,000 meters at station ALOHA in the North Pacific subtropical gyre." *Applied and Environmental Microbiology* **75**: 5345-5355.
143. Koonin, E. V. (1994). "Conserved sequence pattern in a wide variety of phosphoesterases." *Protein Science* **3**:356-358.
144. Koonin, E. V. and M. Y. Galperin (2003). Sequence - evolution - function: computational approaches in comparative genomics. Kluwer Academics, Boston. ISBN-10: 1-40207-274-0
145. Koonin, E. V. and Y. I. Wolf (2008). "Genomics of bacteria and archaea: the emerging dynamic view of the prokaryotic world." *Nucleic Acids Research* **36**: 6688-6719.
146. Kovacs, D., E. Kalmar, *et al*. (2008). "Chaperone activity of ERD10 and ERD14, two disordered stress-related plant proteins." *Plant Physiology* **147**: 381-390.
147. Kovacs, D., M. Rakacs, *et al*. (2009). "Janus chaperones: assistance of both RNA– and protein- folding by ribosomal proteins." *FEBS Letters* **583**: 88-92.
148. Krause, L., N. N. Diaz, *et al* (2006). "Finding novel genes in bacterial communities isolated from the environment." *Bioinformatics* **22**: e281-e289.
149. Krísko, A., Z. Smole, *et al*. (2010). "Unstructured hydrophilic sequences in prokaryotic proteomes correlate with desiccation tolerance and host association." *Journal of Molecular Biology* **402**: 775- 782.
150. Kulakova, L., A. Galkin, *et al*. (2004). "Cold-active esterase from *Psychrobacter* sp. Ant300: gene cloning, characterization, and the effects of Gly→Pro substitution near the active site on its catalytic activity and stability" *Biochimica et Biophysica Acta* **1696**: 59-65.
151. Kumar, P. S., M. Ghosh, *et al*. (2011). "Cold active enzymes from the marine psychrophiles: biotechnological perspective." *Advanced Biotechnology* **10**: 16-21.

152. Kumar, S., J. Dudley, *et al*. (2008). “MEGA: a biologist-centric software for evolutionary analysis of DNA and protein sequences.” *Briefings in Bioinformatics* **9**: 299-306.
153. Kunst, F., N. Ogasawara, *et al*. (1997). “The complete genome sequence of the gram-positive bacterium *Bacillus subtilis*.” *Nature* **390**: 249–256.
154. Kvint, K., L. Nachin, *et al*. (2003). “The bacterial universal stress protein: function and regulation.” *Current Opinion in Microbiology* **6**: 140-145.
155. Kyte, J and R. Doolittle (1982). “A simple method for displaying the hydropathic character of a protein.” *Journal of Molecular Biology* **157**: 105-132.
156. Lander, E. S., L. M. Linton, *et al*. (2001). “Initial sequencing and analysis of the human genome.” *Nature* **409**: 860–921.
157. Larkin, M. A., G. Blackshields, *et al*. (2007). "ClustalW2 and ClustalX version 2." *Bioinformatics* **23**: 2947-2948.
158. Livernois, A. M., D. J. Hnatchuk, *et al*. (2009). “Obtaining highly purified intrinsically disordered protein by boiling lysis and single step ion exchange.” *Analytical Biochemistry* **392**: 7-76.
159. Lloyd, L. J., S. E. Jones, *et al*. (2004). “Identification of a new member of the phage shock protein response in *Escherichia coli*, the phage shock protein G (PspG).”*Journal of Biological Chemistry* **279**: 55707-55714.
160. Lorenz, P. and C. Schleper (2002). "Metagenome—a challenging source of enzyme discovery.” *Journal of Molecular Catalysis B: Enzymatic* **19-20**: 13-19.
161. Lovell, S. C., I. W. Davis, *et al*. (2002). "Structure validation by Calpha geometry: phi,psi and Cbeta deviation." *Proteins: Structure, Function & Genetics* **50**: 437-450.
162. Macek, B., I. Mijakovica, *et al*. (2007). “The serine/threonine/tyrosine phosphoproteome of the model bacterium *Bacillus subtilis*.” *Molecular and Cellular Proteomics* **6**: 697-707.
163. Madabhushi, R. S. (1998). “Separation of 4-color DNA sequencing extension products in noncovalently coated capillaries using low viscosity polymer solutions.” *Electrophoresis* **19**: 224–230.
164. Mahajan, S. and N. Tuteja (2005). “Cold, salinity and drought stress: an overview.” *Archives of Biochemistry and Biophysics* **444**: 139-158.
165. Maitra, N and J. C. Cushman (1994). “Isolation and characterisation of a drought induced soybean cDNA encoding D95 family Late Embryogenesis Abundant protein.” *Plant Physiology* **106**: 805-806.
166. Manco, G., G. Carrea, *et al*. (2002). “Modification of the enantioselectivity of two homologous thermophilic carboxylesterases from *Alicyclobacillis acidocaldarius* and *Archaeoglobus fulgidus* by random mutagenesis and screening.” *Extremophiles* **6**: 325-331.
167. Mardis, E. R. (2007). “The impact of next-generation sequencing technology on genetics.” *Trends in Genetics* **24**: 133-141.
168. Margulies, M., M. Egholm, *et al*. (2005). “Genome sequencing in microfabricated high-density picolitre reactors.” *Nature* **437**: 376–380.
169. Mathe´, C., M-F. Sagot, *et al*. (2002). “Current methods of gene prediction, their strengths and weaknesses.” *Nucleic Acids Research* **30**: 4103–4117.
170. Mavromatis, K., N. Ivanova, *et al*. (2007). “Use of simulated data sets to evaluate the fidelity of metagenomic processing methods.” *Nature Methods* **4**: 495-500.
171. McGuffin, L. J., K. Bryson, *et al*. (2000). "The PSIPRED protein structure prediction server." *Bioinformatics* **16**: 404-405.

172. Medhekar, B. and F. J. Miller (2007). “Diversity-generating retroelements.” *Current Opinion in Microbiology* **10**: 388-395.
173. Medigue, C., E. Krin, *et al*. (2005). “Coping with cold: the genome of the versatile marine Antarctica bacterium *Pseudoalteromonas haloplanktis* TAC125.” *Genome Research* **15**: 1325-1335.
174. Methé, B. A., K. E. Nelson, *et al* (2005). “The psychophilic lifestyle as revealed by the genome sequence of *Colwellia psychrerythraea* 34H through genomic and proteomic analysis.” *PNAS* **102**: 10913- 10918.
175. Metpally, R. P. and B. V. Reddy (2009). “Comparative proteome analysis of psychrophilic versus mesophillic bacterial species: insights into the molecular basis of cold adaptation of proteins.” *BCM Genomics* **10**.
176. Metzker, M. L. (2010). “Sequencing technologies- the next generation.” *Nature Reviews Genetics* **11**: 31-46
177. Miller, M. L., B. Soufi, *et al*. (2008). “NetPhosBac- a predictor for serine/ threonine phosphorylation sites in bacteria.” *Proteomics* **9**: 116-125.
178. Morgan-Kiss, R. M., J. C. Priscu, *et al*. (2006). "Adaptation and acclimation of photosynthetic microorganisms to permanently cold environment." *Microbiology and Molecular Biology Reviews* **70**: 222-252.
179. Morozova, O. and M. A. Marra (2008). “Applications of next-generation sequencing technologies in functional genomics.” *Genomics* **92**: 255–264
180. Mouillon, J., S. K. Eriksson and P. Harryson (2008). “Mimicking the plat cell interior under water stress by macromolecular crowding: disordered dehydrin proteins are highly resistant to structural collapse.” *Plant Physiology* **148**: 1925-1937.
181. Mowla, S. B., A. Cuypers, *et al*. (2006). “Yeast complementation reveals a role for an *Arabidopsis thaliana* Late Embryogenesis Abundant (LEA) – like protein in oxidative stress tolerance.” *Plant Journal* **48**: 743-756.
182. Mulder, N. J., P. Kersey, *et al*. (2008). “*In silico* characterisation of proteins: Uniprot, InterPro and Integr8.” *Molecular Biotechnology* **38**: 165-177.
183. Müller, S. A., T. Kohajda, *et al*. (2010). “Optomisation of parameters for coverage of low molecular weight proteins.” *Analytical and Bioanalytical Chemistry*: **DOI 10.1007/500216-010-4093**.
184. Nardini, M. and B. W. Dijkstra (1999). "Alpha/beta hydrolase fold enzymes: the family keeps growing.” *Current Opinion in Structural Biology* **9**: 732–737.
185. Nardini, M., D. A. Lang, *et al*. (2000). “The crystal structure of *Pseudomonas aeruginosa* lipase in the open conformation. The prototype for family I.1 of bacterial lipases.” *Journal of Biological Chemistry* **275**: 31219-31225.
186. Nichols, D. S., P. D. Nichols, *et al*. (1993). "Polyunsaturated fatty acids in Antarctic bacteria." *Antarctic Science* **5**: 149-160.
187. Nichols, D. S., J. Bowman, *et al*. (1999). "Developments with Antarctic microorganisms: Culture collections, bioactivity screening, taxonomy, PUFA production and cold-adapted enzymes." *Current Opinion in Biotechnology* **10**: 240-246.
188. Noguchi. H., J. Park and T. Takagi (2006). “MetaGene: prokaryotic gene finding from environmental genome shotgun sequences.” *Nucleic Acids Research* **34**: 5623-5630.
189. Nonin- Lecomte, S., N. germain-Amiot, *et al*. (2009). “Ribosome hijacking: a role for small protein B during *trans*-translation.” *EMBO Reports* **10**: 160-165.

190. Nyström, T. and N. Gustavsson. (1998). “Maintenece energy requirement: what is required for stasis survival of *Escherichia coli*?” *Biochimica et Biophysica Acta* **1365**: 225-231.
191. Okuda, S. and H. Tokuda (2011). “Lipoprotein sorting in bacteria.” *Annual Reviews in Microbiology* **65**: 239-259.
192. Olvera-Carrillo, Y., F. Campos, *et al*. (2010). “Functional analysis of the group 4 Late Embryogenesis Abundant proteins reveals their relevance in the adaptive response during water deficit in *Arabidopsis*”. *Plant Physiology* **154**: 373-390.
193. Oswald, T. and U. Rinas (1996). “Chloramphenicol resistance interferes with purification of histidine-tagged fusion proteins from recombinant *E.coli*.” *Analytical Biochemistry* **236**:357- 358.
194. Panda, T. and B. S. Gowrishankar (2005). “Production and applications of esterases.” *Applied Microbiology and Biotechnology* **67**: 160-169.
195. Peck, L. S., M. S. Clark, *et al*. (2005). "Genomics: Applications to Antarctic ecosystems." *Polar Biology* **28**: 351-365.
196. Petterson, E., J. Lunderberg and A. Ahmadian (2009). “Generations of sequencing technologies.” *Genomics* **93**: 105-111.
197. Pevzner, P. A., H. Tang and M. S. Waterman (2001). “An Eulerian path approach to DNA fragment assembly.” *PNAS* **98**: 9748-9753.
198. Piette, F., C. Struvay and G. Feller (2011). “The protein folding challenge in psychrophiles: facts and current issues.” *Environmental Microbiology*: **DOI 10.1111/j.1462-2920.2011.02436.x.**
199. Piette, F., S. D’Amico, *et al*. (2011). “Life in the cold: a proteomic study of cold-repressed proteins in the Antarctic bacterium *Pseudoalteromonas haloplanktis* TAC125.” *Applied Environmental Microbiology*: **DOI 10.1128/AEM.02757-10.**
200. Pop, M. (2009). “Genome assembly reborn: recent computational challenges.” *Briefings in Bioinformatics* **10**: 354-366.
201. Potts, M. (1994). “Desiccation tolerance of prokaryotes.” *Microbiology Reviews* **58**: 755-805.
202. Potts, M., S. M. Slaughter, *et al*. (2005). “Desiccation tolerance of prokaryotes: application of principles to human cells.” *Integrative and Comparative Biology* **45**: 800-809.
203. Privalov, P. L. (1990). “Cold denaturation of proteins.” *Critical Reviews in Biochemistry and Molecular Biology* **25**: 281-306.
204. Privalov, P. L. and G. I. Makhatadze (1990). “Heat capacity of protein: II. partial molar heat capacity of the unfolded polypeptide chaino f proteins: protein unfolding effects.” *Journal of Molecular Biology* **213**: 385-391.
205. Prober, J. M., G. L. Trainor, *et al*. (1987). “A system for rapid DNA sequencing with fluorescent chain terminating dideoxynucleotides.” *Science* **238**: 336–341.
206. Qiu, Y., S. Kathariou and D. M. Lubman (2006). “Proteomic analysis of cold adaptation in a Siberian permafrost bacterium – *Exiguobacterium sibiricum* 255-15 by two-dimensional liquid separation coupled with mass spectrometry.” *Proteomics* **6**: 5221–5233.
207. Raes, J., K. U. Foerstner and P. Bork (2007). “Get the most out of your metagenome: computational analysis of environmental sequence data.” *Current Opinions in Microbiology* **10**: 1-9.
208. Ramagopal, S. and B. D. Davis (1974). “Localisation of the stringent protein of *E. coli* on the 50S ribosomal subunit.” *PNAS* **71**: 820-824.

