Repair eXtreme - DNA repair proteins from Antarctic extremophiles

A thesis

submitted in fulfilment

of the requirements for the degree

of

Doctor of Philosophy in Biological Sciences

at

The University of Waikato

by

Elizabeth Rzoska-Smith

THE UNIVERSITY OF
WAIKATO
Te Whare Wānanga o Waikato

2023

Abstract

The McMurdo Dry valleys (DV) in Antarctica are one of the harshest environments on Earth, with high levels of UV radiation, freeze-thaw cycles, low moisture, and nutrient content. Bacteria in this extreme environment have adapted to slow growth and low-nutrient conditions. It is proposed that bacteria from Antarctic DV systems may possess unique DNA repair enzymes and pathways that enable their survival in such extreme conditions. Exploring these DNA repair systems may offer valuable insights into microbial survival in extreme environments or even extra-terrestrial settings. To investigate this, DNA repair enzymes were identified from Antarctic DV microbial metagenomes using *in silico* analysis and *in vivo* characterisation of recombinantly produced proteins.

Four candidate enzymes were chosen for structural and biological characterisation. Three of these enzymes are characterised as LigB type ATP-dependent DNA ligases, with the typical arrangement of a DNA binding domain, adenylation domain and an OB-fold domain. These ligases can utilise both ATP and ADP nucleotide cofactors for the ligation of nick and mismatch DNA substrates, requiring the addition of magnesium or manganese metal ions. One LigB ligase stands out due to its interesting fusion with an N-terminally located nuclease domain, which resembles an extensively studied MBL-β-CASP domain found in all domains of life. The recombinantly expressed nuclease domain can co-ordinate several types of metal ions and shows nuclease activity against a diverse range of DNA substrates. Nuclease activity favoured single stranded DNA substrates, in a 5’ to 3’ direction, and was particularly active on substrates with abasic sites or 5’ flaps. The fourth protein was initially annotated as a hypothetical protein, however *in silico* analysis revealed that it exhibited homology to NucS nucleases identified in archaea and bacteria. This NucS homolog is a monomeric enzyme, made up of four domains, which is different from other proteins in this family. This enzyme shows a preference for single stranded DNA and has a broad range of nuclease activity on damaged and mismatched DNA substrates. It can utilise both magnesium and manganese metal ions for activity on DNA substrates. Overall, these findings suggest that the

described enzymes will play a role in DNA repair pathways, potentially in high damage environments like the DVs of Antarctica.

Acknowledgements

There are many people who I would like to give thanks to for supporting and encouraging me throughout my PhD journey. I would firstly like to say a big thank you to my primary supervisor Dr Adele Williamson. Thank you for your continuous support, knowledge, and guidance throughout this project. I am grateful for the opportunities you have given me not only to work on this project, but other learning experiences throughout my PhD studies. I would also like to thank my co-supervisor, Professor Ian McDonald, particularly for your guidance during the write up of my literature review. Your extensive knowledge of extremophiles was much appreciated.

To everyone in the C2 labs, past and present, thank you all for all for making my time in this lab so enjoyable. Thank you to everyone who gave me advice over the years, helped edit my thesis, or just let me vent about my difficult proteins, over coffee, I will really miss everyone. To the DNA modifying group, thank you all for your support and advice during my project, I have really enjoyed working alongside all of you. Ronja, it seems like only yesterday that you started in the lab as a summer student, always with a smile on your face. Thank you for all your help with this project. Also, a special thanks to Judith, or as most of us call you, lab mum, you have pretty much watched me grow up, from starting my masters to completing my PhD, and throughout this time you have been a constant friendly face and go to person when I need advice.

Family and friends thank you for all your love and support. To my flatmates over the years, thanks for all the cheering up and encouragement, especially during the multiple lockdowns, I honestly couldn't have asked for a people group of people to flat with during this time. A big shout out to mum and dad for everything you have done for me over my many years at university. I will be forever grateful for your unwavering support and love. Maddie, thanks for being such an amazing sister and for your support while I was writing up. And of course, Andy, you've been there throughout all of it, always kind and caring and continually patient, thank you for being you.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

SI (Systeme Internationale d'Unités) abbreviations for units and standard notations for chemical elements and formulae are used throughout this thesis. Other abbreviations are listed below.

3D	three dimensional
aa	amino acid
Abs	Absorbance
ADP	Adenosine diphosphate
APS	adenosine 5'-phosphate
ATP	Adenosine-5' triphosphate
Bp	base pair(s)
BLA	ST basic local alignment search tool
COG	cluster of orthologous genes
C-terminus	Carboxyl end of protein
DNA	deoxyribonucleic acid
ds	double-stranded
EDTA	ethylene diamine tetraacetic acid (disodium salt)
FPLC	fast protein liquid chromatography
GTP	guanosine triphosphate
His	Histidine amino acid
HMM	hidden Markov model
IMAC	immobilised metal affinity chromatography
IPTG	isopropylthio-β-D-galactosidase
Kb	Kilobase
kDa	kilo Dalton
LB	Luria Bertani
mAU	milli-absorbance units
min	minutes
MQ	milliQ ultrapure water
MS	mass spectrometry
MW	molecular weight
NAD	Nicotinamide adenine dinucleotide

Native PAGE	non-denaturing PAGE
NCBI	National Center for Biotechnology Information
N-terminus	amino terminus of protein
OD600	Optical density at 600 nm
PAGE	polyacrylamide gel electrophoresis
PCR	polymerase chain reaction
PDB	protein data bank
Q4	quenching buffer
RMSD	Root mean square deviation
RNA	ribonucleic acid
rpm	revolutions per minute
SD	Standard deviation
SDS	sodium dodecyl sulphate
SEC	Size exclusion chromatography
ss	single stranded
TAE	tris-acetate-EDTA
TB	terrific broth
TE	tris EDTA buffer
TEMED	tetramethylethylenediamine
TEV	Tobacco Etch Virus Protease
Tris	tris(hydroxymethyl)aminomethane
UV	ultra violet
V	volts
v/v	volume per volume
WT	wild type
w/v	weight per volume

1 Chapter 1

Introduction

1.1 The Antarctic McMurdo Dry Valleys

The continent of Antarctica is described as being one of the most chemically and physically extreme terrestrial environments on Earth (Cary et al., 2010). The Antarctic seasons are extreme, with the summers having endless sun light and a high UV flux, and the winters uninterrupted darkness (Greenfield et al., 2020). The air temperature of Antarctica can fluctuate between averages of −14.8 °C and −30.0 °C (Doran et al., 2002; Greenfield et al., 2020). The annual precipitation in snowfall is only 3.0-50.0 mm water-equivalent because of precipitation shadows produced by the mountains to the south (Fountain et al., 2010; Monaghan et al., 2005).

The McMurdo Dry Valleys of Antarctica comprise a number of west to east positioned, glacially-carved valleys found between the Polar Plateau and the Ross Sea in Southern Victoria Land (**Figure 1.1**) (Cary et al., 2010). These valleys make up the largest permanently ice-free land mass on the Antarctic continent (Cary et al., 2010), and are characterised by a rocky-sandy soil devoid of vascular vegetation (Fountain et al., 2010). The environment of the Dry Valleys are defined by extremely low temperatures, reduced moisture, oligotrophic soils and harsh UV conditions (Wynn-Williams, 1990). This environment may appear to be inhospitable to life, yet somehow a large number of microorganisms have adapted to survive and flourish in these harsh conditions (José et al., 2003; Khan et al., 2011; Wei et al., 2016).

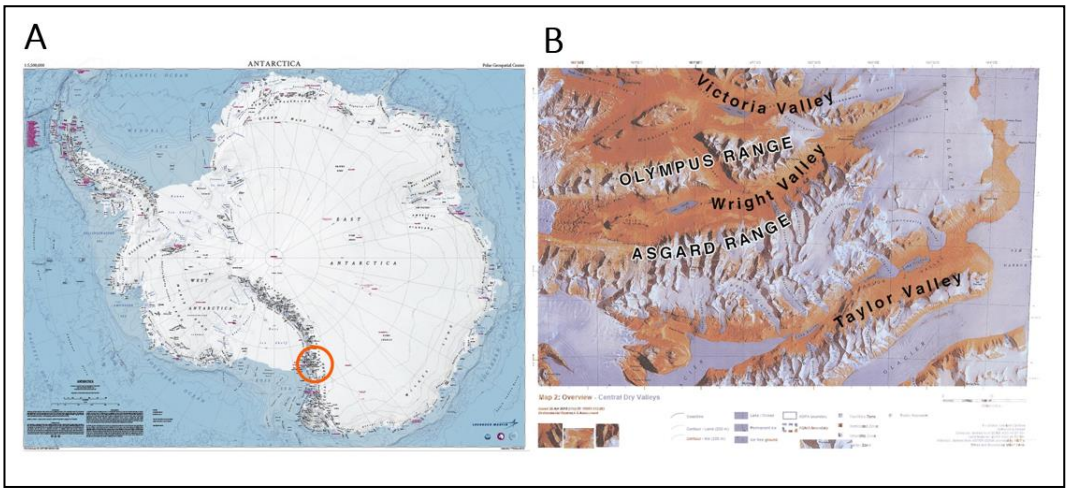


Figure 1.1. Location of the McMurdo Dry Valleys of Antarctica. (A) The McMurdo Dry Valleys (contained within red circle) are made up of a set of large valleys that cut through the Beacon Supergroup in Victoria Land west of McMurdo Sound. (B) The largest three valleys; Victoria, Wright and Taylor, lie between the East Antarctic Ice Sheet and the Ross Sea. Image adapted from (Greenfield et al., 2020).

1.2 The use of metagenomics for new microbial discoveries

In the past, microbial genome sequencing required the availability of a pure culture, limiting sequencing information to only culturable organisms. This requirement is no longer necessary thanks to the development of metagenomic strategies (Handelsman, 2004). Metagenomics emerged in the late 1990s and is defined as the functional and sequence-based analysis of a collection of microbial genomes, isolated from an environmental sample (Handelsman, 2005; Marco, 2010). Metagenomic sequencing studies have revealed previously unknown compositions and diversities of soil microbial communities across a number of soil environments without the need of cultivation (Thompson et al., 2017). Sequencing and functional screening of total metagenome libraries can also be used to reveal novel molecular pathways and proteins from unculturable organisms (Ferrer et al., 2005; Popovic et al., 2017). With the help of microbial metagenomic, several enzymes of industrial importance have been discovered from a range of different environments, such as β-glucosidase and exosialidase, which both have applications in the dairy industry and as a biofuel (Chen et al., 2014; Chuzel et al., 2018).