209. Ranjard, L. and A. Richaume (2001). “Quantitative and qualitative microscale distribution of bacteria in soil.” *Research in Microbiology* 152: 707-716.
210. Rashamuse, K. J., Magomani, V, *et al*. (2009). “A novel family VIII carboxylesterase derived from a leachate metagenome library exhibits promiscuous betalactamase activity on nitrocefin.” *Applied Microbiology and Biotechnology* **83**: 491-500.
211. Ray, M. K., G. S. Kumar, *et al*. (1998). "Adaptation to low temperature and regulation of gene expression in Antarctic psychrotrophic bacteria." *Journal of Bioscience* **23**: 423-435.
212. Reva, O. N., C. Weinel, *et al*. (2006). “Functional genomics of stress response in *Pseudomonas putida* KT2440.” *Journal of Bacteriology* **188**: 4079-4092.
213. Reyes, J. L., M. Rodrigo, *et al*. (2005). “Hydrophilins from distant organisms can protect enzymatic activities from water limitation effects in vitro.” *Plant, Cell and Environment* **28**: 709-718.
214. Reyes, J. L., F. Campos, *et al*. (2008). “Functional dissection of hydrophilins during *in vitro* freeze protection.” *Plant, Cell and Environment* **31**: 1781-1790.
215. Rhee, J-K., D-G Ahn, *et al*. (2005). “New thermophilic and thermostable esterase with sequence similarity to the hormone-sensitive lipase family, cloned from a metagenomic library.” *Applied and Environmental Microbiology* **71**: 817-825.
216. Riesenfeld, C. S., P. D. Schloss and J. Handelsman (2004). “Metagenomics: genomic analysis of microbial communities.” *Annual Reviews in Genetics* **38**: 525-552.
217. Rodrigues, D. F. and J. M. Tiedje (2008). "Coping with our cold planet." *Applied and Environmental Microbiology* **74**: 1677-1686.
218. Roelofs, D., M. G. M. Aarts, *et al*. (2008). “Functional ecological genomics to demonstrate general and specific responses to abiotic stress.” *Functional Ecology* **22**: 8-18.
219. Roh, C and F. Villatte (2008). “Isolation of a low temperature adapted lipolytic enzyme from uncultivated microorganism.” *Journal of Applied Microbiology* **105**: 116-123.
220. Rosenau, F. and K. E. Jaeger (2000). "Bacterial lipases from *Pseudomonas*: Regulation of gene expression and mechanisms of secretion." *Biochimie* **82**: 1023-1032.
221. Rouchka, E. C. and I. E. Cha (2009). “Current trends in pseudogene detection and characterisation.” *Current Bioinformatic*s **4**: 112-119.
222. Rusch, D. B., A. L. Halpern, *et al*. (2007). “The Sorcerer II Global Ocean Sampling expedition: northwest Atlantic through eastern tropical pacific.” *PLoS. Biology* **5**: e77.
223. Rusnak, M., J. Nieveler, *et al*. (2005). “The putative lipase AF1763, from *Archaeoglobus fulgidus* is a carboxylesterase with very high pH optimum.” *Biotechnology Letters* **27**: 743-748.
224. Russell, N. J. (1990). “Cold adaptation of microorganisms.” *Philosophical Transactions of the Royal Society London* **326**: 595-611.
225. Russell, N. (2000). "Towards a molecular understanding of cold activity of enzymes from psychrophiles." *Extremophiles* **4**: 83-90.
226. Salzberg, S. L., A. L. Delcher, *et al*. (1998). “Microbial gene identification using interpolated Markov models.” *Nucleic Acids Research* **26**: 544-548.
227. Sangeetha, R., I. Arulpandi, and A. Geetha (2011). “Bacterial lipases as potential industrial biocatalysts: an overview.” *Research Journal of Microbiology* **6**: 1-24.

228. Sanger, F. and A. R. Coulson (1975). "A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase." *Journal of Molecular Biology* **94**: 441–448.
229. Sanger, F., S. Nicklen and A. R. Coulson (1977). "DNA sequencing with chain-terminating inhibitors." *PNAS* **74**: 5463–5467.
230. Sanger, F., A. R Coulson, *et al*. (1982). "Nucleotide sequence of bacteriophage lambda DNA." *Journal of Molecular Biology* **162**: 729–773.
231. Sasakawa, H., E. Sakata *et al*. (2007). "Ultrahigh field NMR studies of antibody binding and site specific phosphorylation of alpha synuclein." *Biochemical and Biophysical Research Communications* **363**: 795-799.
232. Saunders, N. F., A. Goodchild, *et al*. (2005). "Predicted roles for hypothetical proteins in the low-temperature expressed proteome of the Antarctic archaeon *Methanococcoides burtonii*." *Journal of Proteomic Research* **4**: 464-472.
233. Schmeisser, C., H. Steele, *et al*. (2007). "Metagenomics, biotechnology with nonculturable microbes." *Applied Microbiology and Biotechnology* **75**: 955-962.
234. Schwede, T., J. Kopp, *et al*. (2003). "SWISS-MODEL: an automated protein homology-modeling server." *Nucleic Acids Research* **31**: 3381-3385.
235. Sharma, R., Y. Chisti and U. C. Banerjee (2001). "Production, purification, characterisation, and applications of lipases." *Biotechnology Advances* **19**: 627-662.
236. Sharma, R., S. K. Soni, *et al.* (2002). "Purification and characterisation of a thermostable alkaline lipase from a new thermophilic *Bacillus* sp. RSJ-1." *Process Biochemistry* **37**: 1075-1084.
237. Shih, M., F. A. Hoekstra and T. C. Hsing (2008). "Late Embryogenesis Abundant proteins." *Advances in Botanical Research* **48**: 212-240.
238. Shravage, B. V., K. M. Dayananda, *et al*. (2007). "Molecular microbial diversity of a soil sample and detection of ammonia oxidizers from Cape Evans, McMurdo Dry Valley, Antarctica." *Microbiological Research* **162**: 15-25.
239. Siest, G., C. Courtay, *et al*. (1992). "Gamma-glutamyltransferase: nucleotide sequence of the human pancreatic cDNA. Evidence for a ubiquitous gamma-glutamyltransferase polypeptide in human tissues." *Biochemical Pharmacology* **43**: 2527–2533.
240. Simon, C. and R. Daniel (2009). "Achievemnets and new knowledge unravled by metagenomic approaches." *Applied Microbiology and Biotechnology* **85**: 265-276.
241. Simon, C. and R. Daniel (2011). "Metagenomic analysis: past and future trends." *Applied and Environmental Microbiology* **77**: 1153-1161.
242. Simon, D. M. and S. Zimmerly (2008). "A diversity of uncharacterised reverse transcriptases in bacteria." *Nucleic Acids Research* **36**: 7219-7229.
243. Singh, J., A. Behal., *et al*. (2009). "Metagenomics: concept, methodology, ecological inference and recent advances." *Biotechnology Journal* 4: 480-494.
244. Singh, S., C. C. Cornilescue, *et al*. (2005). "Solution structure of a Late Embryogenesis Abundant protein (LEA14) from *Arabidopsis thaliana*, a cellular stress related protein." *Protein Science* **14**: 2601-2609.
245. Sivashankari, S. and P. Shanmughavel (2006). "Functional annotation of hypothetical proteins – a review." *Bioinformation* **1**: 335- 338.
246. Skolnick, J. and J. S. Fetrow (2000). "From genes to protein structure and function: novel applications of computational approaches in the genomic era." *TibTech* **18**: 34-39.

247. Sleator, R. D. and P. Walsh (2010). “An overview of in silico protein function prediction.” *Archives in Microbiology* **192**: 151-155.
248. Smalås, A. O., H. K. Leiros, *et al.* (2000). “Cold adapted enzymes.” *Biotechnology Annual Reviews* **6**: 1-57.
249. Smith, J. J., L. Ah-Tow, *et al*. (2006). "Bacterial diversity in three different Antarctic cold desert mineral soils." *Microbial Ecology* **51**: 413-421.
250. Smith, L. M., J. Z. Sanders, *et al*. (1986). “Fluorescence detection in automated DNA sequence analysis.” *Nature* **321**: 674–679.
251. Snider, J. and W. A. Houry (2008). "AAA+ proteins: diversity in function, similarity in structure." *Biochemical Society Transactions* **36**: 72–7.
252. Snyder, L. A. S., N. Loman, *et al*. (2009). “Next-generation sequencing – the promise and perils of charting the great microbial unknown.” *Microbial Ecology* **57**: 1-3.
253. Söderberg, M. A., O. Rossier and N. P. Cianciotto (2004). “The Type II protein secretion system of *Legionella pneumophila* promotes growth at low temperatures.” *Journal of Bacteriology* **186**: 3712–3720.
254. Sørensen, H. P. and K. K. Mortensen (2005). “Advanced genetic strategies for recombinant protein expression in *E. coli*.” *Journal of Biotechnology* **115**: 113-128.
255. Soror, S. H., R. Rao and J. Cullum (2009). “Mining the genome sequence for novel enzyme activity: characterisation of an unusual member of the hormone-sensitive lipase family of esterases from the genome of *Streptomyces coelicolor* A3.” *Protein Engineering, Design and Selection* **22**: 333-339.
256. Soufi, B., C. Jers, *et al*. (2008). “Insights from site specific phosphoproteomics in bacteria.” *Biochimica et Biophysica Acta* **1784**: 186-192.
257. Steele, H. L., K-E. Jaeger, *et al*. (2009). “Advances in recovery of novel biocatalysts from Metagenomes.” *Journal of Molecular Microbiology and Biotechnology* **16**: 25-37.
258. Sturgeon, J. A and L. O. Ingram (1978). “Low-temperature conditional cell division mutants of *Escherichia coli*.” *Journal of Bacteriology* **133**: 256-264.
259. Tanaka, S., K. Ikeda and H. Miyasaka (2004). “Isolation of a new member of group 3 Late Embryogenesis Abundant protein gene from a halotolerant green alga by functional expression screening with cyanobacterial cells.” *FEMS Microbiology Letters* **236**: 41-45.
260. Tannacliffe, A. and M. J. Wise (2007). “The continuing conundrum of the LEA proteins.” *Naturwisseschaften* **94**: 791-812.
261. Tatusov, R. L., E. V. Koonin and D. J. Lipman (1997). “A genomic perspective on protein families.” *Science* **278**: 631-637.
262. Tehei, M. and G. Zaccai (2005). "Adaptation to extreme environments: Macromolecular dynamics in complex systems." *Biochimica et Biophysica Acta* **1724**: 404-410.
263. Thanassi, D. G. and S. Hultgren (2000). "Multiple pathways allow protein secretion across the bacterial outer membrane." *Current Opinion in Cell Biology* **12**: 420-430.
264. The *C. elegans* Sequencing Consortium. (1998). “Genome sequence of the nematode *C. elegans*: a platform for investigating biology.” *Science* **282**: 2012–2018.
265. Thöny-Meyer, L. (1997). “Biogenesis of respiratory cytochromes in bacteria.” *Microbial and Molecular Biology Reviews* **61**: 337-376.

266. Tiao, G., C. G. Lee, *et al*. (2011). "Rapid microbial response to the presence of an ancient relic in the Antarctic Dry Valleys." *Nature Communications*: **DOI 1038/ncomms1645**.
267. Tolleter, D., D. K. Hincha and D. Macheael (2010). "A mitochondrial Late Embryogenesis Abundant protein stabilises model membranes in the dry state." *Biochimica et Biophysica Acta* **1798**: 1926-1933.
268. Tompa, P. and D. Kovacs (2010). "Intrinsically disordered chaperones in plants and animals." *Biochemistry and Cell Biology* **88**: 167-174.
269. Tuffin, M. I., D. Anderson, *et al*. (2009). "Metagenomic gene discovery: How far have we moved into novel sequence space?" *Biotechnology Journal*: **DOI 10.1002/biot.200900235.**
270. Tyagi, R., A. V. Shenoy and S. S. Visweswariah (2009). "Characterization of an evolutionarily conserved metallophosphoesterase that is expressed in the fetal brain and associated with the WAGR Syndrome." *Journal of Biological Chemistry* **284**: 5217-5228.
271. Tyson, G. W., J. Chapman, *et al*. (2004). "Community structure and metabolism through reconstruction of microbial genomes from the environment." *Nature* **428**: 37–43.
272. Ulusu, N. N. and E. F. Tezcan (2001). "Cold shock proteins." *Turkish Journal of Medical Science* **31**: 283-290.
273. Upton, C. and J. T. Buckley (1995). "A new family of lipolytic enzymes?" *Trends in Biochemical Sciences* **20**: 178-179.
274. Venter, J. C., M. D. Adams, *et al*. (2001). "The sequence of the human genome." *Science* **291**: 1304–1351.
275. Venter, J. C., K. Remington, *et al*. (2004). "Environmental genome shotgun sequencing of the Sargasso Sea." *Science* **304**: 66–74.
276. Verger, R. (1998) "Interfacial activation' of lipases: Facts and artefacts." *Trends in Biotechnology* **15**: 32–38
277. Vermupalli, R., Y. He, *et al*. (2000). "Over-expression of a protective antigen as a novel approach to enhance vaccine efficacy of *Brucella abortis* Strain RB51." *Infection and Immunity* **68**: 3286-3289.
278. Villeneuve, P., J. M. Muderhwa, *et al.* (2000). "Customizing lipases for biocatalysis: A survey of chemical, physical and molecular biological approaches." *Journal of Molecular Catalalysis- B Enzymatic* **9**: 113–148
279. Virk, A. P., P. Sharma and N. Capalash (2011). "A new esterase, belonging to hormone-sensitive lipase family, cloned from *Rheinheimera* sp. Isolated from industrial effluent." *Journal of Microbiology and Biotechnology* **21**: 667-674.
280. Voloshin, O. N., F. Vanevski, *et al*. (2003). "Characterization of the DNA damage-inducible helicase DinG from *Escherichia coli.*" *Journal of Biological Chemistry* 278: 28284-28293.
281. Wei, Y., J. A. Contreras, *et al*. (1999). "Crystal structure of brefeldin A esterase, a bacterial homolog of the mammalian hornone-sensitive lipase." *Natural Structural Biology* **6**: 340-345.
282. Wimberly, B. T., R. Guymon and J. P. McCutcheon (1999). "A detailed review of a ribosomal active site: the structure of the L11-RNA complex." *Cell* **97**: 491-502.
283. Windsor, A. J. and T. Mitchell- Olds (2006). "Comparative genomics as a tool for gene discovery." *Current Opinion in Microbiology* **17**:161-167.

284. Winkler, U. K. and M. Stuckmann (1979). “Glycogen, hyaluronate, and some other polysaccharides greatly enhance the formation of exolipase by *Serratia marcescens*.” *Journal of Bacteriology* **138**: 663-670.
285. Wise, M. J. (2002). “The POPP’s clustering and searching using peptide probability profiles.” *Bioinformatics* **18**: 38-45.
286. Wise, M. J. (2003). “LEAping to conclusions: a computational re-analysis of Late Embryogenesis Abundant proteins and their possible roles.” *BMC Bioinformatics* **4**: 52.
287. Wise, M. J. and A. Tunnacliffe (2004). “POPP the question: what do LEA proteins do?” *Trends in Plant Science* **9**: 13-17.
288. Withey, J. H. and D. I. Friedman (2003). “A salvage pathway for protein synthesis: tmRNA and trans-translation.” *Annual Review of Microbiology* **57**: 101-123.
289. Wooley, J. C., A. Godzik and I. Friedberg (2010). “A primer on metagenomics.” *PLoS Computational Biology* **6**: 1-13.
290. Wynn-Williams, D. D. (1996). "Antarctic microbial diversity: The basis of polar ecosystem processes." *Biodiversity and Conservation* **5**: 1271-1293.
291. Yang, X. and E. E. Ishiguro (2001). “Involvement of the N-terminus of ribosomal proten L11 in the regulation of the RelA protein of *E. coli*.” *Journal of Bacteriology* **183**: 6532-6537.
292. Zang, X-X. and P. B. Rainey (2007). “Genetic analysis of the histidine utilisation (*hut*) genes in *Pseudomonas fluorescens* SBW25.” *Genetics* **176**: 2165-2176.
293. Zerbino, D. R., and E. Birney (2008). “Velvet: algorithms for de novo short read assembly using de Bruijn graphs.” *Genome Research* **18**: 821-829.