Previously, the belief persisted that Antarctica had limited bacterial diversity because many of these species could not be cultivated in a laboratory (Horowitz et al., 1972). However, through the application of metagenomics,

scientists examined soil samples from Antarctica and discovered that the region harbours a diverse array of bacteria, even at cold temperatures below -20 ºC (Jindal, 2020).

1.3 Microorganisms of the Antarctic McMurdo Dry Valleys

The Antarctic McMurdo Dry Valleys are unable to support higher plant and animal life and instead this environment is dominated by microbial communities that occupy oligotrophic mineral soils (Cary et al., 2010) and exposed rocks (De los Ríos, Cary, et al., 2014). The extreme environmental conditions such as limited moisture, thermal and UV stress reduce colonisation of exposed surfaces and encourage communities to develop within habitats of soil and rocks as a way to avoid these stressors (Pointing et al., 2014). These communities have been found to differ vastly from one another depending on whether they are found in exposed soil, beneath translucent rocks in soil contact as hypoliths, in crack and fissures inside rocks as chasmoendoliths, or in pore spaces within weathered rocks as cryptoendoliths (Pointing et al., 2009).

The soil communities are mainly populated by actinobacteria and other cosmopolitan soil bacteria (Stomeo et al., 2012), while cyanobacterial biofilms dominate the soil-associated hypolithic communities (De Los Ríos, Wierzchos, et al., 2014). The chasmoendolithic and cryptoendolithic settlement of porous rock requires the development of more complex biofilms, supporting cyanobacteria or chlorophyte algae and ascomycete fungi in lichen symbioses to dominate these communities (Yung et al., 2014).

1.3.1 Open soil microorganisms

Mineral soils of the Dry Valleys are exposed to a wide variety of environmental extremes including large seasonal and diurnal variations in temperature, low precipitation and atmospheric humidity, low nutrient availability, high levels of salinity and UV light and strong winds (Stomeo et al., 2012). These soils have been described as microbiologically distinct from all other soils worldwide in a metagenomics study (**Figure 1.2**) (De Los Ríos, Wierzchos, et al., 2014). The microorganisms that inhabit this extreme environment are thought to

have adopted a number of different physiological and adaptive mechanisms in response to these environmental conditions (Casanueva et al., 2010).

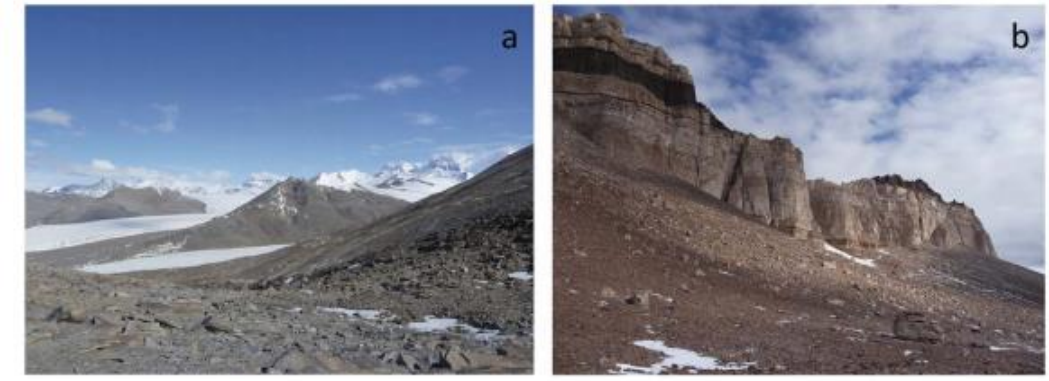


Figure 1.2. General landscape of McMurdo Dry Valleys. General landscape of McMurdo Dry Valleys. **a**. Miers Valley and **b**. University Valley. Image sourced from: (De Los Ríos, Wierzchos, et al., 2014).

Previously it was thought that both biomass levels and microbial diversity were low, as past methods replied heavily on culture-based techniques. Now with advances in technologies, phylogenetic surveys have indicated otherwise. Phylogenetic analyses of soil profiles from different sites around the Dry Valleys have shown that a wide variety of different phyla are present including, psychrophilic and psychrotolerant heterotrophs of the Actinobacteria, Acidobacteria, Proteobacteria and Bacteroidetes groups (Aislabie et al., 2008) as well as numerous genera of the photoautotrophic Cyanobacteria (Wood et al., 2008). These findings also suggest that the soils of the Dry Valleys support relatively low levels of eukaryotic microorganisms, and instead Bacteria are thought to dominate this community (Pointing et al., 2009).

1.3.2 Lithic associated microorganisms

Niches associated with rocks (lithic) offer physical and environmental protection to the microorganisms that colonize them (**Figure 1.3**). The microorganisms that occupy these niches are termed lithobionts, and may inhabit different ecological niches, from the surface of rocks (epilithic), to fissures and cavities within rocks (endolithic) and ventral rock surfaces (hypolithic) (De Los Ríos, Wierzchos, et al., 2014). There are two principle endolithic groups that exist in the McMurdo Dry valleys; chasmoendoliths that live in rock fissures and cracks, and cryptoendoliths found in structural cavities of porous rocks (Coleine et al.,

2020; Friedmann, 1982). Hypoliths are organisms or communities of organisms that live on the underside of rocks or at the rock–soil interface. Hypolithic microbial colonisation in the McMurdo Dry valleys represents a significant source of terrestrial biomass and productivity. These hypoliths have been shown to have significant eukaryal colonisation by fungi and mosses, with cyanobacteria mostly dominating this environment (Khan et al., 2011).

Figure 1.3. Lithic associated microorganisms**. a**. Granite boulder showing epilithic lichen colonization and snow deposits. **b**. *Umbilicaria aprina* Nyl. occupying protected areas of the rock surface. **c**. Lecideoid lichen colonizing weathered granite. Images sourced from: (De Los Ríos, Wierzchos, et al., 2014).

1.4 Survival in this environment

These harsh conditions are known to cause serious harm to microorganisms, especially DNA damage from UV radiation (Fuentes‐León et al., 2020). In the face of these extreme conditions, Antarctic microbes have evolved adaptations which are expressed at the ecological, physiological, metabolic, structural, and genetic levels, to counteract stressors associated with their environment. Adaptions include; metabolic activity at sub-zero temperatures, the ability to survive in metabolically inactive states under unfavourable conditions, and the presence of cell protection strategies (e.g. photoprotective mechanisms) (De Los Ríos, Wierzchos, et al., 2014; Friedmann & Thistle, 1993).

At the cellular and sub-cellular levels, studies have mostly focused on thermoadaptation. Cold adapted organisms can be divided into two overlapping

groups; psychrophiles and psychrotrophs/psychrotolerants (Morita, 1975). Psychrophiles grow optimally at less than 15 ºC (upper limit of 20ºC), while psychrotolerant organisms can survive at temperatures below 0 ºC, but grow optimally at 20-25 ºC (Morita, 1975). These organisms have developed several physiological adaptations to help them survive at these low temperatures; including; increasing the fluidity of cellular membranes, freeze protection, cold-shock responses and adaption of proteins and enzymes (Casanueva et al., 2010). Cell membranes are important for the regulation of cellular homeostasis as they control the function of transport processes. At low temperatures, several changes occur to the fatty acid profile of bacterial cell membranes for the maintenance of membrane fluidity. Typically there is an increase in fatty acid desaturation, a decrease in average chain length, an increase in methyl branching as well as an increase in the ratio of *anteiso*- to *iso*-branching (Casanueva et al., 2010; Chattopadhyay, 2006).

Microorganisms living in environments with extreme temperature lows, have evolved molecular processes to protect themselves from freezing, desiccation and hyper-osmolality. One of the most common methods of protection for cells is the increase of compatible solutes such as; glycine betaine, glycerol, trehalose, mannitol and sorbitol (Casanueva et al., 2010). An increase in these highly soluble polyhydroxylated compounds means there is a decrease in freezing point of the cytoplasmic aqueous phase and potentially a direct stabilisation of cytoplasmic macromolecules, such as enzymes (Borges et al., 2002; Welsh, 2000). A more specific mechanism for controlling the effects of cytoplasmic freezing, is the production of ice-nucleating proteins or antifreeze proteins, which bind to ice and inhibit ice-crystal growth (Kawahara, 2002).

While there is a lot of information in the literature about cold adapted microbes in Antarctica, information on the molecular and physiological responses to other Antarctic soil stress elements, such as UV radiation and desiccation, is lacking.

1.5 Types of DNA damage

1.5.1 Sources of damage

DNA damage can be subdivided into two main groups; endogenous and exogenous. Endogenous damage comes from within the organism. For example; reactive oxygen species, produced by reactions like oxidative deamination, can cause DNA damage. There are many more endogenous cellular processes that can result in DNA damage such as, oxidation of nucleotide bases and generation of DNA strand disruptions from reactive oxygen species, alkylation of bases, hydrolysis of bases, bulky adduct formation, mismatch of bases, due to errors in DNA replication, monoadduct damage resulting from a change in one nitrogenous base of DNA, and diadduct damage, resulting from a change in multiple nitrogenous bases of DNA (Friedberg et al., 2005).

In exogenous DNA damage, the source of damage comes from outside the organism, caused by external agents such as UV radiation from the sun, X-rays and gamma rays, viruses, thermal disruption and desiccation (Friedberg et al., 2005). UV radiation is part of the spectrum of electromagnetic radiation emitted by the sun. DNA is one of the key targets for UV-induced damage in a variety of organisms, ranging from bacteria to humans. During dehydration, DNA damage occurs through covalent modifications, crosslinking, and double-stranded breaks. Because DNA protection and repair mechanisms are slowed down, DNA damage accumulates causing cell death (Humann & Kahn, 2015).