Appendices

Appendices

Appendix I

Supplied on disk

1. Raw data of Solexa sequence reads (r7_s_8)
2. End-sequences of fosmid clones
3. *De novo* contig assembly (CLC genomics workbench)
4. *De novo* contig assembly (Velvet)

Appendix II

ORF predictions by SoftBerry for fosmid clones LD4, LD7 and LD13.

CLONE LD 4

```
>GENE     1      1132   -      1767    316    211 aa, chain +
MHISASLITDSVVQAANWLQKGQLLAYPTESVWGIGCNAYDKEAVQRILDIKQRPPAKGM
IVVTDSAARIAPLLERLNDQQRHTVLESWSHSPQALAQQAHTWLLPINAPLKIPIPSWVT
GAHDSIAVRVIAHPLIRQLCAQMVTATNPYGFIVSTSCNPSGLPPARSLKEAQAYFAESE
FAEQVGYLQGATLGYQLPSQIHDALTGQVIR
>GENE     2      2057   -      3061    667    334 aa, chain +
MNILISGGSGFLGSAFSTELMNRYRAQDKEIHITWLTRDSSQQHPDGINLMTYDELATFD
GSDSSNSSNKGFDVILNLAGAGIADSRWSDERKETLLASRVKPTESLLAFIARTSHKPTL
LVSGSAIGWYGTQGDKPLTESSAYETDFSHRLCDDWEQLALKATEYGVPVAIVRTGVVIH
PDGGMLGKLLLPFKLGVGGQLGDGKQIMSWVSRDDWVGAAIFIIERHLADKVNAQQDSSS
TTDDYTLQTASDTSVVVYNLTSPNPVTNHTFTKTLGSWLHRPTFFTLPSPLLKLMFGEMS
TLLIDGQKVLPQALLDAGYTFKQPTLKQALEQQS
>GENE     3      3265   -      4311    503    348 aa, chain +
MLALTTSIILLWFIQPKRRSGAIATIAIVFIINNIFLAYGLTEFWLDRFHIYLVIGILQG
FMVYAALITIAIALLYCRLLKQPRRPKLIRGIGIITYVGIVSFAVFNAYSPVVHRLTVTT
DKPMKQPMNIALVSDTHLGRWFGNRQLDKLVTLIDEQNADMIVMAGDIMNDTTIAYDKTN
MHEHLSKLSAPLGVYAVLGNHDYLGYEKRIAAAVTKAGITVLDNQNVRLNESVSLVGRSD
DNDPTRLSASKLLAKVDTDKPVIFLEHRPTAIDEIKGLPIDLHLSGHTHGGQIFPLTTLM
KWFKPLVHGTKNIEDTHFLVTSGYGFGPVPFRLGTRSEIWMVTLQQAE
>GENE     4      4431   -      5576    582    381 aa, chain +
MRYMFFITIIALLQLFSLGAALSLQWWLQPWSVSIPLLPRIIFGVIFVISNGLLLLSVSK
LLANSYRWVSGWMLFMHFMMLTALLTSLLYGGYWLITMLTGVSLIDTDMVAVGLRALALL
IFVAMFIYALYSAYIPVVRKQTISINKPMQGQLRIAVASDLHLGRLFGNRAIDRLRHLVV
QSQADILLMPGDIMDDNTKAFNDYDMENNLAQLCASLPYGIYATLGNHDLYGHEKPISHS
LRRAGVHLLNDEVLHLLHKEQSIWLVGRFDNHKRQRAATTDLLAQVDTAEPIILLDHRPS
DIVEHSQMPIDLQVSGHTHNGQIFPANFIVQAINRLGYGYEAINESHFVVSSGYGFWGIP
FRLGSRSEVWLITLTGQAQAG
>GENE     5      5551   -      6945    693    464 aa, chain -
MPVLPIPSVNALVSKTANTIKNLRSRASTSSSELNLLSTSPLNHRSASKNIDRTTARYKK
HTLHYVLKALGYLPTPFLEKFNSTLHGPNTKQYLHADAHLRLILALSNKLKQPLDIEKLP
ILRQKFAADTVAMQAPAVWDQSDSTSVRWQDKAIMNADDGEMTVRCYQHYLQNEADLDIN
```

KQTTNQQTVNQKNTDKTVMLFIHGGGFCIGDIDTHHEFCHAVCTQTGWSVVSVDYRLAPE
HPAPTALRDCLSAYAWVAEHCHTLGALPSRIVLAGDSAGGCLAISVAQQVSAPDASQWLN
LGLDNDKITQLLQSLPRPLAQLPLYPVTDIEAEHPSWALYGQGLLLDHNDVEVFDAAYMQ
QSAIAQSHPLVSPMYGDNTQMCPSYIVVAELDILRDEALVYAEELQNKGIKVETHTVLGA
PHGFIHLMSVHQGLGDETDYIINEFGSFVRQLLTRDFNQPEPVL
>GENE 6 7432 - 8037 681 201 aa, chain +
MSEQVPNQEFDVKNPKAAQSNIDREESVLEETMNEFDPENNSGANVTIENEIDIDAFHAR
IAELEGEVKQAKEGTARANAETYNAQKRMEQEADKSKKFALQKFAKELLEIVDNLERAIE
SADANDPVTEGVKLTHKALLDVLNKNGVQVVEPQGEKFNADLHEAVGIDAEAEADIVGTV
LQKGYSLNGRLLRPAMVRVGQ
>GENE 7 8305 - 10245 1683 646 aa, chain +
MGKVIGIDLGTTNSCVAVMEGDSVKIIENAEGTRTTPSIVAYKDDETLVGQSAKRQAVTN
PNNTLFAIKRLIGRRFDDKVVQKDIGMVPYKIAKADNGDAWVEVNGKKLAPPQVSAEILK
KMKKTAEDYLGESVTEAVITVPAYFNDSQRQATKDAGKIAGLDVKRIINEPTAAALAYGM
DKKQGDSTVAVYDLGGGTFDVSIIEIADVDGEQQFEVLATNGDTFLGGEDFDSALIEYLV
AEFKKDQDVNLKGDSLAMQRLKEAAEKAKIELSSAQSTEVNLPYITADSNGPKHLVITIS
RSKLESLTEELVKRTMGPCKIALEDAGLKASDIDDVILVGGQTRMPLVQKQVQDFFGQEP
RKDVNPDEAVAAGAAIQGAVLSGEKTDVLLLDVTPLTLGIETMGGVMTPVIEKNTMIPTK
KSQVFSTAEDNQPAVTIQVYQGERKIANQNKQLGRFDLTDIPPAPRGLPQIEVSFDINAD
GIMNISATDKGTGKAQSIQIKADSGLTDEEVEQMVRDAEANAEEDEKFANLAQVRNEADG
RIHAVQKALKEAEDKVTDEEKSAVETAISELETAAKEDDHDDIKAKLEALDNAFLPVSQK
IYADAGAGAEGMDPGQFQQETESADGTQTDDDVVDAEFTEVNEDKK
>GENE 8 10682 - 11215 700 177 aa, chain +
MKTLQKTLLALAVGSLVSVGAQAAVTYGSGYTGTPYVGVKVGKFDLDTNGASDPTAYGVY
GGYNFDPNFGVEAEYVGSDDANYYNGDVDAKSYGVYGTYRYAFADTPVYAKGKLGIARTK
IEGDSRLPYASVSDSDTRIAGGIGLGYSVNPNFGVEAEYDILSSDANLMTVGAHLKF
>GENE 9 11455 - 12003 719 182 aa, chain +
MNTLQKALIALSVGSLLSVSAQAAVNYAGQPYVGVKAGKFMVDVDGLDDPTAYGVYAGYN
FDSNFGAEVEYVGSSDTDIDTSTRLVESEYDLKTYGAYGTYRYQFPNTGLYAKGKLGFAK
AEVDVSASNYLGYSNSVSDSDSGLAGGIVLGYDFNPNMSIEAEYDYVAEDITLLTLGANL
KF
>GENE 10 12249 - 12860 373 203 aa, chain +
MIETIWMIEPWHWLVLGFLLMIAEMFIPTFASLWFGAAAVIVAALSWLLPIPDVVQILIW
LTLSVLFMFAWVKYIKPLSINRTKAGLGGSVIIGETGMIVLKPPVGGLGLVRFNVPIVGA
DEWSCRTLDETVEVGDRVVVTGLSGNELIVAPTKRLTAPSVNQPNITPTIENQPVTTKAK
ASRLNLNKPDLDRARLKKKFIDK
>GENE 11 13051 - 13932 800 293 aa, chain +
MNSIISSMNSLTIVMLVLVALIAFTVFKGVRIVPQGYKWVVQRLGKYNQTLEPGLTLIIP
FIDRVAYKVTTKDIVLDIPSQEVITRDNVVIIANAVAYINIVRPDKAVYGIEDYEYGIRN
LVQTSLRSIIGEMDLDSALSSRDEIKALLKHAISEDIADWGITLKTVEIQDINPSDTMQM
AMEEQAAAERQRRATVTRADGQKQAAILEADGRLEASRRDAEAQVVLAKGSEESIRLITG
AMGKEEMPIVYLLGEQYIKAIRQLAESDNSKMVVLPADILSTVKGMVGDKLKV
>GENE 12 14224 - 14763 467 179 aa, chain +
MTKQNVTTKLIDISAFVQGTPLPFSVLDNKHINAAYPEEKLEIRGGGYGSDAAAHPTNAN
QFYALTDRGPNADFEGSAGSGKQFLLPDYTPRIGLFELQADSQIIKIKDILLKDTQGNSI
SGLPNPKALGGTNEIPYDIAGEVMTVHPHLPFDENTNPIRTDINGLDPEGLAALTDGSF
>GENE 13 14887 - 15660 421 257 aa, chain +
LPTEFAKRRANRGMESLTITPDQTTLVGVMESSLDNPDKSGRGSNLARIVTINLSSGQIA
QYLYRLDKAQHVTSGIVALSKHEFYIIEHDRKFPLQDKSTQKVIYKINIAQATDIERLAD
GNPASKVDIKQDNAFGLTINEQTLEQFVASDDNSWQTFEKMNIKPVIKTLVVDVMATLDY
PHDKLEGLWLRQDGSLGLLNDDDFSMTDSAITDLVKTDFSLTASSATKLKSHVEQKYLDS
QKTLEDANRLYLVMPNT
>GENE 14 15729 - 17267 863 512 aa, chain -
MPEHNDIRQQPASSDTEQPTTVRPLVQPDNPYASTRGSIMTKQLPAFDEALINKYNRSGP
RYTSYPTALEFLPIPDGLETKILQNREARAPLSLYFHIPFCRHLCYYCACNKIITKKNSD
SGDYLQYLIAEIKHKRRLLSSPESDRNPLVKQLHLGGGTPTFLRDEEMVQLWEFLQTQFD
FLPENEGDYSIEIDPRELSGETLKTLRNLGFNRVSLGVQDLDETVQIAVNRIHSAELIEN
VLTEARELGFHSINIDLIYGLPHQTPATFDKTVQRIIEMSPDRLSVFNYAHLPERFKAQR
QIKDEDLPGPSAKLTMLGNTINTLTEAGYQYIGIDHFAKPDDELAVAQREGHLHRNFQGY
TIMGDCDLLGFGVSSISQIANANTRYILQNDTDLQVYQDTIDAAKNNLAIMPAVKVIKTS
IKDRLREYVIMNLLCHDYIDFRDVNQKFGIDAITYFIDEIQQLGAMQADKLIDMDAAGIR