1.5.2 Single strand damages

Discontinuities in one of the strands of the DNA double helix are referred to as DNA single-strand breaks (SSBs), which are frequently accompanied by damaged or mismatched 5’- and/or 3’- termini at the sites of the breaks. Such SSBs can be caused by a range of external and internal stressors (Hossain et al., 2018).

Highly reactive chemicals known as alkylating agents are capable of introducing alkyl groups into biologically active molecules, leading to the prevention of normal functioning. These agents can be found in the environment

or may be produced during metabolic processes. One example of an environmental alkylating agent is aflatoxin, a fungal toxin generated by *Aspergillus flavus* (Eubanks, 2005). Alkylation is a process that can cause DNA damage by transferring alkylating agents onto the nitrogenous base of DNA, resulting in altered base-pairing and potentially mutagenic effects. The formation of O6-methylguanine (**Figure 1.4**), an adduct resulting from exposure to methylating agents, can potentially cause mutations and even lead to cell death if left unaddressed (Kleibl, 2002).

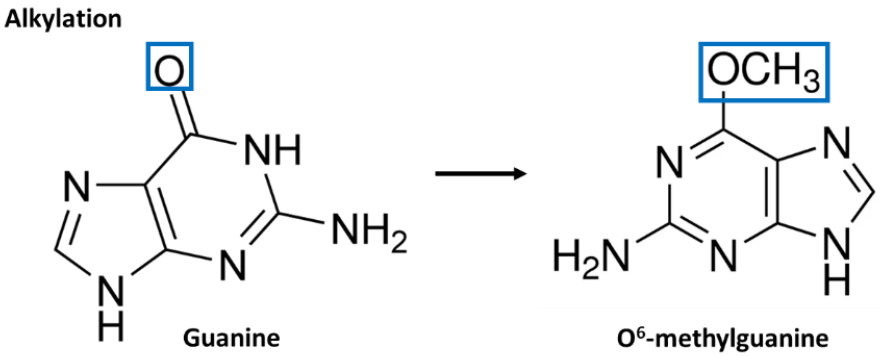


Figure 1.4. Schematic of the formation of O6-methylguanine from Guanine, due to alkylation.

UV is a major genotoxic agent in prokaryotic and eukaryotic microorganisms. The most frequent lesions induced by UVB or UVA in DNA, result from the covalent bonding of two adjacent pyrimidine bases, known as pyrimidine dimers. Pyrimidine dimers, like most bulky DNA lesions, are cytotoxic and can lead to cell death or be at the origin of mutagenesis (Douki et al., 2017; Kemp & Sancar, 2012). The most frequently generated photoproducts are *cis-syn* cyclobutane pyrimidine dimers (CPDs). They occur from the cycloaddition reaction between C5-C6 double bonds of adjacent pyrimide bases. Another type of pyrimidine dimer are pyrimidine (6-4) pyrimidone photoproducts (6-4PPs) (**Figure 1.5**) (Cadet et al., 2015; Kemp & Sancar, 2012).

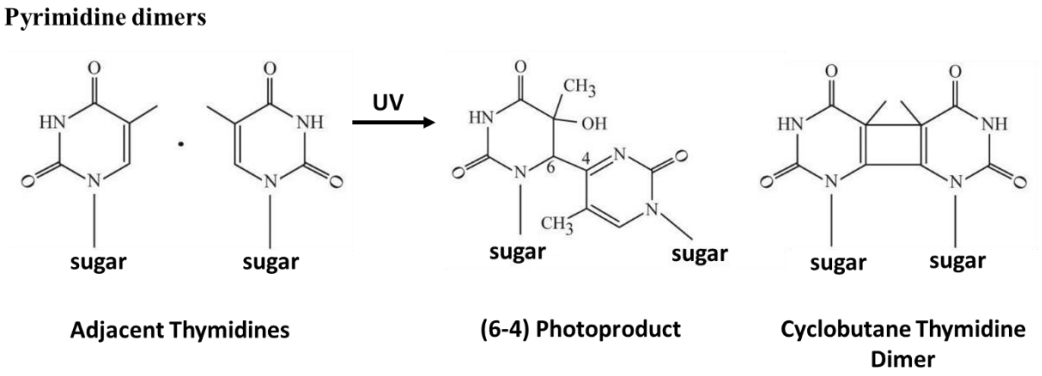


Figure 1.5. Diagram of pyrimidine dimer products, (6-4) photoproduct and cyclobutane thymidine dimer, caused by UV radiation. Image adapted from (Kemp & Sancar, 2012).

Living cells are subjected to a constant barrage of oxidative damage from a variety of sources, including environmental sources such as chemicals, UV, and ionising radiation, as well as endogenous processes like replication stress and metabolism. These factors can produce reactive oxygen species (ROS) that can harm essential cellular components such as DNA, proteins, and lipid membranes (Dupuy et al., 2020; Nikitaki et al., 2015). Intracellular ROS include the superoxide anion (O_2^-), the direct product of oxidases and respiration, while hydroxyl radical (*OH) and hydrogen peroxide (H_2O_2) are generated via Fenton reactions and processes, respectively. The generated O_2^- may be scavenged by NO and form peroxynitrite (-OONO). Exogenous ROS including singlet O_2^-, O_3 and *OH radical are produced during radiolysis of H_2O by ionising radiation (Hegde et al., 2012). Oxidative genomic damage induced by reactive oxygen species includes oxidized bases, apurinic/apyrimidinic (AP) sites and single-strand breaks (**Figure 1.6**) (Gonzalez-Hunt et al., 2018; Poetsch, 2020; Thompson & Cortez, 2020).

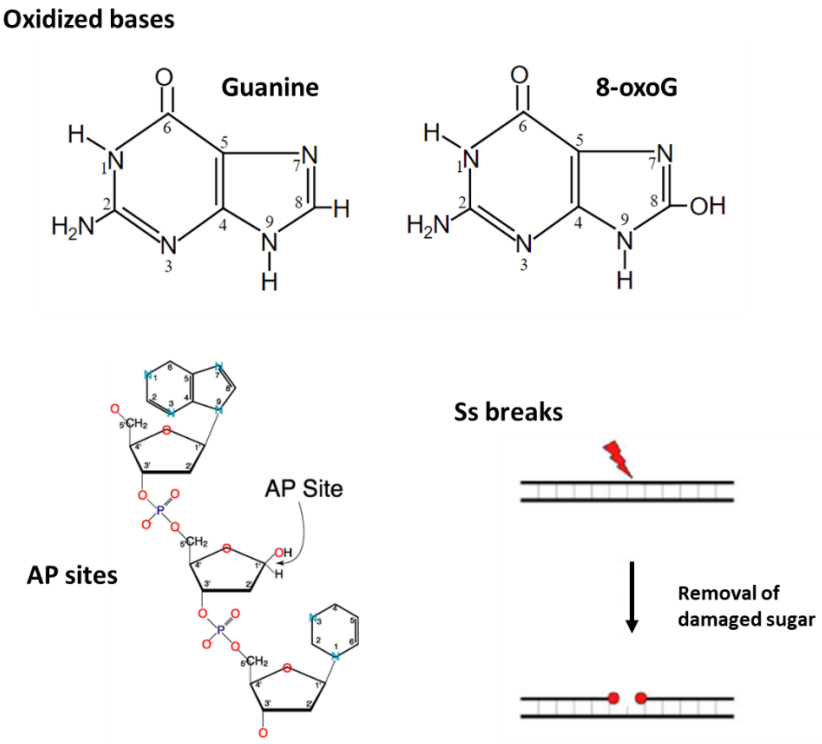


Figure 1.6. Figure showing results of oxidation damage to DNA. Common products of this damage are; oxidized bases e.g. 8-hydroxyguanine (8-oxoG), apurinic/apyrimidinic (AP) sites, due to loss of a purine or pyrimidine base and single stranded (ss) DNA breaks (Poetsch, 2020; Thompson & Cortez, 2020).

DNA damage caused by deamination of nucleobases is a prevalent type of hydrolytic damage. This can happen spontaneously through hydrolysis, as a result of nitrosative stress, or due to the activity of cellular deaminase enzymes (Shi et al., 2021). During deamination, the amino group is removed from a nucleotide base in DNA. This results in the loss of the exocyclic amino group from cytosine and adenine bases, creating uracil and hypoxanthine, respectively (**Figure 1.7**) (Khan, 2014). These bases have similar base-pairing properties to thymine and guanine, and if not repaired, can lead to point mutations during DNA replication (Hofreiter et al., 2001).

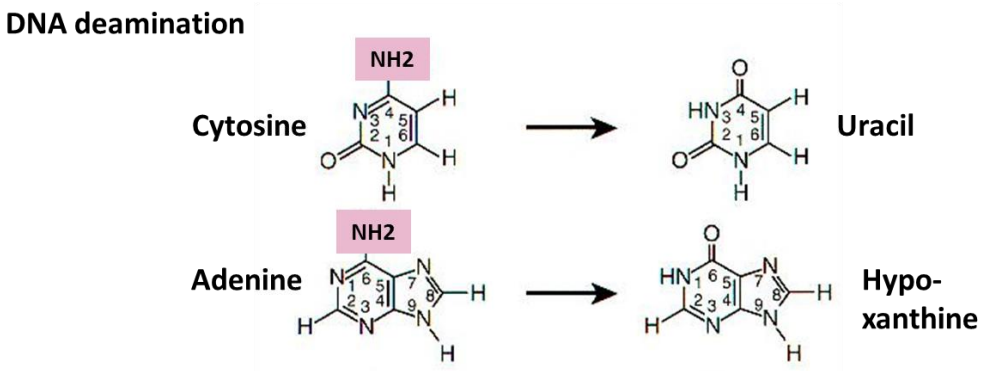


Figure 1.7. Figure showing DNA deamination. Hydrolytic DNA damage can result in the deamination of cytosine and adenine, leading to the creation of uracil and hypoxanthine bases, respectively (Khan, 2014).

Mutagens can originate from physical, chemical, or biological sources. Radiation and UV exposure are examples of physical mutagens, while chemical mutagens, such as ROS, can induce mutations either directly or indirectly. Many of these DNA damages described above, can cause high rates of mutations within cells, which may result in errors during replication, or prevent replication from occurring altogether (Watford & Warrington, 2017).

1.5.3 Double strand damages

Double-strand breaks (DSBs) pose a significant threat to genomic stability and integrity. They occur when both strands of the DNA double helix are severed, and can result from exposure to external agents such as UV, ionising radiation and certain chemicals, as well as from internal processes such as DNA replication and repair (Mourad et al., 2018). Bacteria, for instance, can be exposed to ionising radiation in their environment, which can cause breaks in the phosphodiester DNA backbone directly or indirectly by producing highly reactive hydroxyl radicals that react with DNA and produce single-stranded breaks. These SSBs, generated by either mechanism, can spontaneously convert into DSBs (**Figure 1.8**) (Cannan & Pederson, 2016; Đermić, 2015; Enderle et al., 2019). The presence of unrepaired or incorrectly repaired DSBs in DNA can be particularly dangerous for cells, as it can cause deletions, translocations, and fusions in the DNA, collectively referred to as chromosomal rearrangements, as described by (Mourad et al., 2018; Singh et al., 1999).

DNA crosslinking damages refer to the covalent connection of two nucleotide residues from the same DNA strand (intrastrand crosslink) or from the complementary strand (interstrand crosslink (ICL)) caused by crosslinking agents (**Figure 1.8**) (Enderle et al., 2019; Huang & Li, 2013). These agents can occur naturally or synthetically. Natural crosslinking agents include psoralens, mitomycin C, nitrous acids, among others. Synthetic ICL agents are a diverse group of bifunctional alkylators such as nitrogen mustard, carmustine, platinum compounds, and diepoxybutane (Lawley & Phillips, 1996; Noll et al., 2006). As

an exceedingly genotoxic and cytotoxic DNA lesion, the presence of an unrepaired DNA ICL can result in lethality in monocellular organisms (Lawley & Phillips, 1996; Magana-Schwencke et al., 1982).

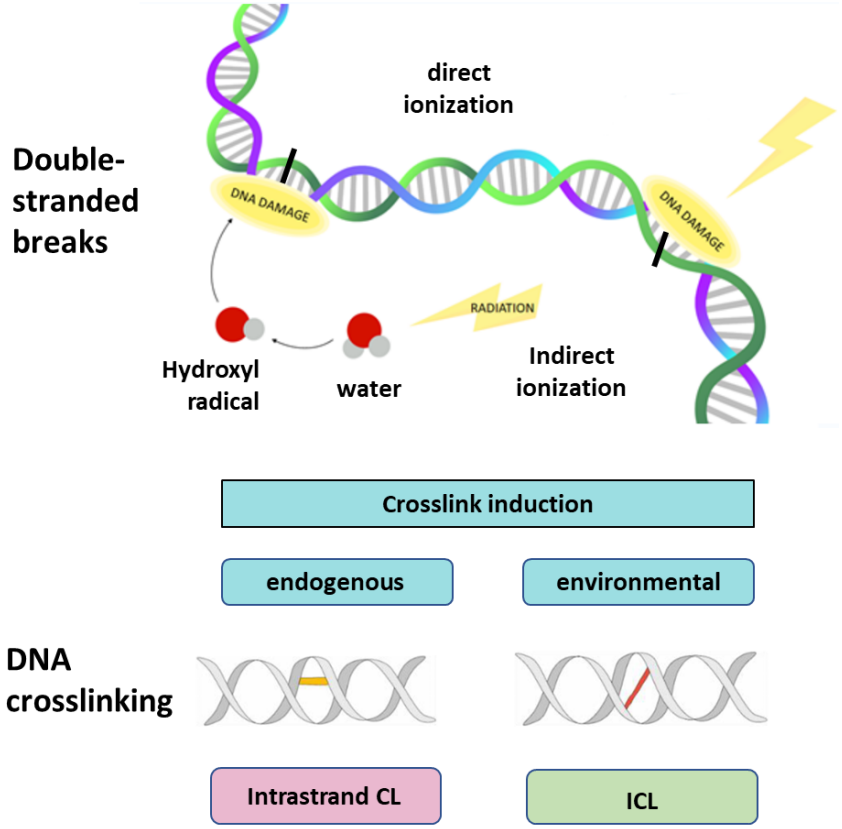


Figure 1.8. Diagram of double-stranded DNA damage. Diagram of double-stranded DNA damage. Double-stranded breaks can occur directly or indirectly from ionizing radiation. Diverse types of crosslinks form, due to different induction sources (Enderle et al., 2019).

1.6 Known mechanisms of DNA repair in Bacteria

Organisms, particularly those living in extreme environments, will be exposed to endogenous and exogenous DNA damaging agents. If the damaged section of DNA is not repaired, this could result in mutations, disease, and cell death. The cellular response to DNA damage includes processes that can detect the damage and signal a response to activate/recruit proteins involved in DNA repair and those that are directly involved in the correction of the damage by DNA repair mechanisms.

There are many different types of DNA damage, such as single or double strand breaks, base modifications, crosslinks, and mismatches. There are also

many different DNA repair pathways. These repair pathways are directed to specific types of damage, and the same DNA damage may be targeted by more than one pathway. The main players in DNA repair are mismatch repair (MMR), nucleotide excision repair (NER), base excision repair (BER), homologous recombination repair (HR) and non-homologous end joining (NHEJ). These pathways each require a number of proteins, which interact with one another to repair the damage. On the other hand, some DNA damage, like O-alkylated bases, can be repaired by the action of a single protein (Fleck & Nielsen, 2004). This action is more common with direct reversal pathways, which is explained below.

1.6.1 Direct reversal of base damage

Most damage to DNA must be repaired through the removal of damaged bases, followed by re-synthesis of the excised region. However, some lesions in DNA can be repaired through the action of direct reversal of the damage (Cooper & Hausman, 2000). These processes do not require a template, since the type of damage they counteract can only occur in one of the four bases. Direct reversal mechanisms are specific to the type of DNA damage and do not involve breakage of the phosphodiester backbone (Watson et al., 2004). Three major mechanisms of direct DNA repair have been identified to date: (1) photolyases reverse UV light induced photolesions, (2) O^6-alkylguanine-DNA alkyltransferases (AGTs) reverse a set of O-alkylated DNA damage and (3) the AlkB family dioxygenases reverse N-alkylated base adducts (**Figure 1.9**) (Yi & He, 2013).

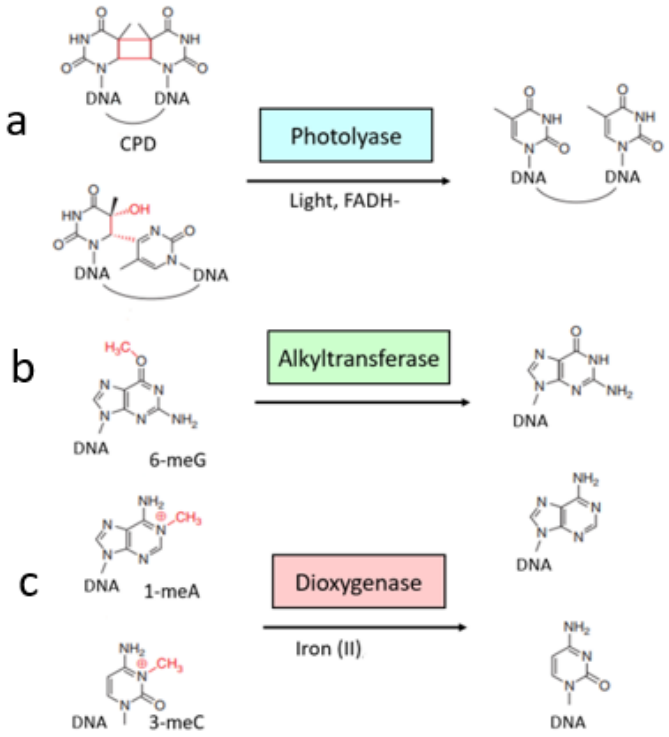


Figure 1.9. The three direct DNA repair pathways for base damage: (**a**) photolyases reverse UV light-induced photolesions; (**b**) O6 -alkylguanine-DNA alkyltransferases (AGTs) reverse a set of O-alkylated DNA damage; and (**c**) the AlkB family dioxygenases reverse N-alkylated base adducts. Image adapted from (Yi & He, 2013).

1.6.2 Repair of single strand damage

In single strand damages only one of the two strands in the double helix has a defect, meaning the complement can be used as a template to guide the repair. Several excision repair pathways exist where damage to one of the two paired molecules of DNA is repaired by removal of the damaged section of DNA and replacement with undamaged correctly paired nucleotides which are complementary to the undamaged DNA strand (Watson et al., 2004). These pathways are base excision repair (BER), nucleotide excision repair (NER) and mismatch repair (MMR).

1.6.2.1 Base excision repair (BER)

This type of repair pathway exists to repair DNA damage involving structurally non-distorting and non-bulky lesions produced by alkylation, oxidation or deamination of bases (Krwawicz et al., 2007). Damaged single bases or nucleotides are most commonly repaired by removing the base or the nucleotide involved (Fleck & Nielsen, 2004). Base excision repair (BER) can be

initiated by three different factors, the first being the enzymatic activity of a damage-specific DNA N-glycosylase, the second is by non-enzymatic hydrolytic depurination and the last is by single-stranded breaks (SSBs) with ends other than 3'OH and 5'-P (Krokan & Bjørås, 2013; McCullough et al., 1999; Wallace, 2014).

The main proteins involved in BER are a DNA glycosylase, an AP endonuclease, a dRpase, a DNA polymerase and a DNA ligase. These key proteins are conserved from prokaryotes to eukaryotes (Krwawicz et al., 2007). The first step in the BER pathway involves the identification and removal of a damaged base, by a DNA glycosylase. This DNA glycosylase cleaves the N-glycosidic bond between the damaged base and the deoxyribose sugar, generating an apurinic/apyrimidinic site (AP site) (Krokan & Bjørås, 2013; Krwawicz et al., 2007). Different DNA glycosylases are recruited depending on the type of base damage. There two classes of DNA glycosylases: those only having glycosylase activity or monofunctional (e.g., 3-methladine DNA glycosylase) and those that have an associated AP lyase activity or bifunctional (e.g., MutY DNA glycosylase). When monofunctional DNA glycosylases generate an AP site, this AP site is cleaved by apurinic/apyrimidinic endonucleases (APE). The APE cuts at the 5' of the abasic site, resulting in a 3'-OH terminus and a 5'-abasic sugar that is then removed by a dRpase (e.g., RecJ) generating a 5' phosphate end. The remaining nucleotide gap is then filled by specialised DNA polymerases (e.g., DNA Pol I) using their 5' to 3' exonuclease activity, they then synthesise the new strand using the complementary strand as a template. Finally the nick is sealed by a DNA ligase (e.g., LigA) (Fleck & Nielsen, 2004; Wozniak & Simmons, 2022). Bifunctional DNA glycosylases display intrinsic AP lyase activity, meaning they cleave the sugar-phosphate backbone 3' to the AP site. This generates a 5'-phosphate end and a 3'-α, β-unsaturated aldehyde on the 3' end, which requires further cleavage by an APE prior to the gap-filling and ligation (Krwawicz et al., 2007; McCullough et al., 1999) (**Figure 1.10**).