VLPKGRLLGRNVAMVFDEYLDKKHQNRFSKVI
>GENE 15 17527 - 18471 454 314 aa, chain -
MTDINYPHYLQQLGFTKSESDKPITPDLETLARLQLAHLTQYPFQSLTTILDRPVDLEDS
SIYDKLVKRRAGGYCYELNGLFLSLLRHLGYEAKIITGIVIIDNQLERQNARTHMVIMVT
IDGQNYLVDVGFGGLVPSAPLLFAYNNQESSNQETDDKNASQIQTTPHGRYRIIKDDSFE
NQSENPILLNAKVNYERYILCCEIKSEWQMLYVFDLLPQIKIDMIVGSWCISTYPKSPFK
SRLMASRLDEQGIRHTLLNNQYRRHQLGKPSQSRALADIDELLTLLKQVFFMDIACELTP
SERQKLTIFLENNG
>GENE 16 18517 - 18840 115 107 aa, chain +
LRIGTNSLFVLWKTSVVSSQLLDMLSSFLISIILYIPPTTYNVKCKEKDDDDIIKNYPDT
CALKVVIDRKLLQSLKEDELNNFTTVIHFLKNLRAMVLLSAETGSER
>GENE 17 19345 - 21066 937 573 aa, chain +
MNIYDLSEYQLYKLKSMDPNLSPDWREVIQHILPKLDLESQNSLYKNILEPMGISFNTNE
ELIYKHPATLKETIRDIQTHNNDLLAIVSNMYKIIDSRVDYYDAIQLADEIEAIFGYLDN
LNIRHLLYEHKNRKRIRIAFLYDLARWIDTVELKVSPGLRKLDSRIVISYLKEVFIKQKI
QGQDFRKWDSSDLSFQEFTHLPFFIRSEGEKRNFFVVEGREYWFLIGNTDEPKKNPYSLR
RFLHEECSEKYVYLTHVVINKDEMRNSQYLSRASHAMSRFYTLDLGTPDTLLSFVKEAQS
FRKRYLKPLLKERLEQTGGSTEAIIKERMITYEKQVSVLILQKIPRVQAILYSKKDQDYL
FYHLDKLIKKMIENVQDFRLQPLVMHSTSSEILLIKLMALRKLINRSYDFTFYQDLSIEE
RSEAMSIPLFMVKEKLSETKASIKELKDLKEKIDNYFLIKENGSFWKRIVWGKRPSYTLE
DVTKEKLMLEKEVFMFIVRMAKSQNKGMIYIEFEFDEIINKNYRHYALADGQLGISRLPR
VLRLPEEKREFNIEYISQVVNHNIFEANQLWHM
>GENE 18 21351 - 22943 561 530 aa, chain +
MSFAQVSTRSVVGLHAPQVIIEVHLSQGLPALTIVGLPEAAVRESKDRVRSAILNSGFQF
PNRRLTINLAPADLPKDGARLDLPIAIGILAASDQLDPDVLSGFEFIGELALNGNLRQVT
GSLAVARAIKAETLVLKVSKASDTSKDAGQEPDQSQRLSPSTTPQLIVPIDNGAEASRVA
GIEILGAQSLKAVCDHLQSLADPSAKDGRLEVVTPSPVQQHAGYQVDLADVKGQHHARRA
LEIAAAGGHSLLFTGPPGSGKTLMASRLPTILPDLSAEDALEVASTYSVADSDYDYGTRP
FRQVHHTISAVALVGGGSRPKPGEITLANKGVLFLDELPEFDRAVLEALRQPLEAKQITI
SRANSQMTFPANFQLVAAMNPCPCGYDGDTSVRCRCRPEQIKRYQDKLSGPLLDRIDLHI
TVPALPIADLQNSRTGESSEQVRARVSAAHKHQLTRQKKVNNELSPSEIDEHVPLGEGEQ
QLLQLAQQRLNLSARGYHRVLRVARTIADLADSIDVTSAHVSEALSYRSK
>GENE 19 22956 - 24713 1121 585 aa, chain -
MKETFDMTAPIITVPIITVPSDIAIAQNTILRPINAIAEQLGLSAKDIEPYGHYKAKIDP
AAVFAMPAKAKRSKLILVTAINPTPAGEGKTTVTIGLADALNRLHTQQNNGKKTVVAIRE
PSLGPVFGMKGGAAGGGYAQVLPMEDINLHFTGDFHAIGAANNLLAALLDNHIYQGNELN
IDPKQIFWRRAVDMNDRQLRNIVSGLGKRTDGMVREDGFDITVASEVMAIFCLAADLPDL
KQRLGNILVAYNKDKQPVYARDLHAQGAMAALLKEAIKPNLVQTIEGTPAIVHGGPFANI
AHGCNSVIATRVAMHLSDYTLTEAGFGADLGAQKFCDIKCRLSGLTPDAAVIVATIRALK
YNGGVAKDSLTDENLSALKQGLPNLFKHIENMQEVYGLPVVVAINHFVSDTDAEVDIVRQ
ACREKGVEVALTQVWEKGGDGGEALANTLLTLLDSHDSQPSQFRLAYDSNTSMANKIRTV
AQRIYGANDIEISSLAQTKIERLEALNLDRMPICIAKTQYSLSDNAKLLGRPINFDIHVR
DISISSGAGFIVVICGPIMKMPGLPKRPSAERIDVDDAGHITGLF
>GENE 20 24697 - 25401 556 234 aa, chain -
MSKHFKEQKQLLIFDFDGTLIDSVPDLADATNTMLTTLGKSTYPIDTIRNWVGNGSRLLV
ERALVGKVEVAEGELTQEQADHAEQVFFDAYSNISDSKTVAYPDVNIGLNKLHKAGFTLA
LVTNKPIQFVPKILQTFGWQDLFAEVLGGDSLSVKKPNPAPLLHVCTALGVTPDKAVMIG
DSKNDILAGQNANMDTLGLSYGYNYGQDIRDFHPTQTFDDFASLIEYLLDERDL
>GENE 21 25577 - 26077 379 166 aa, chain -
MREFDYDLDYKHLDLRAQPELYRVGRGEQGVLLVEPYKSEILPHWRFATPDIATESSETI
YQMFLDYLAADDFVGADMARKFIQMGYTRARRYANHKGGKKYKGPVPDDKKGQSGAHGRE
ELPRQEEDPIKAESARIFKQKWDLCRENKTYLTMKAAHRKRYQEVE
>GENE 22 26373 - 26723 243 116 aa, chain +
MDKIDIEKPVWSLPHSKALVTILGWLLIGMLSINSAHSAWFERLIPQNGTFTVSDVDRDV
AFVIDDFIYDARKFCYMSVGDRVVFFEGRHGIDYRSTVYNLNSRERCELLLRDRVE
>GENE 23 27068 - 27490 375 140 aa, chain -
MPLEAIAINNLNTPLIKKPTRKHELAERLQALYDADDNQPDGKEVLNIIEQGFKRGQILY
VGMFNDKPIAAVGCFDDGQTDAKRLQYLTVHPENRKRAIDAKFIKLVYDAEVKKGVRQFV
PVDSDIHQIMSEYDLLQVKG
>GENE 24 27820 - 28464 465 214 aa, chain +
MTVNLATISDANQNLHGRPARIATVERNTAETQVTCTVNLDGTGQGTVDTGVPFLDHMID

QIKRHGLFDLDIKCTGDTFIDDHHSVEDTGITVGQAFAKALGDKKGIRRYGHFYAPLDEA
LTRAVVDLSGRPGLHMDIPFTRSHVGTFDVDLFSEFFYGFVNHSWMTVHLDNLKGKNSHH
QIESTFKAFARALRMACEYDERALNTLPSTKEAF
>GENE 25 28633 - 29268 294 211 aa, chain +
MTKIALLDYGMGNLHSASKALSAVGAEVSITNDPKVVAAADKIFFPGVGAMRDCIAGMNE
AGIDDVIRHAIFNKPVMAICVGMQALFEQSAENGGTKCLSILDGTVEAFDPTWKDERGAT
IKVPHMGWNTINGMDLNHPLWKGVENNAHFYFVHSYYCAPTDSSQVAAVCDYGQPFCASI
IKDNLFATQFHPEKSHTAGLQLLKNFVEWDI
>GENE 26 29459 - 29878 169 139 aa, chain +
MPSQIWKTPVDFTAFKIDIIETTGLAKDALHIYVGIGVYLLCLLVLRPIIKKQSIRSFMA
LIFVTGIALLGEYLDNRHIIRPRGFFALGIVDIKASLHDLINTCLLPYGFHALNKWTTIF
HPTNKITPITKRHKKTDYS
>GENE 27 29909 - 31267 993 452 aa, chain -
MKFSKFGQKFTQPTGISQLMDDLGDALKSDKPVNMLGGGNPAKIDAVNELFLETYKALGT
DNDTGEANSSAIISMANYSNPQGDAVFIDALVSFFNRHYDWNLTSENIALTNGSQNAFFY
LFNLFGGAFSDNKLDEQSQDKENQSIDKSILLPLAPEYIGYSDVHIEGQHFAAVLPHIDE
VTHDGEEGFFKYRVDFEALENLPALKEGRIGAICCSRPTNPTGNVLTDEEMAHLAEIAKR
YDIPLIIDNAYGMPFPNIIYSDAHLSWDNNTILCFSLSKIGLPGMRTGIIVADAKVIEAV
SAMNAVVNLAPTRFGAAIATPLVENDRIKQLSDNDIKPFYQQQAKLAVRLLKEALGDYPL
MIHKPEGAIFFWLWFKDLPITTLELYERLKEKGTLIVPSQYFFPGVDTSHYQHAHECIRM
SIAADEQTLKDGIAVIGEVVRELYDEAKSKNA
>GENE 28 31552 - 32499 641 315 aa, chain +
MKAVLLDEKRFVADLARPTPEKITDYVTFEETPQDNKTIIERCQDADIMINGSLRIEREV
IEALPNLKLIQLLSVGSNHIDKQACEDNDVKVLNAPDFASATVAEHTMMLLLSAMRASIH
YHNKVKSGEWKEKGRADIIDAEAIDLEGLTIGMIGIGDIGKRITKLAKAYDMNVLWAERQ
GRKPRNDDYTDFETVLSNSDVITIHCPLTEETKHLINQDTLSKMSKKPLLVNVARGKIVD
SEALASAVKQEQILGYATDVFENEPADEDDPIVQLANEGHPRIILTPHMGAGSRASQVKL
WKIITRQINEFIENN
>GENE 29 33296 - 33433 87 45 aa, chain -
MLNLLALINILIVYLTNDYGAAQISILTLEWSFGALNHNFTYSST
>GENE 30 33583 - 35052 508 489 aa, chain +
MTLINESFFQNKQKKRNYLHFDKKYSPSFLYSYIINPSNIIQHSFYPFVSYNLHDRKIKK
GYIRIAKAKHCPIKGYIPTTPIFKPPISSPNIIQHSHYPFISYKKAKKYAVKSNKVRLIN
YASHLDSAIYAYYTELLSPSYEESLIRNNLENTVLAYRKIERTVDCKTVSKCNIHFSKDV
FSIVSQKKDCIVLCFDISKFFDNLDHQILKDNWCSLLNVKHLPEDHYKVYKSLTKFASVD
KELLYKELGLSLNSRTLHKRHKTLCETKDFRSRVRGKGLISTNYTSRGIPQGSALSGFLS
NVYMMQFDKNIKKYLENINSVYFRYSDDMIFIVDDINELQLVEFIKCEIGKLKLAINDNK
TQRVVFENGIANVDVNNLSYNNPSKLQYLGLLYDGKQVFLRDTGISRYNHKLRKAVRMRS
AHYRKLKQNNNQNGTTIYRRTLYSRFTYIGKRNYISYVFRVSEVHNSKNVKRQVKGHFKL
FKEYLNERV
>GENE 31 35264 - 37096 1035 610 aa, chain -
MTMITPSYLGETIEYSSLHACRSTLEDPTVAMYGVITAFVTKQQANNAIRMAMGEPVKQI
KVSYYRLLLATSIIASIAAVVLSLGLLTVGQPYLAAILPADIGLAINPISIIKTLVIAII
LTLLITQRGLSPLSTTKPATLLNQSANQQLGKQPWYQRLPLLWYGLMLIGLYLFFAYEVG
SWILSAQLLIGLIAFVAVFWALARAWLWLLTKLAMNTDISWMKRIAIHNLARKGNQSALF
FVTLSLSVAVLTLITTLNHSINAQFINAYPEDAPNLFLLDVQSDQHEDIDALIGAPVSYY
PVIRARVVTANDVPAQDIEPADGFDEPTRVFNLSYADTVMDTEFITDAVEDNELYSPIKT
IKADGQQDQNSQIAPLSILDTAASMLNVGMGDQVRFNIQGIEIVGQITSIRSRYERGPSP
YFYFLFEPSVLSAAPQIQFATAHVAEDAIPELQGKLVRQFPAVTTIDGTAIAEQVQELVV
QMSRLVYVFTLLALLTGVMVLISSLLSTSQDRMQESASFRLLGMQKRDLYMLNILELGVL
GISAATFAVIIASVGAWAAITQWFNLRFSVPWTSLGIGGVALVALLFGIAIIYVRLVIGR
GIMARVRAMI

CLONE LD7

>GENE 1 129 - 632 330 167 aa, chain -
MAQKQSILGRVTQMTRANINAALDRAEDPEKMLDQLVRDYTDSIAEAEDAVAQTIGNLRL
AEADHADDIAASHEWGHKALTASKKADSLRSSGNTEDAQKLWDPRVPSSNSPYSESYYNS
LAVVLQRRDWENPGVTQLNRLAAHPPFASWRNSEEARTDRPSQQLRS
>GENE 2 730 - 1425 662 231 aa, chain -
MDGQGWLQISELLLAFTLSSLVGLERQLRGKAAGLRTQALVGTTSALLMLVSKYGFNDVL
LAGHVVLDPSRVAAQIVTGVGFLGAGLILTRQGNVKGLTTAAAVWETAAIGMAAGAGLWL
LALVVTFLHFVTVYAFTALTRRLPGAAVTSVRIELTYVEQHGLLRTLLGRVTELDWKLTG
LSPHETPDGDAAVTSVYLDLSGNGNPADIVRTLAITDGVRRAGVISEEELS
>GENE 3 1600 - 2190 167 196 aa, chain +
MVTHVLKFARASTSRWRNNGGVTHEIARRKNKIDTNAFDWRISVADVTAPGPFSPFPGID
RVRVFCEGQEMSVTVDGLTHDLSRWAPFHFSGDADVSGAVPRGATRDLNVMTNRALFRTT
ADVDEFAGSHTVVAPERTDVVVVVVEGSMKLPGKESCECGDLGIYDAFTMSGPGTVVVKG
SARLVTVQFHLIDDTP
>GENE 4 2392 - 3639 702 415 aa, chain +
MTAEASPLVDTLADRVRHEVRLQRIDPKSNAAAVESIAVRVIQDHEGRSLTGAVQALEDP
DMLVSRLVADVAGFGLLQQYFDDPDVEEIWINEPSRIFVARDGKHELTASILSVEQVREL
VERMLGSSGRRLDVSNPFVDAMLPGGHRLHVVLDGISRNFTAVNIRKFVAKAHSLDDLVV
LGTLTQPAADFLNAAVIAGLNIVVSGGTQAGKTTLLNCLAAAIPGGQRLVSCEEVFELQC
GHPDWVAMQTRQAGLEGNGEIALRSLVKEAMRMRPSRIIVGEVRAEECLDLLLALNAGLP
GMASIHANSARQALVKLCTLPLLAGENIGSRFVVPTVGSSVDLVVHTGIDASGHRAVQEI
VAVTGRVENDIIESEMIFCIRGGQLVRGSGMPARRDRFEQVGIDLDDLMGAPWVP
>GENE 5 3777 - 4466 364 229 aa, chain +
MSLAVLLATFLAMVVISGSLIVAAVFGVLAAGGPAAVIRGNAQRRQREFAELWPDAVDNL
ASAVRAGLSLPEALQQLGERGPEPLRPPFAAFGRDYQSSGRFHEALDLLKDRLADPVGDR
VVEALRIAREVGGGDLGRMLRTLSGFLRDDLRTRGELESRQSWTVNGARIAVAAPWLVLL
LMSFDRDVIGRFSSGPGLLVLGIGAVTCICAYRLMMWIGRLPVERRILA
>GENE 6 4739 - 5365 356 208 aa, chain +
MQQVMWGLSGFALAAVLSALVYATRSTSVMALLIMCCAGFLGGVLARDQKLSSEVKHYEE
RLSEEFPTVADLLALAVAAGEGPAASLERVIRVCHGDLAVELGRVLAVIRTGTPLVRAFD
QLAARTGVATIARFAEGLAIAVERGTPLVDVLHAQAADVRESTRRELIETGGRKEVAMMI
PVVFLILPITVLFAFFPGFIGLHLTSGF
>GENE 7 5387 - 5554 76 55 aa, chain +
MRSIFNRYLDRGDRGDVPGWVMITVMTAGLVAMLTAVAGPQLRSMLSQALSSVGG
>GENE 8 5623 - 5943 195 106 aa, chain +
MPLVLGILQVGLVLHVRNTMVAAASEGARYGAAVDATAADGATRAQHLIRTSIADRYARN
VSAVEVVRSGRQQVVVSAHTSVPALGLGGPGFALVVRGHATKEVAR
>GENE 9 5940 - 6380 236 146 aa, chain +
MSTRAHTDHGSALIEFVWLAIVLLLPMVYILIAVFDVQRAAYGVSAASKSAARAFLLAPD
EVSARHRAEEAASLALADQDVRAGAVTIVCTPSQSSCLTPGSSVRVVVRVTQKLPLTPSF
LGDQIAAFTVDSTHVEPYSRYRMARQ
>GENE 10 6377 - 6793 148 138 aa, chain +
MSQRPETGQMAVMIVGFFVVIGLLAVVVINASAAYLQHQQLANAADGAALTAAQTVAEDS
IYRKGVEVGDLPLNSQSASAAVGAYLRGATGIRWQVVLQQRQVNVQLARRLRLPLVPPGW
MDNTLVTARASAVLRVHQ
>GENE 11 6861 - 7967 966 368 aa, chain +
VAATDFPEELDSIKATLTSIEKVLDLDSMRAEIADLQQQVGAPDLWDDQENAQRVTGRLS
LLQSELERTTNLRQRIEDLEVLVQLGQEEDDADSLAEAETELKRIHKTIDALEIRTLLSG
EYDERDALISIRSGAGGVDAADFAEMLMRMYTRWAERHHYSIEVYDTSYAEEAGIKSATF
AIKAPYAYGTLSVEAGTHRLVRISPFDNQGRRQTSFAAIEVVPVLEQTDEIELPEDEVRV
DVYRSSGPGGQSVNTTDSAVRLTHIPTGTVVSCQNEKSQLQNKASAMVILKAKLLALKKA
EERATLDEMRGDVQASWGDQMRNYVLNPYQMVKDLRTEYETGNTQAVLDGEIDDFIEAGI
RWRRSHAV
>GENE 12 8044 - 8733 637 229 aa, chain +
VIRFDKVTKTYPGQKRPALDSLDLEIQKGEFVFLVGASGSGKSTFLRLVLKETTPTQGRV
YVAGKEINKLSSWKVPKLRRQVGTVFQDFRLLPNKTVTENVAFALQVIGKSRAEINRLVP
ETLEMVNLEGKGHRMPDELSGGEQQRVAIARAFVNRPMILIADEPTGNLDPTTSVGIMKL

LDRINRTDTTVIMATHDSTIVDQMRKRVIELVDGKIIRDEARGIYGYQS
>GENE 13 8742 - 9662 675 306 aa, chain +
MSFRHTLAELGAGLRRNKSMTISLVVTMSVSLLLASLGLLIQSQADRTERYFGDRLQLQI
NLCTKNSPGPNCLGGVATNEQKLAVKAALSENPEVKDFETRTPADNYDQARALLGQTDTG
RKQLATLGPEAFPESYFVTLQNPREFDGVVSQVSGMDGVGNVNSLRKLLGPLFEMLDKMR
WAALGTSLLLIVAAILQVSNTIRMTAYARRREIGIMRLVGASSWHIQLPFILESMMAAVI
SAIVAGGGLMAFMYFVVYGYLRDTLGQITTWVRWQDAIAVMGYTTILALVLALIPTLVMT
RKYLDV
>GENE 14 9777 - 11108 875 443 aa, chain +
MHFTGNPSVRNTMLAAAISCALVTALVSPATADSKSDLEKQRRGVSGDIGNAQKSYDQSS
KQYAGAVNALKKAQGRLDSAQTHLGETRGQLAAAAAKDAQMQRQLEASQAALEKALAELK
SGEKSLAASEVQVKQFTLESLQEGDRGLRAFGDLLRGSSPSAFSERMSLNSSVGDAQLAT
MEHLAASKVMLGLKRDKVGQLRDKVALKRKEAAANLAQKEVLEAAAQEQTVQVGELVGKR
SSAKKSANKILAGDAVKLRELERDKDRLSSQLRALAAAEAAKAAKAARRKAQGKPTNNGG
GGGGGGNSSGGGGGTLSRPVYGPTTSPYGMRTHPVTGVYKLHDGTDFGVGCGTPIHAAAS
GTVISRYFNAGYGNRLIINHGWMRGANVVTAYNHATRYIVGQGQRVSRGQTIGYVGSTGY
STGCHMHFMVLVNGSTTNPMGWL
>GENE 15 11142 - 11621 464 159 aa, chain +
MSKETGRKLIAQNRKARHDYTIEDTYEAGLVLTGTEVKSLRAGRASLVDGFADIEDSEIW
LLNVHIPEYTQGTWNNHATRRKRKLLLNRTEIEKIEHRVTQRGLTIVPLSLYFKDGRAKV
EIALAKGKKTYDKRHSLAEKQATREVQREIGRRAKGMDR
>GENE 16 11622 - 12494 559 290 aa, chain +
VVSLRDEDVPAVARSLGLPGIFDVHVHFMPEPILKKVWAYFDAAGPLLGRPWPITYRLSD
DERVARLRAMGIKHFSALSYAHKPGIAPYMNDWTRDFAAATPEALWSGTFYPEPEAATYV
PDLIAQGVQIFKAHLQVGNFAADDPLLDPVWASLADSLTPVVLHAGSGPAPGAHTGPDGV
AAVLQRHPDLSLIIAHLGAPEFEEFFGLAEKYVNVRLDTTMAFVDFFDAPFPPHLMPQVL
DLQPKILLGTDFPNIPYHYAHQIEVLQQLGLGDDWMRDVLWNNAAGLFHL
>GENE 17 12532 - 13056 260 174 aa, chain +
MNIAKEVDLSGGVLVGHDGSRFANKALSWALEYAGAFGHDVTVVRAWVMTTAPRPKTWES
GYVPPLADFAAATLETLESDIAPLRGEFADIAVSCQAVHGSAAKLLLEASARADLLVVGA
RGRGGFLGLSLGSVSEKLARFAPSTVVVVRGDDDDPTPAADIEYDGSVDFDRSA
>GENE 18 13587 - 13742 89 51 aa, chain -
MLATSAVAAAAMAAGLKERETDMASPMTGTLPDPSTLCAAWENVNSTQEST
>GENE 19 13675 - 15549 1136 624 aa, chain +
MSVSRSFKPAAIAAAATALVASMLLALGTPSSASQPPQAQTTHKGHQGPQPQAVARGAGG
AVSSVDANASKIGVEVLRKGGNATDAAVAMASALGVSEPYSAGIGGGGYFVHFDAKSGSV
ETIDGRETAPAGITHDAFIDPDTGKPYPFTPELVTSGVSVGVPGTLATWDEALDRWGSTS
LRKALKPSIQLARKGFTVDETFRNQTLDNAERFAAFPSTAKIYLPGGDAPKVGSRLRNRD
LAKTLHLIGKRGPKAFYRGRLAEEIAQTVQRPPKSEDTDLPVPPGSMTVRDLAKYKVVDR
APTKVDYRGFNVYGMPPSSSGGTTVGEALNIVENFQLDKGDVAQALHLYFEASALAFADR
GAYVGDPDYVDVPTKRLLSQSFADSRACNIDPDQAAVKPVQAGALSGSDCTTQNHDEAAD
TENVSTTHLSVVDRRGNAVAYTLTNEQTGGSGIVVPGRGFLLNNELTDFTAVYDAKDPNR
IEPGKRPRSSMAPTIVTKDGDVRLVIGSPGGSTIITTVLQILINRIDLDMSLPAAIAAPR
ASQRNTEAVTAEPEFIEQYRDVLEPYGHILKPSGDAFTSLAEIGAAAGVEVDRRGNMTAA
AEPVRRGGGTAAVVCTAARRHSCR
>GENE 20 15613 - 15903 257 96 aa, chain +
MSALSEQEINERLEKHRDWTVEDGALHREFTFEGFPAAMAFMVQASRQIDAMNHHPEWSN
VYNKVDVRLTSHDEGGITDQDFTLAGIFDELGSPPA
>GENE 21 16082 - 16987 431 301 aa, chain +
MSPLTGLTMSEGRPDRPIVVTKIDNTASANPQHGVNKADLVVEELVEGGLTRLAAFYYSN
TPTHVGPVRSARATDIGIASPVNAELVASGGAPKTNKRIKAAGIKFHSEDAGAKGFSSDP
AKSRPYNRAINVQTLLKGRNATKIPGPYFTWASKKSTDKKSDSSAAPSAKASTTPKKATS
AAVRFSPSSTTQWGFKGGKWSRTNGISEKEFKADTMVVLFSKVGDAGYRDPAGNKVPETI
FDGSGSMFLFHGDTVTEGTWEKKGLGSTITMKDKSGAPVGVEPGNVWIELVPKGDGNVSV
N
>GENE 22 16989 - 17693 429 234 aa, chain -
MRPAQKVLRTARKTTRRQDYSSSTKRALVKHATALFTDHGYAGTSLDEVVAAARVTKGAL
YHHFPSKLALFEAVFMRCQDDATAQIDKALKSSRDPWERAQIGLRTFLNTCQQPNYRRIC
LQEAPVALGHERWQEAERESSYGLVERIVSDLLDELGSEKELAETFGVIFYGAMRTAGEF
VADAADPVQASANVEMVIGSLLGGLRLTPPFGEGENGDDTGQAGVSTDAAENQK

>GENE 23 17831 - 18337 474 168 aa, chain +
MTWDAYNRRKEALREVLAVADSRRDTTAHQLIADNDSAREAFGSASELMLDVQMNWYQRL
SGQLDRALTDGADDLEELVITAWSDAAAQAPGARVLIDAGEYMPELQRALTNEAVLLART
AGIRSSEHDQAEAGAQLRSMAKAAVVEIPEIPDTPAGLFTRIRSALAA
>GENE 24 18408 - 19229 505 273 aa, chain -
LPEWPQATVRAGFPWLYDAHMAEDAASGREFTVDELAARAGMTVRNVRAYASRGLIAAPR
LEGRTGYYSNQHLQRLHLIRVLMDRGFTLASIEKNLLNTSSAIGDHALDLVDVLHSPAQE
EEPEVMSRDALAALAGVPRDDALIESLAQLGLAEWINVDEVTLLRPAIVRAGASAVRMGL
SPATVLAMLPLLQTPLRSIADEFVRSVRDEVWQPFADRGMPPDQWPAVLEVIESLLPVAA
QAVLALFREQLAQSVEQAMGEQIALASNNAIPQ
>GENE 25 19187 - 19735 283 182 aa, chain +
MENLRGRLLVATPAIDAGLFRRAVVLMLDHDRDGALGVVINRPLDSGVGEVLPQWADCVN
EPGCLFAGGPVAADSALAVGVLGSDDVPVGWRPMCGRLGLVDLEGPVEVSLALAGMRVFA
GYAGWSPHQLEEELAEGSWMVVPARDDDLMSAAPEDLWRRVLARQPGELRIWASYPDDPS
LN
>GENE 26 19800 - 20114 127 104 aa, chain +
VANAVKSQQTESQPSPIRPSGGSQTVLDERTKSVPEAGDHERFSHYVNKNELTEAMVMGT
SVIALCGKAWVPSRDPKRFPVCPECKEIWKTMKPGDDGDSSSDS
>GENE 27 20104 - 21807 885 567 aa, chain +
VTPSPARALRAWQSDAFAQYQRAQPRDFLAVATPGSGKTTFALTIAADLLHRRVVDRVVV
VAPTDHLKNQWAHAATRVGISLDPQMAGRGALSADFKGFAVTYAGLAANPTAFRIRIERS
RTFVILDEVHHAGDALAWGDAVLEACEPATRRLCLTGTPFRSDDNPIPFVRYEPGFDGIS
RSVADYAYGYGTALRDGVVRPVLFMAYSGDMHWRTRAGDEVAARLGEPLTKDVTAQALRT
ALDPEGSWIPTVLVAADRRLQEVRRHVPDAGGLVIASDQEQARAYAAILTGLTGIKPTLV
LSDQVGASKRIDEFATGNSTWMVAVRMVSEGVDIPRLAVGVYATTTSTPLYFAQAVGRFV
RTRRRGETASIFLPSVPFLLALASDLEIERDHVLGRPVKDEDDLMAAEIELLRAQESESA
SDALGDFKALGSDASFDRVLYDGGEFGHEGIVAAGSDEELDFLGIPGLLEPAQVADLLRH
ARSRRATKTAAKNAVIDDAATFERQSDLRRELNGLVGAWHHRTGQPHGVTHSQLRSTSGG
PPSAQATSVQLQARIDLLRTWALKESS
>GENE 28 21810 - 22757 641 315 aa, chain +
MGSMPLRTRIFSLLMNLISKPIEDVPDDQIPALRAGRIKLQDTRAGRFLFGAEDPGVSIA
TQTISPEGHEIVLRIYRPSAGTLPVVVNFHGGGWVQGNNGQSGWLASRVAAQAGVVVVSV
EYRLAPESPFPAAVEDSWAALRWVHDNAQSLDVDAERIAVMGDSAGGNLAAVVALLARDA
GGPALRLQVLIYPSVEMYEKWPSELRNAEAPVLTSKNMRAFSRIYLAGADGTVFTASPIR
AESHLGLPPALIQTAEFDPLLDNGVKYAEVLTAAGVPVTQTTYPGAIHGFVSLPGATTAA
PHALNEIVRSLREAF
>GENE 29 22760 - 23641 464 293 aa, chain +
VRRNYWQRRIAALDPEVDYEEIVRIVAHHEFPWDIQQALSFALFRTYAVPSIGRLLFETG
QFTTDTQRRHDDTVLILDAIASDGMESPGGRAAVRRMNDMHGSYAISNDDMRYVLSTFVV
MPSRWIEIYGWRAGTDGEQLADVRYYRRLGALMGIKDIPATFAEFADLMDSYERDHFGQD
AGGKAVAEATLDLFGSFYPWLPRRLMRVFSLSIMEPHLRAAFGFRSPPIAVVNASRAALK
LRAKVVRWLPARSTPKHGRDLREVKSYVGGFRVEDLGTFPAACPVAPADEPRA
>GENE 30 23625 - 24209 369 194 aa, chain +
MSPVLDHYVEWWRKDRRHRRLLRPINRMRLNRALAQHGAYARGIVHGEALDMLIDGRLQI
GKDAFLEPQVWLTGGETGRIIIGSGSFLNLGVMIAALDLVEIGDHCMIANGCVITDANHR
FDDLTRPVTWQGFDSKGPTRIGANCWLGANVVVTSGVTIGERCVIGANSVVTGEVPPFSV
VAGNPARVMRRAET
>GENE 31 24266 - 27118 1622 950 aa, chain +
MQTPKQSGSMRRAHRKRLLWSVGLAAGAVIMASVALPTLAEKAAYDKVLRGGDPLMNPEA
LKMAADGHNHDHSDPRTKNLVSRSLATGVKDPTTPAQAAASHEAVAKQRAEADPKIVPGD
APAARRDVPEDIYAMAQGCYAVQDAASGKWVNRAGSGYQAAASSMNEAEPFHFQATALGK
YLLFDSAKDFVARTSGDPTGAIFSDSVGQAAKASPEADWTVSKGGDLFQFSLGDPDSGMA
VDGGGTVGLGVATGFRLHTIEGCQEWPEAQTNVTGQPHKGISDMQEVRGYLDAHTHGMAY
EFLGGRLHCGKPWDAYGVEVALSDCPDHTATGGNGALVDAAMGGGVSHDPVGWPTFKDWP
APNSLTHEGTYYKWMERAWRGGQRLFVNLLVENGQLCKVYPLKKNSCDDMDSIRLQAKRM
HEFENYIDAQSGGPGEGWYRIVTDPYEARKVINDGKMAVVMGIETSVLFGCTAKADVPSC
SKQQIDDQLDAVYDMGVRQMELVNKFDNALSGVAGDAGTTGAAINGANFLETGSFWKMEK
CDDNGEGVEDREQVAAPSSGPQQDALFGSIKGVLDRLPIAVPVYGSGPHCNQRGLTDLGD
HTIRRMVDKKMLFDPDHMSVKGRVSSLDLLEELNYSGVISSHSWSTPDAYPRIYRLGGVV
TPYAGDSNGFVKKWKKHLDWADGRYYFGFGFGADINGLGAQGNPRGADVPDPVTYPFQGI
GGVTIDKQVSGQRVYDINKDGVSHYGLYPDWMQDLRKIDGNTIVDDMERGPEAYLQMWER