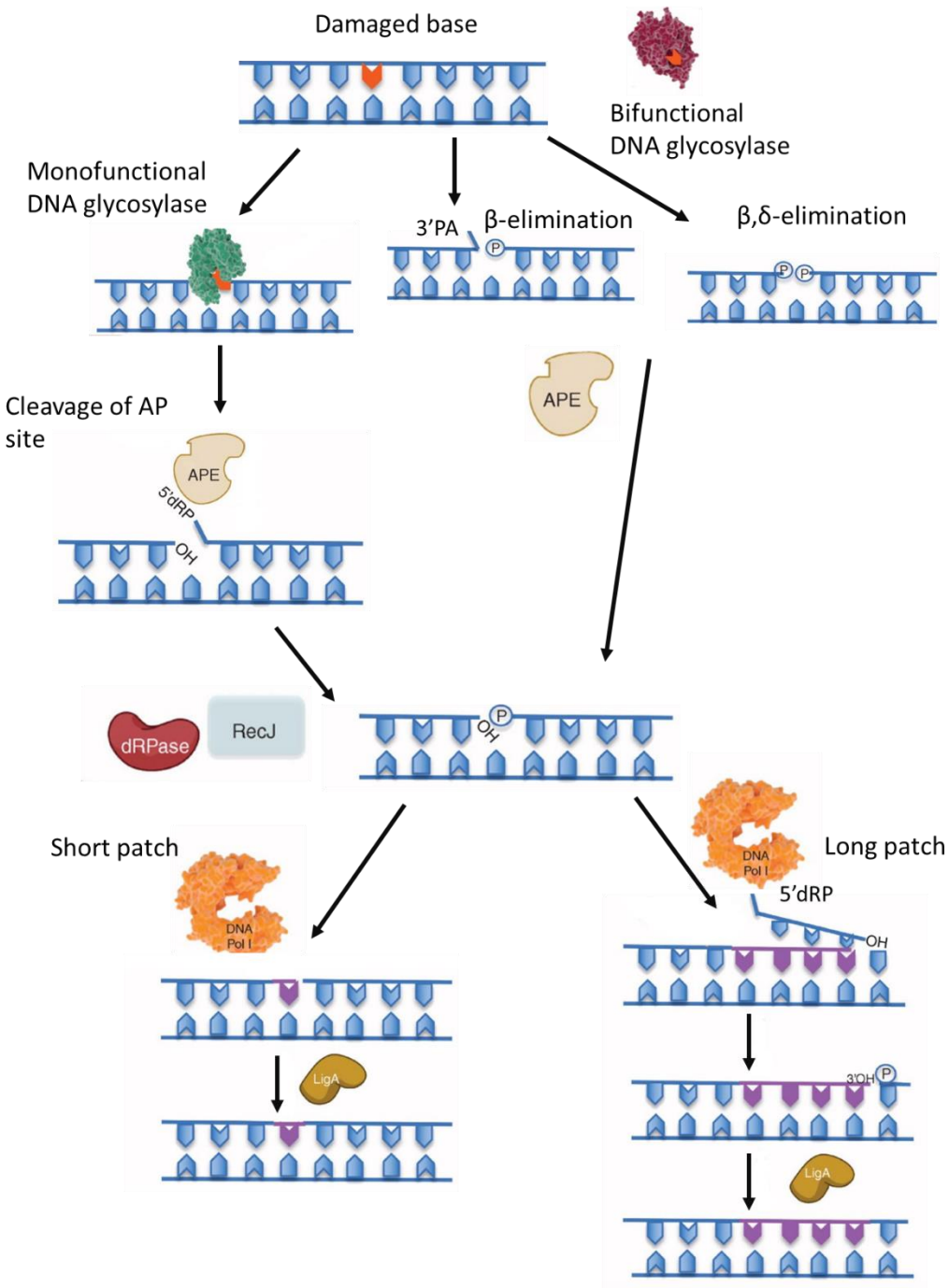


Figure 1.10. Schematic of bacterial base excision repair (BER) pathway. BER in bacteria is initiated by a DNA glycosylase. There are two groups of DNA glycosylases that are involved in BER: monofunctional and bifunctional. Monofunctional DNA glycosylases generate abasic (AP) sites, that are further excised by APE generating 3' OH and 5' dRP ends. The 5' dRP end is removed by a dRPase, generating a single nucleotide gap, to be filled in by a DNA polymerase. The remaining nick is then sealed by a DNA ligase. Bifunctional DNA glycosylases can cleave the AP sites and generate 3' PA (3'-α, β unsaturated aldehyde) and 5' phosphate ends. Some are even able to further process the 3' PA ends via δ-elimination resulting in a 3'-phosphate end. AP endonuclease further processes the 3' phosphate end to 3' OH end. This is followed by further processing by the same DNA polymerase and ligase, as discussed above (short patch). DNA polymerase I can also displace the strand and polymerise tracts of DNA longer than one base producing flapped substrates, that require further processing before ligation can occur (long patch). Figure adapted from (Krwawicz et al., 2007).

1.6.2.2 Nucleotide excision repair (NER)

Nucleotide excision repair (NER) repairs damaged DNA which commonly consists of bulky, helix-distorting damage, like pyrimidine dimerization caused by UV light (Reardon & Sancar, 2006). NER is highly evolutionarily conserved and is found in nearly all cellular organisms (Reardon & Sancar, 2006). In this pathway damaged regions are removed in 12-24 nucleotide-long segments in a series of steps which consist of damage detection, damage verification, incision, excision and DNA ligation (Kisker et al., 2013).

In prokaryotes, NER is mediated by Uvr proteins, which form a multienzyme complex (UvrABC). The subunits for this enzyme are encoded in the *uvrA*, *uvrB* and *uvrC* genes. This enzyme complex is capable of repairing various types of DNA damage (Grossman & Yeung, 1990; Reardon & Sancar, 2006). This repair pathway can be initiated in two ways. First, damage can be detected by UvrA, which is working with UvrB. Alternatively, the damage may first be encountered by an RNA polymerase (RNAP) that stops at the damaged site (**Figure 1.11**) (Kisker et al., 2013). A transcriptional repair coupling factor (TRCF) can dislodge the stalled RNAP, and recruit UvrAB machinery to the damaged site. The following steps are the same regardless of the initiation factor. The damaged section is passed between UvrA and UvrB, which separates the two DNA strands to identify the position of the lesion, initiating the release of UvrA. The remaining UvrB protein forms a tight scaffold on the DNA, in preparation for the arrival of UvrC protein. UvrC contains two nuclease domains and can cleave the phosphodiester bonds 8 nucleotides 5’ and 4-5 nucleotides 3’ to the damaged site. The post incision complex is then displaced by the dual action of UvrD (helicase II) and DNA polymerase I, which work alongside one another to excise the damage containing oligonucleotide and eventually fill in the resulting gap, using the complementary strand as a template (Caron et al., 1985; Husain et al., 1985). UvrB and UvrC proteins are turned over as a result. The final step is accomplished through the action of a DNA ligase, which seals the newly created repair patch (Kisker et al., 2013).

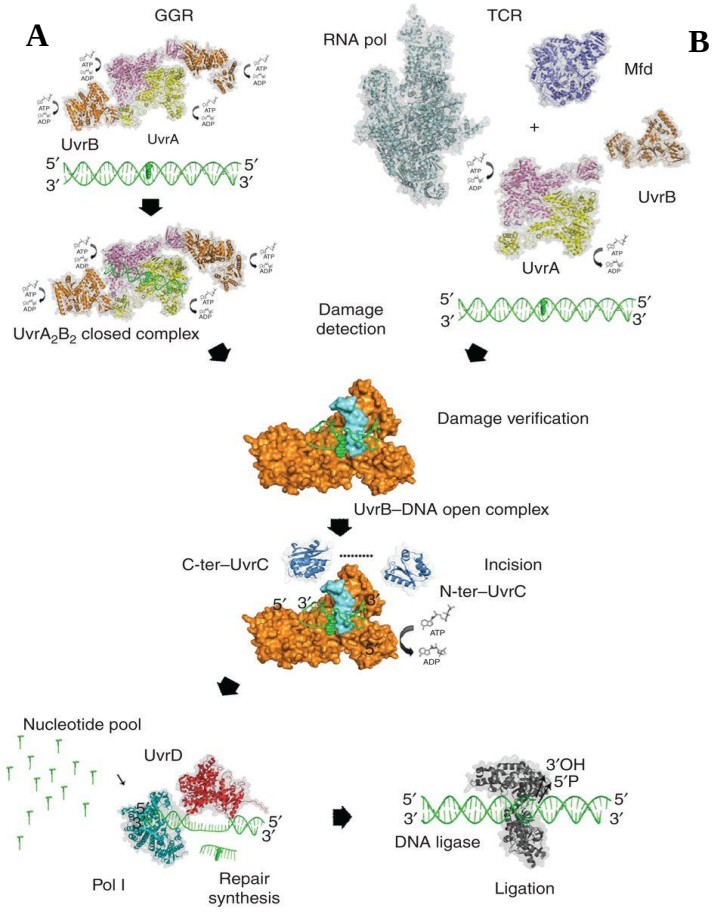


Figure 1.11. Schematic representation of the prokaryotic NER pathway**. A**) represents global genome repair (GGR). **B**) represents transcription coupled repair. Both mechanisms then converge into the same pathway and proceed with damage verification UvrB and 3′ to 5′ catalysis by UvrC. The cut strand is then removed, and the strand is repaired by DNA polymerase I and DNA ligase. Image sourced from: (Kisker et al., 2013).