ADGVSNDACRDDRALKPVSAITSIKDGSSVQTVLEQAGQPHSRLGSTFSYCAKTDTGKSS
TVEVDFAANKVSAIKATATADDEPSAEPSGSAGPAKPDAPANANAAAPAAAAAGTDDTGE
AVSATGDETKSAADNGWMPGTGGPALGVILMALGMIAAGTTVMIRQRKRR
>GENE 32 27384 - 28430 826 348 aa, chain +
MVFTGLLVGAALGYAMQRGRFCVTGAFRDLWVNKNSRWFSAFMLVVAIQSVGVFALDSLG
VITLADKPFPWLATIVGGFIFGYSIVYAGGCATGTYYRSGEGLIGSWIALTGYAVFAAIT
KTGALAPLNDAIHKPVLETTTIHGALGISPWVLVVALVIGVAVWVRYHRSKPALVMATLP
PKHSGLRHIVAERPWTAYGSAVVIAIIAIIAWPLSSATGRNDGLGITTPSANLVNFLVSG
DTTLVDWGVYLVVGILIGSFIAAKASGEFRLRVPDAKTVVRSIFGGAGMGIGASLAGGCT
IGNAMVKTATFSIQGWVALAFMVLGTGAAAYQTILKKQPASERTLIDA
>GENE 33 28456 - 28680 273 74 aa, chain +
MAVILETAGQVCPFPLVEAKDAIGDLPVGEELVINFDCTQATDAIPRWAAENGYPVTNYE
RVGDAAWTITVRKA
>GENE 34 28805 - 29851 619 348 aa, chain +
MANLKLSVLDLVPVRTDQTTGDALAASVKLAQSADTLGFTRYWVAEHHNMLAVAASSPPV
LIAHLAAHTQRIRLGSGGVMLPNHAPLAVAEQFALLEAAHPGRIDLGIGRAPGSDPVTSM
ALRGPAGRGDADIQNFPQYLDDVVALMGTAGAKIALRGQDYVLRATPHAVSEPALWLLGS
SPYSAHLAAAKGLPYVFAHHFSGQGTEEALQMYRSEFKPSDQAAQPRTFLTVNAVVAETT
KEAQALALPNMQHMARLRTGAPLGAMDLVEDALDADMTPQQEAMVDAGLRRSVIGSPAEA
ARQITELATQFGVDEVMIHPVASVRRGAAADQALGRERTIELLAAELL
>GENE 35 29879 - 30367 343 162 aa, chain -
LFAQLLLWLSISSAVAVVAGIAATRLVYTDMLRDRRANGIQRTALARSYGAMFAERARDN
SAFVSAINGRLADRDRTIRELDGVIRLADKRAEVAIIRANDSAERLTVATARVGDLEEQL
EIRNSELLDELAPWHGAELETVIDLLGWEERGIVAAARSKQA
>GENE 36 30550 - 31101 473 183 aa, chain -
MAKTLIIVDVQNDFCEGGSLGVAGGAAVATNVAELIASGAYDIVVATKDHHIDPGAHFSD
HPDFVDSWPPHCVVGTDGEHLHTPLSADLFTETFLKGEYEAAYSGFEGKSVSGVALADWL
RREREVTAFDVCGIATDFCVRATAVDGARLGFDVTVLMDLTVAVSSDNLLSVRQMFTEVGV
TTG
>GENE 37 31104 - 32438 984 444 aa, chain -
LFYAVGVTYETSTTAPSTALLTDRYELTMLQATLADGTAERQSVFELFARRLPEGRRYGV
VAGIGRLLDALENFRFGDAELEFLRSAGVVDEQTLEWLSNFRFSGDIWGYPEGEVYFPGS
PVLIVESSFAEGVILETLFLSILNHDSAIASAAARMISAAGGRPCIEMGSRRTHESAAVA
CARAAYVAGFASTSNLQAGRDYGIPTVGTSAHSFTLVHDNERDAFAAQIKSLGKDTTLLV
DTYDVAEAVRIGVELAGADLGAVRLDSGDLVQMARTVRDQLDALGATGTRITVTSDLDEY
AIQGLAVAPVDGYGVGTSLVTGSGHPTCGMVYKLVARAGADGAMTSVAKKSKDKLSIGGR
KFALRKRNGRNVAQTEVLGIGAPSVSDGNDRSLLLQFVDGGTRVHHDTLDEARERLRRAL
KELPMQAMQLSKGFPAIETTYEGE

CLONE LD13

>GENE 1 327 - 1826 1022 499 aa, chain +
VLAPTRELALQVATAFESFAAQMPSVNVVAIYGGAPMGPQLKAIRNGAQVIVATPGRLVD
HLSRNGGLLSTIKFLVLDEADEMLKLGFMDDLEVIFNAMPDERQTALFSATLPASIRGIA
EKHLRNPQQVKIASKTQTVARIDQAHLMVHADQKVNAILRLLEVEDFDAMIGFVRTKQAT
LDIAAALEAKGYKVAALNGDIAQNQRERVIDSLKDGRLNIVIATDVAARGLDVPRITHVF
NIDMPYDPESYVHRIGRTGRAGRDGRALLLVTPRERRMLQVIERVTGQKVAEIQLPDAKV
ILQARIRRLTQDLAPRVKDKEKQNVELLAHLTTELNCSAEDLALALLAKTTEGQAFTLDG
VEREQPVLAPRGRDSRDRDDRGGRGRDDRGGRSEGGRGDRGEGRGGEGGRERRAPLALSE
GKVRCRTALGIRDGVAARNLLGAILNEGGLSREAIGRIQIRESFSLVELPEEGLENLLGK
LKDTRVAGKALKLRRYRED
>GENE 2 2014 - 2793 320 259 aa, chain +
MTLHICILEADDLHPALQESFIGFGQMFKQLFNTQDVAVDCQVFNVVRGEYPDSNQQFDA
YLVTGSKADSFASDPWIANLRTYLHQRFVQGDVLLGICFGHQVLALVLGGDTQRSNKGWG
LGVHRYRLEHKPTWLPSVTDEFQLLISHRDQVTALPKGAALLASSEFCENAAFMLGQQVL
CFQGHPEFTHDFSRSLLDIRQSIYCPDEYQAACQSLEHQHDGQAVAQWMLCFIKAAKEGR
NAISSEPNPATQRDSLCLA
>GENE 3 2781 - 4334 1149 517 aa, chain +
MLSLTVQELSQYTVALLLGIIVLGIALQQFIRAQKLQQQSADNQTALRDAQQQLHEQDTD
LQVLGSQQDQLSAQLQLREADYQQLKTEHSAVQQQLLQGRGIAEAAQAGYRELKEQQQVR
EQALARQQTVYLELQDEHQRLQQEYASLRGLSAQKEQHLLEQQQLLRDSREQLKLEFEQL
ATQIFDARGQALTQTSQQSLQAMLKPFREQIDGFRQKVEDIHHKDVQQQASLQQQLLQLK
ELNQQITQEAHELSTALRGQKKAQGNWGELILENVLERAGLQLGVDFAREVSFTTSEGRK
RPDAIVYLPQNKHLIIDAKVSLNAYLRYVNAEDEALRAQALREHVSAFSARVRELAERDY
AALPGINAPDMVVMFVPIESAFADAVRADEGLLQRAIEQNILIATPSTLLASLTIVRQLW
RFEEQSRSTAELAERASKVYDKLRIFLGSMDGIGNSLDRAQEAYRKACDQLVSGRGNLIK
QASDFQQLGVSVKTEIAQQWQDRARLELTHAEQDSAD
>GENE 4 4331 - 4777 204 148 aa, chain -
MQIFSSPWLIVAVVCLLGLATYAALLWRRVGLAEQQRQQQRAQQKAQRHDDLIILSEGFL
SEQMPWAEGCIRIKVILDHYDYELGMQPDYQVLHTVFSATENIPTHDAWRALSSAEKQPF
TQLLSELELQHKQESMRAVQQLLSHLKG
>GENE 5 4772 - 4975 88 67 aa, chain +
LHIKLRDTFYTFSGCRMGMGHCDQFLSAGYAQAHGEHDCNAGRVHEVLLLEHGRGDHAGS
QGRERVL
>GENE 6 4988 - 5485 217 165 aa, chain -
MSYLATIKKTSAYLIFITLLSGVGGCASSGSSGFKDPDVQLVDVELIHAKLLEQQFMLHF
RVDNPNSKSLPMRGMDYRILLNNTPLATGSNSQWLTVPAHDYAYFKIPVHTNLWRHMKVV
LRMLENPDQPIHYALHAEVKTGLMFSKKINILRHGDIIPGDYIRE
>GENE 7 5728 - 6294 114 188 aa, chain -
MLHPLYRTVCLFSTCLKTYNSLFKRFLLYGLPMTAISSCPCNPNIAYQDCCGRYHAGATA
STAQALMRARYSAYVTHNIEFIKSTSLPAQQEQLDMQAIAEWSKNSKWLGLEVLSETVAQ
DQRHATVEFIAHWHDAQGRQQHQETSLFIKPAEHRYFYDPNVPLKAERNTPCPCGSRLKF
KKCCASYF
>GENE 8 6293 - 8437 1496 714 aa, chain +
MLSDELKKTIQGAYTRFLDSKGLKARYGQRLMIAEVAKVLGNIAVDEEGKRSGDPAVVAI
EAGTGTGKTVAYSLAAIPVARAAGKRLVIATATVALQEQIVDKDLPDIIKNSGLHFTYAL
AKGRGRYLCLSKLDALSQHGEAEQATAQLFAEDGFHIEVDSDSEKLFSEMITKLASNRWD
GDRDSWPQVLEDVTWSRVTTDHTQCTGRHCANFQQCSFYKAREGMTKVDVIVTNHDMVLA
DLALGGGAVLPDPRETIYVFDEGHHLPDKAINHFAHFTRLRSTADWLGQVEKNLTRLLAQ
NPLPGEFGRLLEPVPEMAKSLRTQMQFMFNFCEQLAEFRPNTDPDSYEKPRYRFIGGVVP
KELSELGVELKKEFIKLTDIFSRVAELLKKAMDDEPGLGVPSHQAEEWYPLFGSLLARAQ
NNQELWTAFTAEDPEKSPPMARWLSLSDAGGAFDIDVNASPILAAETLRRHLWNVAHGAL
VTSATLTALGKFDRFNMRAGLPFESKKEVVPSPFKYAEAGVLRVPDLRADPRDSEAHTQA
IIRELPAILEGAQGSLVLFSSRRQMKDVFEGVDHDWRKRVLIQGGLSRQETLAKHKSQID
KGEASVLFGLASFAEGIDLPGAYCQHVVIAKIPFAVPDDPVEAALAEWIEARGGNPFMEI
AVPDASLRLIQACGRLLRNEQDTGSVTLLDRRLVTQRYGKAILNSLPPFRREID