1.6.2.3 Mismatch repair (MMR)

Mismatch repair (MMR) systems exist in essentially all cells to correct post-replicative errors that have escaped the 3'-5' exonucleolytic proofreading activity by replicative DNA polymerases, but this system can also recognise base damages (Stojic et al., 2004). Mis-incorporated bases are identified due to their failure to form Watson-Crick base pairs, while the base damage is identified due to a weakened base pairing as well as a slightly distorted helix (Dalhus et al., 2009). Base mismatches, small insertion/deletion loop mismatches and base damages, which might have arisen due to DNA damaging agents or through replication errors are removed at the start of this repair pathway (Fleck & Nielsen, 2004).

These systems consist of at least two proteins. One will detect the mismatch and the other will recruit an endonuclease to cleave the newly synthesised DNA strand near the region of damage (Berg et al., 2012). In most bacteria, the main players in this pathway are the Mut proteins: MutS, MutL and MutH (**Figure 1.12**) (Fleck & Nielsen, 2004; Jun et al., 2006). Some bacterial species are missing the MutH gene, so their pathways will vary slightly. Mismatched bases are recognised by MutS, while MutL interacts with and stabilises the complex. MutH nicks the non-methylated strand, enabling discrimination between the newly synthesised strand and the template strand. Other proteins involved in this repair pathway are: helicases like UvrD, endonucleases like RecJ and ExoI, DNA ligases, DNA polymerase III and DNA binding proteins (Fleck & Nielsen, 2004).

The MMR system is bidirectional, meaning nicking and degradation can occur from either the 5' or 3' side of the mismatch. Strand discrimination can be mediated by the β-sliding clamp, or it is achieved by recognition of nicks, gaps or free 3' ends that are present in the growing strand during replication. Towards the end of this pathway, the newly synthesised strand is degraded, which removes the mismatch. MMR is completed after DNA synthesis by the replication machinery and ligation of the remaining nick (Fleck & Nielsen, 2004).

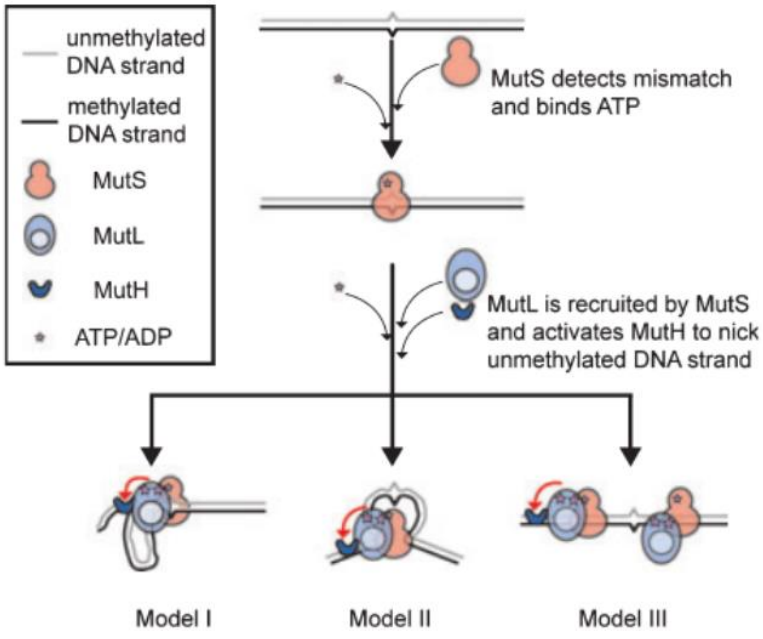


Figure 1.12. Models for the assembly of the DNA mismatch repair complex in a schematic drawing. A mismatch base is detected by MutS and ATP bound MutS recruits MutL. Image sourced from: (Jun et al., 2006).

1.6.3 Repair of double-strand breaks

Double-strand breaks (DSBs) result from disruption of the phosphodiester backbone on both strands of the DNA double helix (Pastwa & Błasiak, 2003). These DSBs may arise through DNA replication errors or after exposure to DNA-damaging agents, such as ionising radiation or radio-mimetic chemicals (Doherty et al., 2001). DSBs in the DNA helix can result in genome rearrangement or mutations in cells, threatening the continued survival of these cells (Doherty et al., 2001; Shuman & Glickman, 2007). Three mechanisms exist to repair DSBs: non-homologous end joining (NHEJ), microhomology-mediated end joining (MMEJ) and homologous recombination (HR), with NHEJ and HR being the major repair pathways (Shuman & Glickman, 2007).

1.6.3.1 Homologous recombination (HR)

In this repair mechanism, one or both of the DSB ends is degraded by an exonuclease to leave a 3' single-stranded tail which becomes coated with the RecA protein. The protein RecA (in bacteria) then locates and pairs the homologous sequences and promotes strand invasion (**Figure 1.13**). DNA polymerase then copies the sequence information from the homologous copy, extending from the invading 3' –OH end and causing formation of a displaced single-stranded 'D'loop on the undamaged strand (Shuman & Glickman, 2007). The resulting recombination intermediates are separated by the action of helicases and the Holliday junction resolvase, and ultimately correct the DSB by transferring a short segment of template DNA sequence to the original chromatid. The remaining single-strand nicks in the repaired duplex are then sealed by the replicative DNA ligase (LigA in bacteria) (Lusetti & Cox, 2002). In bacteria, the source of the homologous DNA that is used as a template in this repair process is typically a fully or partially replicated chromosome, meaning this mechanism is only effective in repairing damage in actively dividing cells.

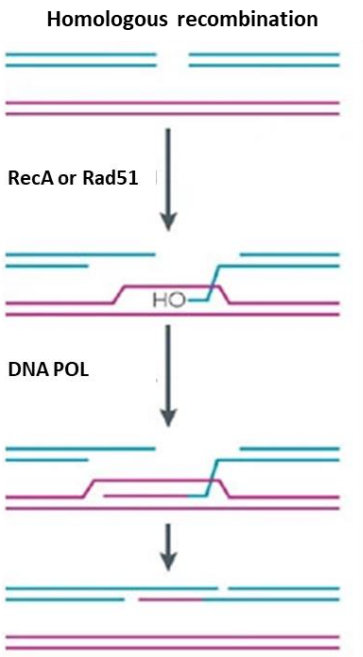


Figure 1.13. In homologous recombination (HR), the DNA duplex that sustains the double-strand break (DSB) (cyan) is resected at one or both ends by a 5' to 3' exonuclease. In homologous recombination (HR), the DNA duplex that sustains the double-strand break (DSB) (cyan) is resected at one or both ends by a 5' to 3' exonuclease. This generates a 3-OH single-stranded extension that invades the intact homologous sister chromatid (magenta) in a reaction that is catalysed by the bacterial RecA protein. Image sourced from: (Shuman & Glickman, 2007).

1.6.3.2 Non-homologous end joining (NHEJ)

In some species of bacteria, broken DNA ends can be joined directly following a repair mechanism that is analogous to the non-homologous end joining (NHEJ) pathway that is seen in eukaryotes (Lieber, 2010). NHEJ is a relatively simple process that does not require a homologous DNA template, so it can operate in situations where there is only one chromosomal copy available (Shuman & Glickman, 2007). The pathway of NHEJ involves the direct ligation of broken DNA ends after minimal processing such as removal of 3' phosphates, to make them compatible for ligation (Lieber, 2010). NHEJ is advantageous as it can take place any time during the cell cycle, but it is at a disadvantage when it comes to repair fidelity, which is low due to removal or addition of nucleotides during end processing (Lieber, 2010).

In bacteria, NHEJ uses Ku and a multifunctional DNA ligase to process and re-join broken DNA ends (**Figure 1.14**) (Shuman & Glickman, 2007). Ku is a dimeric protein complex that functions as a molecular scaffold and effectively aligns the DNA, while still allowing access of polymerases, nucleases and ligases

to the broken DNA ends to promote end joining (Aravind & Koonin, 2001; Shuman & Glickman, 2007). Once the DNA DSBs are juxtaposed by Ku, the ends can then be re-selected and sealed by a specialised DNA ligase, LigD in bacteria, that is unique to NHEJ (Shuman & Glickman, 2007).

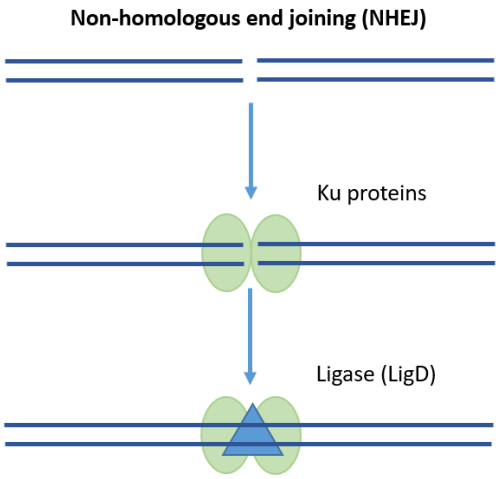


Figure 1.14. In non-homologous end joining (NHEJ), there is no requirement for a homologous sister chromatid. Rather, the DSBs are juxtaposed by the end-binding protein Ku (depicted as a pair of green spheres) and then sealed by a specialized DNA ligase that is unique to NHEJ: Lig4 in Eukarya or LigD in Bacteria. Image sourced from: (Shuman & Glickman, 2007).

1.6.3.3 Microhomology-mediated end joining (MMEJ)

Microhomology-mediated end joining (MMEJ) is a DSB repair mechanism that involves the alignment of microhomologous sequences (5-25 bp) internal to the broken ends before joining and is associated with deletions flanking the original DSB (McVey & Lee, 2008; Sfeir & Symington, 2015). In MMEJ, repair is initiated through end resection by the MRE nuclease, resulting in single stranded overhangs (Truong et al., 2013). These single stranded overhangs then anneal at microhomologies, which are described as short regions of complementarity between the two strands (McVey & Lee, 2008). After annealing, any overhanging bases (flaps) are removed by nucleases, such as flap structure-specific endonuclease 1 (FEN1) and gaps are filled in by DNA polymerase theta. This gap filling ability of polymerase theta helps to stabilize the annealing of ends with minimal complementarity (Sfeir & Symington, 2015). This is then followed by recruitment of a DNA repair protein and a DNA ligase to the site for ligating the DNA ends, leading to intact DNA (**Figure 1.15)** (Sinha et al., 2016).

The pathway of MMEJ repair results in frequent deletions and occasionally insertions. Because of this, MMEJ is frequently associated with chromosome abnormalities such as deletions, translocations, inversion and other complex rearrangements (Decottignies, 2013).

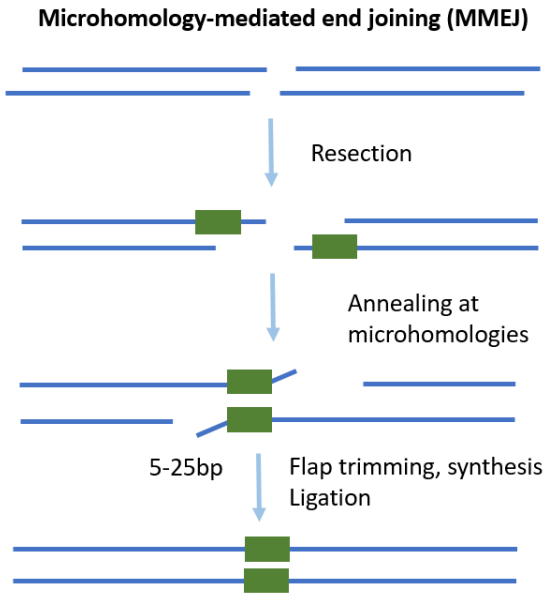


Figure 1.15. Microhomology-mediated end joining (MMEJ), a particular form of alternative non-homologous end joining (alt-NHEJ) that requires 5-25 nt of homology internal to the ends to align them for repair; and single-stranded annealing (SSA) involves annealing between more extensive homologies provided by direct repeats flanking the double-strand break (DSB). Image adapted from (Sinha et al., 2016).

1.6.3.4 Repair of cross-linked strands

DNA-crosslinks are one of the most severe types of DNA lesions. Crosslinks (CLs) can be subdivided into DNA-intrastrand CLs, DNA-interstrand CLs (ICLs) and DNA-protein crosslinks (DPCs) (Enderle et al., 2019). Bacteria use NER to repair intra-strand crosslinks, while a combination of NER and HR is used to repair inter-strand crosslinks (**Figure 1.16**) (Burby & Simmons, 2019; McHugh et al., 2001; Noll et al., 2006). The UvrA protein recognizes crosslinks in genomic DNA to initiate repair, which is described in **Section 1.6.2.2** as NER. In the case of ICLs, unlinking the two strands can occur through one of two known pathways, followed by repair of the resulting monoadduct. The primary mechanism of ICL separation involves incisions to one strand by endonucleases associated with NER. The resulting gap and strand break are further processed by translesion DNA synthesis and HR prior to repair of the monoadduct through a second round of NER. Another ICL repair pathway was recently discovered in

both eukaryotes and prokaryotes, where DNA glycosylase cleaves one of the N-glycosidic bonds that link the modified nucleotide to the DNA backbone, generating an AP site on one strand but leaving the backbone intact (Bradley et al., 2020).

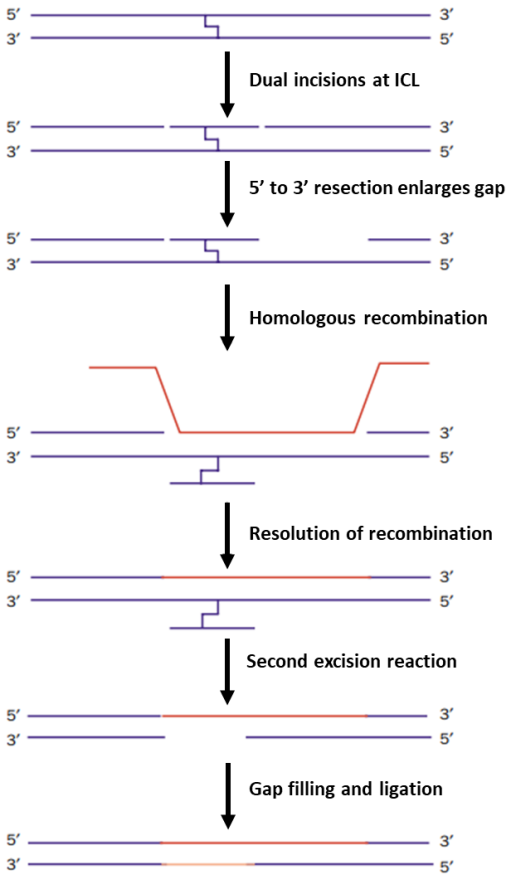


Figure 1.16. Schematic of ICL repair in Bacteria, first proposed by Cole in 1973 (Cole, 1973). Following incisions around the ICL, a 5'-3' exonuclease enlarges the gap. Homologous recombination into the gap, then allows a second excision resynthesis event on the complementary strand. The gap is filled and joined by a DNA ligase. Figure was adapted from (McHugh et al., 2001).

1.6.4 Translesion synthesis (TLS)

Translesion synthesis (TLS) is a DNA damage tolerance pathway that allows the DNA replication machinery to replicate past DNA lesions, such as thymine dimers or abasic sites. Replicative DNA polymerases are generally not capable of replicating damaged DNA (Lindahl, 1993). Cells across all domains of life, are equipped with specialised DNA polymerases that are able to replicate damaged DNA, in a pathway known as translesion synthesis (TLS) (Bridges, 2005). Most of the TLS enzymes belong to the Y-family of DNA polymerases

that display a low fidelity of DNA synthesis, a low processivity, and are devoid of 3'-5' proofreading activity (Fujii & Fuchs, 2004). TLS can also be carried out by repair polymerases that do not belong to the Y-family. In *Escherichia. coli*, the B-family Pol II is responsible for -2 deletion bypass products (Becherel & Fuchs, 2001) and in Gram-positive Bacteria *Bacillus. subtilis* and *Streptococcus pyogenes*, the type C replicative polymerase DnaE, is also an error-prone TLS enzyme (Le Chatelier et al., 2004).

When replicative DNA polymerases (e.g., Pol III) encounters a replication-blocking lesion, it either stops one nucleotide before the lesion or it may sit idle at the lesion site. As soon as Poll III stalls at a lesion site, it dissociates from the template and the replicative DNA helicase (DnaB) continues to open the parental duplex, generating a stretch of single-stranded DNA downstream from the lesion. The single-stranded DNA might first be covered by single stranded binding proteins (SSB), this SSB-DNA filament is then converted into a RecA-nucleoprotein filament (RecA), through the action of recombination mediator proteins (Fuchs & Fujii, 2013). RecA activates Pol V, which then accesses the 3'-OH end of the nascent strand freed by the dissociating Pol III (Shinagawa et al., 1988). After Pol V has synthesised past the DNA lesion, Pol III regains access to the nascent strand and resumes elongation (Fuchs & Fujii, 2013).

1.7 The prokaryotic SOS response

Organisms have a set of diverse genetic programs that are used to alter cellular mechanisms in response to environmental cues (Simmons et al., 2008). When prokaryotes are exposed to DNA damaging agents, this event can result in the induction of the SOS response, which is a global regulatory network targeted at addressing DNA damage (Erill et al., 2007).

SOS systems contain a set of different damage-inducible (*din*) genes, governed by the products of the *lexA* and *recA* genes which act, respectively, as inducer and repressor of the system (**Figure 1.17**) (Dallo & Weitao, 2010; Erill et al., 2007). RecA is recruited on ssDNA by presynaptic complexes RecBCD or RecFOR. RecCD recognises DSBs or double-strand ends (DSE). Its helicase and

nuclease activities result in the formation of a ssDNA substrate for RecA. RecFOR recognises DNA nicks and gaps, and recruits RecA to this ssDNA patch. RecA binds ssDNA in the form of a nucleofilament that catalyses the auto-proteolysis of the repressor LexA (Baharoglu & Mazel, 2014). Under normal conditions, transcription of these genes is blocked by the SOS repressor protein LexA, which binds to specific sequences (SOS boxes) located within their promoter region (Simmons et al., 2008). The proteolysis of LexA leads to the derepression of this regulon (Baharoglu & Mazel, 2014).

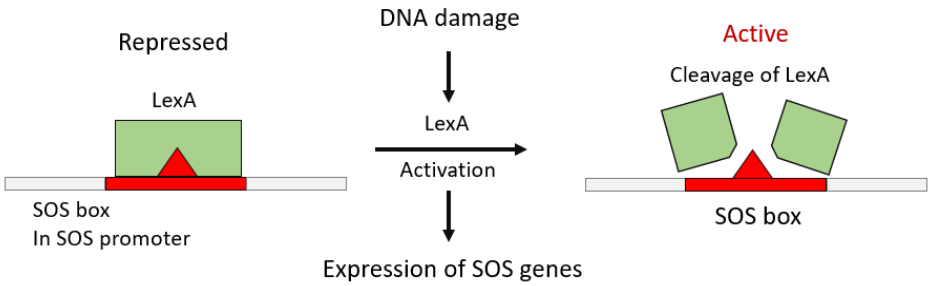


Figure 1.17. LexA binds to the SOS box, blocking transcription of SOS genes. When activated by DNA damage, the co-protease RecA causes LexA to self-cleave and vacate the SOS box, allowing expression of SOS genes. Image sourced from: (Dallo & Weitao, 2010).