>GENE 9 8466 - 9611 730 381 aa, chain +
MHIDGSSDSAFSTVLDVFTQLLEDPEQRGASLCVQVAGETVIDVWGGVIDRHAEQLWQRD
TLVNIFSSGKPVAAVVLLQMVAEGRLQLDTPLAEYWPEFAVQGKRQITLRHILSHQSGLS
AVLEPLAPEALFDWERMIAAVENTPPWWIPGDAHGYAPMTYGWMVGELIRRVDGCEPGEA
IAKRIAQPLQLDLYLGLTERELPLVGDVMRMKGVFADAASLRLMKAMGNEPQGMTAKAFA
NPSSMMSSTNKLEWRQMQQPAANAHSTARGLAGFYTGLLQGQLLDGEMLNEMLREHSSGM
DLTLHTKTRFGLGCMLEQVDELPASYSLGARSFGHPGAGGTLGCADPEREVSVAFVTNSL
GASVLVDPRAQKINAMLKKCL
>GENE 10 9702 - 10571 523 289 aa, chain -
MANNLMTAQLIDGKQIAADIRKNIAQQVQDRLSQGLRVPGLAVILVGNDPASEVYVAHKR
KDCEQVGFQSQAYDLPATTTQDELLNLIDTLNTDDSVDGILVQLPLPAHLDSSLLLERIN
PYKDVDGFHPYNVGRLAQRMPLLRSCTPKGIITLLEHTGVDLHGLDAVIVGASNIVGRPM
ALELLLAGCTTTITHRFTTNLEEHVRRADLVVVAVGIPNLVKGEWIKPGAIVIDVGINRQ
ADGKLIGDVGFDEAIKRAAWITPVPGGVGPMTRACLLENTLQACEHNEK
>GENE 11 11305 - 11526 96 73 aa, chain +
VHMLLAVFNWCASGQVRLFMCMLGYLLVAVGESLLEIAWVYHSAPYKTRIYAGFCLNTRL
FNSYGFNQVCCGV
>GENE 12 11610 - 12008 259 132 aa, chain +
MFAEHYLLVKNIHITLVLLSGSLFVLRGLWVLLAGSGSVLQKKVNRLSYVIDTGLLLAAF
ALLMILNYAPLSAAWLQAKLLLLVLYVVLGALAFRAKYSLSMRWLAYCAALLCFAGMYYS
ARLHQPFAGLLS
>GENE 13 12426 - 13064 391 212 aa, chain +
VTDAYLETVDLACERDWRLLFEHLQVSVRPKDMLQVSGPNGSGKTSLLRLMSGLMRPTAG
EVLIQGVSIQQKRNELASNLLWLGHAAGIKGLLTAEENLTWLSALHHGASREQIWQALAA
VGLAGFEDVPCHTLSAGQQRRVALARLYLENVPALWILDEPFTALDKQAVTQLELHLAEH
CNNGGMVVLTTHHPLQNVPDTFRELDLGQVVV
>GENE 14 13061 - 13732 443 223 aa, chain +
MSNVFALMLLRETRIMFRRPAELVNPLVFFAIVIALFPLAVGPESQLLQTISPGLLWVAA
LLAVLLSLDGLFRSDFEDGSLEQWVVSPHPLVVLVLAKVLAHWLYCGLALVMLSPLLALM
LGMPGDKIPTLLLSLLLGTPVLSLLGAVGAALTVGLKRGGILLALLILPLYIPVLILGSG
VIQAALQGLPTAGYLLWMATLTMLTLTLAPFAIAAGLRISVGE
>GENE 15 13834 - 14589 456 251 aa, chain +
MNWTWFHKLGSPKWFYEISGRWLPWLSVSAALLIAVGLVMGLAYAPADYQQGNSFRIIYI
HVPAAFLAQSTYISLAVAGIVGLVWKMKVADVALQQAAPIGAWMTVIALVTGAVWGKPTW
GAWWVWDARLTAMLILLFLYFGIIALGHAITNRDSAAKACAILAIVGVVNIPIIKYSVDW
WNTLHQPATFTLTEKPAMPMEMWLPLLIMTIGFYCFFAAVLLVRMRTEVLRRESRTRWAQ
AEVARQIGRRV
>GENE 16 14759 - 15226 468 155 aa, chain +
MNPIRKKRLIIIGAILLGVVATVALGLTALQQNINLFYTPTQIANGEAPQDARIRAGGLV
KKGSLTRSEDSLTVDFIVTDGDADTGIQYRGILPDLFREGQGIVALGRLNEKGVLIADEI
LAKHDENYMPPEVSSALEKTGMLKHYEDGQKESKK
>GENE 17 15223 - 17226 1281 667 aa, chain +
MSNAALYLPELGHLALILALCFAVVQSFFPLVGAWRGDHKWMSLGQPAAWGQFVFTLIAF
ACLTWAFMIDDFSVAYVASNSNSALPWYYKFSAVWGAHEGSLLLWVVILAGWTFAVAVFS
RQLPEEMLARVLSIMGLISIGFLLFLIMTSNPFERLLPQMPMDGRDLNPLLQDFGLIIHP
PMLYMGYVGFSVAFAFAIAALLGGKLDAAWARWSRPWTLVAWAFLGIGIALGSWWAYYEL
GWGGWWFWDPVENASFMPWLLGTALIHSLAVTEKRGVFKNWTVLLAIAAFSLSLLGAFLV
RSGVLTSVHAFAADPERGIFLLVFLLVVVGASLTLFVMRAPAVKSKVGFGFWSKETLLLI
NNIILVVATAMVLLGTLYPLVLDSLTGAKLSVGPPYFNALFVPLMGLLMFAMAVGMITRW
KNTPGKWLIKMLAPVLILSAVLSVIGSVLYRDFNAAVLALLFVCAWVLLASARDILDKTR
NKGLWRGMRSLTRSYWGMQIGHLGMVFMAIGVVLVSQYSDERDLKMAPGDSLEMAGYHFV
FEGAEHYEGPNYISDKGSVRIFEGEREIALLHPEKRLYIVQQMPMTEAAIDPGFTRDLYV
AMGEPLENGAWAIRVHIKPFIRWIWLGAFLTAFGGVLSATDRRYRVKVTKKVKDTLGLSA
QGKTTHV
>GENE 18 17219 - 17758 450 179 aa, chain +
MSKTRLVAVFAVLLGVVLFAAMAMFGIKNDPSELPSALVGKPFPEFSTHSVDDLGAVITR
EDLLGRPALVNVWATWCISCKIEHPVLNELSKQGVVIHGINYKDENPAALRWLEDFLNPY
QLNISDPKGTLGMDLGVYGAPETFMIDKKGIIRHKFVGVVDKRVWREKLAPLYQELLDE

>GENE 19 17755 - 18228 402 157 aa, chain +
MKRFIYALGLTLACFGTAQASIDTYDFATQAERERYRVLVEELRCPKCQNQNIADSDAPI
AMDMRDQIFKKLEKGETNEEIVGFLVDRYGDFVRYKPPVNSSTMVLWYGPAALLVFGFAM
VAIIVFRRRRATRTEQNDKQLSGDEQSRLSDILKQHK
>GENE 20 18243 - 19478 890 411 aa, chain +
MTQFWIYAVLLLLLALLLLLVPVLRGRKDQTEEDRTALNIALYEEHIAELEAQYVGGAMT
AEQLAEGRIEADRELLDDTDIGRPKQSANLGSALPLIAALLVPVVGLGLYFVWGSSDKVA
LTMSLSEQPKTAQEMIERLEETVRLQPESVDAWYFLGRTYMSEQRPKEAARAYERTIELV
GRQPDLLGQLAQASYFANGNRWNAELQGLVDEALAQDPNEATSLGLVGIAAFEDSRFQDA
VDAWTQLLKGIDPQDQSYQAIQAGIERARAAIGTSSPQPQVSSTAPAQTPATDAAAAGDY
KITVEIAISDELAAQVAASDTVFVFARAAGGAPMPLAAKRFTVAELPARIVLTDADSLMP
NLKLSSVDSIELKASISSGGDAMQAQWKSEPLAVDAADTEAKNTLLINQKN
>GENE 21 19608 - 19790 140 60 aa, chain +
MIRKLLTAEGYEKLKDEFTYLVRKHRPEITEIVSWAASLGDRSENADDNSMLSGGVIISI
>GENE 22 19801 - 20148 270 115 aa, chain +
LREIDRRIRHLTKLFDVAQAVSYDPVQEGKVYFGAWCELENDDGETLRFRIVGDEEVYGR
QDYISLKSPMAKACLGKSVDDEVTVQTPNGEMHWYIIKIEYNVEGDVESSAEPSE
>GENE 23 20581 - 21333 259 250 aa, chain +
MENYHESFKKYESALLECTKLSQECAGIPSPTSSHFYASLLFTKLCNCAHSIGRLAPKPD
QIGKDAHWDYSSVASLTRDLIECYLTFYYLCIDKCSSEEWNARWQLMNLHDHLSRVKMFN
ALGMDYEEKEEAKNVKNDVIEKLKSNKWFRKLSDKQQTHFLKGKNAFFKSQDEILTASGG
NVSDFRFKYIFASNHTHTFPMGFYRMADGNRGRGVESQVEIQYTGLCLEWVSEYLLKAKA
EFGGKFENQK
>GENE 24 21444 - 21968 324 174 aa, chain +
MVITKTPRLVLREFTSNDVGALVEILSDPEVMEFSTNGPCTEDDTRKFVDWCLDSYQEHG
FGQWAIVDSSSEAIIGFCGLSQVDLNGVQEVEIAYRLARTAWSKGLASEAAAAVLAYGFT
ECHIDSVIAIVADRHMVSARVAEKVGLKIDTLTKYRDWDVRIYRKSLALNSKIT
>GENE 25 22035 - 22460 204 141 aa, chain +
MSTNVLLSIAAVLNAIVALLHIGCIYYGATWYRFMGAGEGMAALAERGSIQPTIITSFIV
LVFSIWTAYALSGAGVISQLPLLRWVLSAITAVYLLRGLAGFFFYSNPLGRTPEFWLWSS
AICLTLGLVHLFGLKQVWAQI
>GENE 26 22548 - 23243 269 231 aa, chain +
MLHHKLPVRQIGASLADLAHSYIIKCPALLNDVRSLISQSSTDNRFSLSAISSSISAMEV
FLNEITQLGNGYKAHNHGMNNPLVQMATALENAEKQRKATFVKLQIAYKSLAGKGVKCGD
LTPFQKLKIAIDVRSELAHPKSSTLTISPNGIHLPQKEQKLINKLKSNGFNISNDPLDWE
RVVNTKEFSLWVYQAVISTMLLVFDAWPYANSIESFKELYSLKLFKPEDWE
>GENE 27 23324 - 24646 404 440 aa, chain -
LNTRTPSGDDVTQRSDMKPAFSNDLFEHVLQEDNLSAAWKRVRANKGAAGIDGMTIDEFP
AWVRSGNWKALKQQLVTGCYQPSPVRRVEIAKPDGGTRQLGIPTVTDRVIQQAIAQVLTP
IFDPDFSEHSFGFRPSRNGQQAVKQVQSIIKEGRRFAVDVDLSKFFDRVNHDLLMTRLGD
KVKDKRLLKLIKRYLRAGFIDNQLLGESRVGVPQGGPLSPLLANIMLDSLDKELEKRGHK
FARYADDFTIVVKSQRAGERVLRSISQYLQNRLKLVVNTDKSRVVKTNESQFLGFTFKAN
RIHWHPKTLLKFKQNVRKLTNRNWGVSMKYQLFKVSQYLRGWINYFGIASGYQHCVELDH
WIRRRVRMAYWRQWRKPRTKVGNLMRLGVHVQAAVACGITSKGPWRSSKTPGINQALSNA
YLKSQGLYELRDGWIKLHHS
>GENE 28 25468 - 25857 169 129 aa, chain +
LVDMKQRQRSAKAANTDLVFTHSEIKDCQAAVSSALTRIFLKLNPGDKNINIKALIKHLD
EIKSCSQSAKWDEINKKIDFVREDSQKLLKQEWDRVKKGEPAFIWAKRIALSVFFGAFSL
GGYILWSFI
>GENE 29 26022 - 26417 128 131 aa, chain +
MSSRAIQNLFACYSEQLKNIKQPLKNLKGIDALTHCRTRALGASYYRCKLNHAEIEQLHS
CRHRSCYVCAHKQRLEWIEKQKARLLNVPHFHVVFTLPHEYLPLWRYNEALFARILFKAS
QETLLQLLAQK
>GENE 30 27147 - 27626 275 159 aa, chain -
DILARLMNAHPEFRMAMKDGELVIWDSVHPCYTVFHEQTETFSSLWSEYHDDFRQFLHIY
SQDVACYGENLAYFPKGFIENMFFVSANPWVSFTSFDLNVANMDNFFAPVFTMGKYYTQG
DKVLMPLAIQVHHAVCDGFHVGRMLNELQQYCDEWQGGA

Appendix III

Nucleotide and amino acid sequences of lipolytic genes; *DEase*I, *DEase*II and *DEase*V and dWHy1

***DEase*I**

```
1          ATGGGATCAATGCCGCTACGGACACGAATCTTCAGCCTGCTGATGAACCTCATCTCGAAA
1           M  G  S  M  P  L  R  T  R  I  F  S  L  L  M  N  L  I  S  K

61         CCAATCGAAGACGTCCCTGACGACCAGATACCGGCATTGCGCGCGGGGCGCATCAAATTG
21          P  I  E  D  V  P  D  D  Q  I  P  A  L  R  A  G  R  I  K  L

121        CAAGATACCCGAGCCGGCCGGTTCCTCTTTGGCGCCGAGGACCCCGGCGTCAGCATCGCC
41          Q  D  T  R  A  G  R  F  L  F  G  A  E  D  P  G  V  S  I  A

181        ACGCAGACGATCAGCCCCGAAGGCCACGAGATCGTGCTGCGCATCTATCGGCCCTCCGCG
61          T  Q  T  I  S  P  E  G  H  E  I  V  L  R  I  Y  R  P  S  A

241        GGCACGTTGCCGGTCGTGGTCAACTTCCACGGCGGCGGATGGGTGCAGGGCAACAACGGC
81          G  T  L  P  V  V  V  N  F  H  G  G  G  W  V  Q  G  N  N  G

301        CAATCAGGATGGTTGGCGAGCCGCGTCGCGGCCCAAGCCGGTGTGGTGGTTGTTTCGGTC
101         Q  S  G  W  L  A  S  R  V  A  A  Q  A  G  V  V  V  V  S  V

361        GAATACCGGTTGGCTCCTGAAAGTCCCTTTCCGGCTGCTGTGGAGGATAGCTGGGCCGCC
121         E  Y  R  L  A  P  E  S  P  F  P  A  A  V  E  D  S  W  A  A

421        CTTCGTTGGGTGCATGACAACGCCCAAAGCCTGGACGTCGATGCCGAACGGATAGCCGTC
141         L  R  W  V  H  D  N  A  Q  S  L  D  V  D  A  E  R  I  A  V

481        ATGGGCGACAGCGCCGGCGGCAATCTGGCCGCTGTGGTGGCCCTGCTGGCCAGAGATGCC
161         M  G  D  S  A  G  G  N  L  A  A  V  V  A  L  L  A  R  D  A

541        GGCGGGCCGGCGTTGCGCCTCCAGGTGTTGATCTACCCGTCGGTCGAAATGTATGAGAAG
181         G  G  P  A  L  R  L  Q  V  L  I  Y  P  S  V  E  M  Y  E  K

601        TGGCCCTCGGAGTTGCGCAACGCCGAAGCCCCCGTCCTCACCTCCAAGAACATGCGCGCC
201         W  P  S  E  L  R  N  A  E  A  P  V  L  T  S  K  N  M  R  A

661        TTTTCGCGGATCTATCTTGCCGGCGCTGACGGTACGGTGTTCACCGCTTCTCCGATCCGG
221         F  S  R  I  Y  L  A  G  A  D  G  T  V  F  T  A  S  P  I  R

721        GCCGAGTCGCACCTCGGACTGCCGCCGGCTCTCATTCAGACCGCTGAATTCGATCCGCTG
241         A  E  S  H  L  G  L  P  P  A  L  I  Q  T  A  E  F  D  P  L

781        CTCGACAACGGCGTCAAATATGCCGAGGTGCTGACGGCCGCGGGCGTGCCGGTCACGCAA
261         L  D  N  G  V  K  Y  A  E  V  L  T  A  A  G  V  P  V  T  Q

841        ACGACCTACCCCGGAGCCATCCACGGTTTTGTCAGCCTGCCGGGGGCGACCACGGCGGCG
281         T  T  Y  P  G  A  I  H  G  F  V  S  L  P  G  A  T  T  A  A

901        CCGCATGCGCTCAACGAAATCGTCCGGTCGCTGCGCGAGGCATTCTGA
301         P  H  A  L  N  E  I  V  R  S  L  R  E  A  F  *
```

DEaseII

```
1          ATGCCTGTCCTACCAATCCCATCCGTGAATGCGCTGGTAAGCAAAACAGCTAACACTATA
1           M  P  V  L  P  I  P  S  V  N  A  L  V  S  K  T  A  N  T  I

61         AAAAATTTGCGCAGCCGAGCAAGCACAAGCTCATCAGAGCTTAACCTCTTATCAACCAGC
21          K  N  L  R  S  R  A  S  T  S  S  S  E  L  N  L  L  S  T  S

121        CCGTTAAATCATCGCAGCGCTTCTAAGAATATTGATAGAACCACCGCGCGTTATAAGAAG
41          P  L  N  H  R  S  A  S  K  N  I  D  R  T  T  A  R  Y  K  K

181        CACACACTGCATTATGTACTAAAAGCTCTAGGCTATCTGCCCACACCTTTCCTTGAAAAG
61          H  T  L  H  Y  V  L  K  A  L  G  Y  L  P  T  P  F  L  E  K

241        TTTAATAGCACGTTGCACGGTCCCAATACTAAGCAATATCTCCATGCGGATGCGCACCTA
81          F  N  S  T  L  H  G  P  N  T  K  Q  Y  L  H  A  D  A  H  L

301        CGGCTGATTTTGGCATTGAGTAATAAGCTTAAACAGCCGCTAGATATCGAAAAACTGCCT
101         R  L  I  L  A  L  S  N  K  L  K  Q  P  L  D  I  E  K  L  P

361        ATTCTGCGCCAAAAGTTTGCAGCCGATACGGTCGCTATGCAAGCACCAGCGGTATGGGAT
121         I  L  R  Q  K  F  A  A  D  T  V  A  M  Q  A  P  A  V  W  D

421        CAGTCAGATAGCACTAGCGTACGCTGGCAAGATAAGGCCATAATGAATGCCGATGACGGT
141         Q  S  D  S  T  S  V  R  W  Q  D  K  A  I  M  N  A  D  D  G

481        GAGATGACGGTACGTTGCTATCAGCATTATTTACAAAATGAGGCTGATTTAGATATTAAT
161         E  M  T  V  R  C  Y  Q  H  Y  L  Q  N  E  A  D  L  D  I  N

541        AAACAAACCACTAATCAACAAACCGTTAACCAGAAAAATACTGATAAAACAGTGATGCTG
181         K  Q  T  T  N  Q  Q  T  V  N  Q  K  N  T  D  K  T  V  M  L

601        TTCATTCATGGGGGTGGGTTTTGTATTGGAGATATCGACACTCATCATGAGTTTTGTCAT
201         F  I  H  G  G  G  F  C  I  G  D  I  D  T  H  H  E  F  C  H

661        GCGGTGTGTACGCAGACAGGCTGGTCAGTGGTCAGTGTCGACTATCGCTTGGCACCAGAA
221         A  V  C  T  Q  T  G  W  S  V  V  S  V  D  Y  R  L  A  P  E

721        CATCCGGCACCAACGGCGCTTAGAGATTGTCTGAGCGCTTATGCTTGGGTGGCTGAGCAT
241         H  P  A  P  T  A  L  R  D  C  L  S  A  Y  A  W  V  A  E  H

781        TGCCATACTTTGGGTGCATTGCCATCACGTATTGTATTGGCAGGTGACAGTGCTGGCGGT
261         C  H  T  L  G  A  L  P  S  R  I  V  L  A  G  D  S  A  G  G

841        TGCTTGGCCATTTCGGTCGCGCAGCAAGTGTCCGCGCCCGATGCGTCGCAGTGGCTGAAC
281         C  L  A  I  S  V  A  Q  Q  V  S  A  P  D  A  S  Q  W  L  N

901        TTGGGATTGGATAATGACAAAATTACTCAGTTGCTACAAAGTTTACCGCGTCCATTAGCG
301         L  G  L  D  N  D  K  I  T  Q  L  L  Q  S  L  P  R  P  L  A

961        CAGTTGCCCTTGTATCCGGTGACCGATATCGAAGCTGAACATCCGAGCTGGGCATTATAT
321         Q  L  P  L  Y  P  V  T  D  I  E  A  E  H  P  S  W  A  L  Y

1021       GGTCAAGGGTTACTGCTGGATCATAATGATGTCGAGGTATTTGACGCCGCTTATATGCAG
341         G  Q  G  L  L  L  D  H  N  D  V  E  V  F  D  A  A  Y  M  Q

1081       CAAAGCGCTATCGCTCAGTCGCATCCACTGGTCTCGCCCATGTATGGTGACAATACGCAA
361         Q  S  A  I  A  Q  S  H  P  L  V  S  P  M  Y  G  D  N  T  Q

1141       ATGTGTCCCAGTTATATCGTTGTAGCTGAACTGGATATTCTGCGGGATGAGGCGCTGGTC
```

```
381         M  C  P  S  Y  I  V  V  A  E  L  D  I  L  R  D  E  A  L  V

1201        TATGCTGAGGAATTACAAAATAAAGGCATCAAAGTTGAGACTCATACCGTCCTTGGTGCC
401          Y  A  E  E  L  Q  N  K  G  I  K  V  E  T  H  T  V  L  G  A

1261        CCACATGGCTTTATCCATTTGATGAGTGTCCATCAAGGACTTGGTGACGAGACAGATTAC
421          P  H  G  F  I  H  L  M  S  V  H  Q  G  L  G  D  E  T  D  Y

1321        ATTATTAATGAGTTTGGCAGCTTTGTACGCCAACTGCTTACCAGAGATTTTAACCAGCCT
441          I  I  N  E  F  G  S  F  V  R  Q  L  L  T  R  D  F  N  Q  P

1381        GAGCCTGTCCTGTGA
461          E  P  V  L  *
```

DEaseV

```
1           ATGCATATTGATGGCAGCAGTGACAGCGCATTTAGCACTGTTCTTGATGTGTTCACGCAA
1            M  H  I  D  G  S  S  D  S  A  F  S  T  V  L  D  V  F  T  Q

61          CTACTTGAAGACCCTGAGCAGCGCGGTGCATCTTTGTGCGTACAAGTCGCAGGGGAGACA
21           L  L  E  D  P  E  Q  R  G  A  S  L  C  V  Q  V  A  G  E  T

121         GTCATTGATGTCTGGGGTGGGGTGATTGACCGCCATGCCGAGCAACTGTGGCAGCGCGAT
41           V  I  D  V  W  G  G  V  I  D  R  H  A  E  Q  L  W  Q  R  D

181         ACCTTGGTGAATATCTTTTCCAGTGGTAAGCCAGTTGCGGCAGTGGTCTTACTGCAGATG
61           T  L  V  N  I  F  S  S  G  K  P  V  A  A  V  V  L  L  Q  M

241         GTGGCAGAAGGGCGTTTGCAGCTGGATACACCGCTGGCAGAATATTGGCCAGAATTTGCT
81           V  A  E  G  R  L  Q  L  D  T  P  L  A  E  Y  W  P  E  F  A

301         GTGCAGGGTAAGCGGCAGATTACTTTGCGCCATATTCTTAGTCATCAGTCAGGTTTGTCT
101          V  Q  G  K  R  Q  I  T  L  R  H  I  L  S  H  Q  S  G  L  S

361         GCAGTGCTGGAGCCGCTAGCGCCGGAAGCGCTTTTTGACTGGGAGCGGATGATTGCTGCC
121          A  V  L  E  P  L  A  P  E  A  L  F  D  W  E  R  M  I  A  A

421         GTAGAAAATACCCCGCCGTGGTGGATACCCGGTGATGCGCACGGTTATGCGCCGATGACT
141          V  E  N  T  P  P  W  W  I  P  G  D  A  H  G  Y  A  P  M  T

481         TACGGGTGGATGGTTGGCGAGTTGATTCGTCGTGTTGATGGCTGTGAGCCAGGCGAAGCG
161          Y  G  W  M  V  G  E  L  I  R  R  V  D  G  C  E  P  G  E  A

541         ATTGCCAAGCGTATAGCTCAACCGTTGCAGTTAGATTTGTATTTAGGGCTGACTGAGCGT
181          I  A  K  R  I  A  Q  P  L  Q  L  D  L  Y  L  G  L  T  E  R

601         GAATTGCCGCTAGTGGGCGATGTGATGCGGATGAAGGGTGTCTTTGCTGATGCGGCCTCT
201          E  L  P  L  V  G  D  V  M  R  M  K  G  V  F  A  D  A  A  S

661         TTGCGCTTAATGAAAGCCATGGGTAATGAGCCGCAAGGTATGACGGCCAAAGCTTTTGCT
221          L  R  L  M  K  A  M  G  N  E  P  Q  G  M  T  A  K  A  F  A

721         AATCCATCATCGATGATGAGCAGTACTAATAAGCTTGAATGGCGTCAGATGCAGCAGCCA
241          N  P  S  S  M  M  S  S  T  N  K  L  E  W  R  Q  M  Q  Q  P

781         GCTGCGAATGCGCACAGTACCGCACGGGGTTTGGCAGGGTTTTATACCGGTTTACTACAA
```

```
261         A  A  N  A  H  S  T  A  R  G  L  A  G  F  Y  T  G  L  L  Q

841         GGGCAATTGCTTGATGGTGAAATGCTGAACGAGATGCTGAGGGAGCATAGTTCTGGTATG
281          G  Q  L  L  D  G  E  M  L  N  E  M  L  R  E  H  S  S  G  M

901         GATTTAACTCTGCATACTAAAACCCGTTTCGGTTTAGGTTGCATGCTAGAGCAGGTTGAT
301          D  L  T  L  H  T  K  T  R  F  G  L  G  C  M  L  E  Q  V  D

961         GAGTTGCCCGCCAGTTATAGCTTAGGCGCGCGCAGTTTTGGGCATCCCGGAGCTGGAGGC
321          E  L  P  A  S  Y  S  L  G  A  R  S  F  G  H  P  G  A  G  G

1021        ACATTAGGTTGTGCTGATCCTGAGCGTGAGGTGAGTGTGGCGTTTGTGACCAATAGTTTA
341          T  L  G  C  A  D  P  E  R  E  V  S  V  A  F  V  T  N  S  L

1081        GGTGCCAGCGTATTAGTGGACCCACGGGCGCAAAAAATTAATGCCATGCTGAAGAAGTGT
361          G  A  S  V  L  V  D  P  R  A  Q  K  I  N  A  M  L  K  K  C

1141        TTGTAG
381          L  *
```

---

dWHy1

```
1           ATGAGCTATTTAGCAACTATAAAAAAAACATCGGCATATTTGATTTTTATCACATTGTTA
1            M  S  Y  L  A  T  I  K  K  T  S  A  Y  L  I  F  I  T  L  L

61          AGTGGTGTTGGTGGCTGTGCATCGTCTGGTAGCAGCGGCTTTAAAGACCCTGACGTCCAG
21           S  G  V  G  G  C  A  S  S  G  S  S  G  F  K  D  P  D  V  Q

121         CTCGTTGATGTTGAACTGATACACGCTAAGTTGCTTGAGCAACAATTTATGCTGCACTTT
41           L  V  D  V  E  L  I  H  A  K  L  L  E  Q  Q  F  M  L  H  F

181         CGTGTCGATAACCCTAATTCAAAAAGTTTGCCAATGCGCGGTATGGACTACCGTATTCTG
61           R  V  D  N  P  N  S  K  S  L  P  M  R  G  M  D  Y  R  I  L

241         CTAAATAACACACCTTTGGCCACAGGTAGCAATAGCCAATGGCTGACAGTTCCTGCACAC
81           L  N  N  T  P  L  A  T  G  S  N  S  Q  W  L  T  V  P  A  H

301         GACTACGCCTACTTTAAAATACCCGTGCATACCAATTTATGGCGGCATATGAAGGTTGTT
101          D  Y  A  Y  F  K  I  P  V  H  T  N  L  W  R  H  M  K  V  V

361         CTGCGCATGCTAGAAAACCCTGACCAGCCGATTCATTACGCGCTGCATGCAGAGGTAAAA
121          L  R  M  L  E  N  P  D  Q  P  I  H  Y  A  L  H  A  E  V  K

421         ACTGGCCTTATGTTCAGCAAAAAAATCAATATCCTTCGTCACGGTGATATCATTCCCGGC
141          T  G  L  M  F  S  K  K  I  N  I  L  R  H  G  D  I  I  P  G

481         GACTATATTCGCGAGTAA
161          D  Y  I  R  E  *
```

Appendix IV

Publications

1. Marla Tuffin, Dominique Anderson, Cal Heath and Don Cowan. (2009). Metagenomic gene discovery: How far have we moved into novel sequence space? *Biotechnology Journal* **4**: DOI 10.1002/biot.200900235

Oral Presentations

1. Anderson, D. E., Meiring, T., Taylor, M., Tuffin, M. I., Cowan, D. A. Gene discovery in Antarctic Dry Valley soils. MERCK Young Scientist Award, Johannesburg. August 2009.
2. Anderson, D. E., Meiring, T., Taylor, M., Tuffin, M. I., Cowan, D. A. Metagenome sequencing and gene discovery. Cape Biotech Forum, Somerset west. March 2010.
3. Casanueva, A, Anderson, D, Tuffin, IM, Cowan, DA. Molecular adaptations to psychrophily: the impact of 'omic' technologies. Extremophiles-2010, Ponta Delgada, September 12-16.

Poster Presentations

1. Dominique Anderson, Mark Taylor, Marla Tuffin, Craig Cary and Don A. Cowan. Metagenome sequencing and gene discovery. SASM, Durban. September 2009.
2. Dominique Anderson, Mark Taylor, Marla Tuffin and Don A. Cowan. Metagenomic sequencing and *in silico* gene discovery. SASM, Cape Town. November 2011.

International travel

1. University of Bergen, Norway. 4th October -31st October 2009.
2. Antarctica. January 2011.