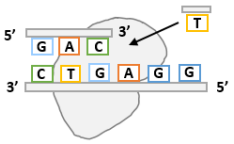
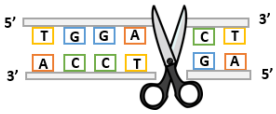
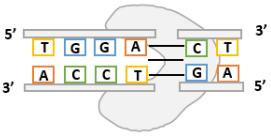
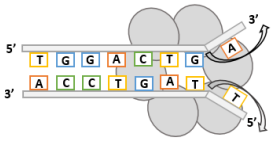
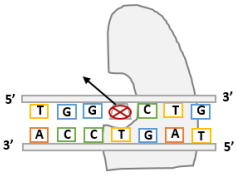
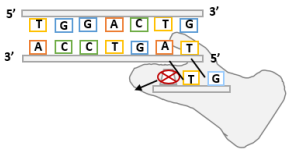
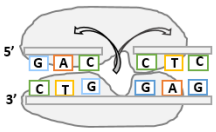
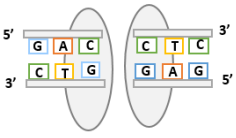
The number and type of genes found in the regulon vary among Bacteria. For example there are 40 genes belonging to this regulon in *E. coli*, compared to 33 genes from *B. subtilis* (Au et al., 2005). The affinity of LexA for these genes is variable and depends on the LexA box sequence itself, and on the number of LexA boxes present at the promoters of these genes (Zhang et al., 2010). LexA is a late induced gene and can stop the SOS induction when the genotoxic signal disappears and LexA cleavage is no longer favoured (Kovačič et al., 2013).

1.8 DNA repair proteins in prokaryotes

The enactment of DNA repair within prokaryotes involves a wide range of structurally and functionally diverse proteins. Many of these repair proteins function across more than one repair pathway and are often involved in functional repair complexes with other repair proteins. These proteins are mostly well conserved between prokaryotes and eukaryotes, with many prokaryotes been discovered to have homologs of eukaryote repair proteins. There have been recent discoveries of novel DNA repair proteins, found only in particular prokaryotes,

which have been attributed to the extreme environments these organisms can occupy. Due to the complexly and number of repair proteins that have currently been discovered, the general enzymatic functions of key enzymes involved in DNA repair systems in prokaryotes, and the main role these proteins play are reviewed below (**Table 1.1**).

Table 1.1. The key enzymes involved in DNA repair in both prokaryotes and eukaryotes.

Key enzymes in DNA repair:		Main role in DNA repair:
Polymerase		The synthesis of new base pairs on DNA strands, that have been damaged.
Nuclease		Remove damaged sections or single bases from DNA strands.
Ligase		Join back together single or double DNA strands that were broken apart from damage.
Helicase		Unwind sections of DNA to expose areas of damage.
Glycosylase		Flipping mechanism to remove damaged bases from DNA strands.
Recombinase (RecA)		Plays multiple roles in DNA repair, including SOS repsonse and homologous recombination.
Topoisomerase		Can unwind DNA strands and introduce breaks to open up DNA strands for repair.
Scaffolding proteins (e.g. Ku end binding and SSB)		Ku proteins bind to double strand DNA breaks and are involved in the NHEJ pathway.

Many of these enzymes involved in DNA repair systems are also involved in DNA replication pathways, while others are only involved in DNA repair. For example, the DNA polymerases which are categorised into families based on similarities between domain structures. Pol I, II and III belong to the A, B and C families, respectively (Fuchs & Fujii, 2013). Pol I is involved in both DNA repair and replication, where it is involved in Okazaki fragment maturation and DNA repair synthesis during nucleotide excision repair (NER) (Fuchs & Fujii, 2013). Whereas the Y-family of specialized DNA polymerases are only involved in the translesion synthesis pathway (Fuchs & Fujii, 2013).

1.8.1 DNA ligases in bacterial repair

DNA ligases play an essential role in the sealing of breaks in the phosphodiester backbone of double-stranded DNA. This process is important for the replication and survival of all organisms. The repair of DNA breaks by DNA ligases occurs in a three-step reaction (Shuman, 2009). The first step involves adenylation of the DNA ligase through attack on the α-phosphorus of ATP or NAD^+ by a conserved lysine residue, forming a covalent ligase-adenylate intermediate, in which AMP is connected by a phosphoramide (P-N) bond to a lysine side chain (Lehnman, 1974). This is followed by DNA binding and transfer of the adenyl group to the 5' phosphorylated end of the donor strand, to form a DNA-adenylate intermediate. The final step results in the formation of a new phosphodiester bond. This proceeds through the reaction of the adenylated donor end, with the adjacent 3' hydroxyl acceptor joining the two polynucleotides together and releasing AMP (**Figure 1.18**) (Shuman, 2009).

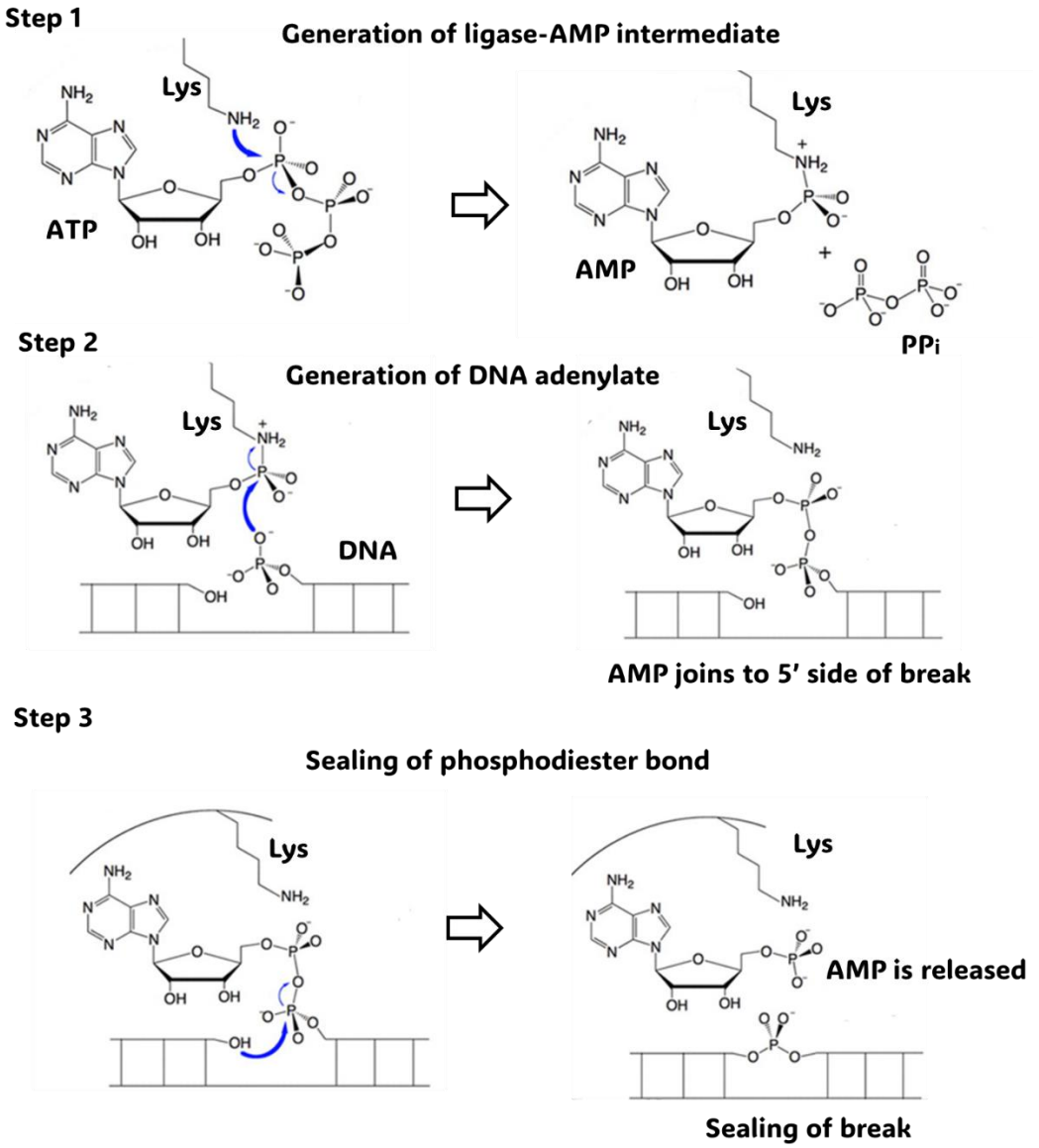


Figure 1.18. Three-step pathway of nick sealing by DNA ligase. Figure is adapted from (Shuman, 2009).

DNA ligases belong to the superfamily of nucleotidyltransferases and share similarity to the GTP-dependent mRNA capping enzymes (Shuman & Lima, 2004; Williamson et al., 2014). DNA ligases can be divided into two groups, according to their nucleotide substrate requirement. The first being the highly conserved NAD-dependent enzymes, which can be found in all bacteria and are involved in joining of Okazaki fragments during DNA replication (Ogura & Wilkinson, 2001; Williamson et al., 2014). The second group includes the structurally diverse ATP-dependent enzymes, which can be identified through all domains of life (Martin & MacNeill, 2002; Williamson et al., 2014).

NAD-dependent DNA ligases possess a unique N-terminal Ia domain which is important for utilization of this substrate (Sriskanda & Shuman, 2002; Williamson et al., 2014). ATP-dependent DNA ligases all share a common catalytic core, including all six of the conserved nucleotidyltransferase motifs and contains the adenylation domain, where AMP cofactor is covalently bound and an oligonucleotide-binding domain, which will bind the minor groove of the DNA duplex upon nick binding, and assists in the adenylation reaction (Doherty & Suh, 2000; Shuman, 2009).

All known bacteria encode a highly conserved NAD-dependent DNA ligase (LigA) (Gottesman et al., 1973; Kaczmarek et al., 2001). Some bacterial species, including *E. coli, Salmonella typhimurium, Shigella flexneri, Yersinia pestis,* and *Pseudomonas putida*, also have a second NAD-dependent ligase (Sriskanda & Shuman, 2001). The assumption that bacteria encode only NAD-dependent DNA ligases was overturned when an ATP-dependent ligase was discovered in the respiratory pathogen *Haemophilus influenzae* (Cheng & Shuman, 1997). ATP-dependent ligase homologues co-exist with NAD-dependent enzymes in multiple bacterial species, including major human pathogen like; *Neisseria meningitidis, Y. pestis, Vibrio cholerae, Pseudomonas aeruginosa,* and *Mycobacterium tuberculosis*. Of interest is *M. tuberculosis*, which encodes three distinct ATP-dependent ligase homologues (LigB, LigC and LigD, plus an NAD-dependent ligase (LigA) (**Figure 1.19)** (Gong et al., 2004).

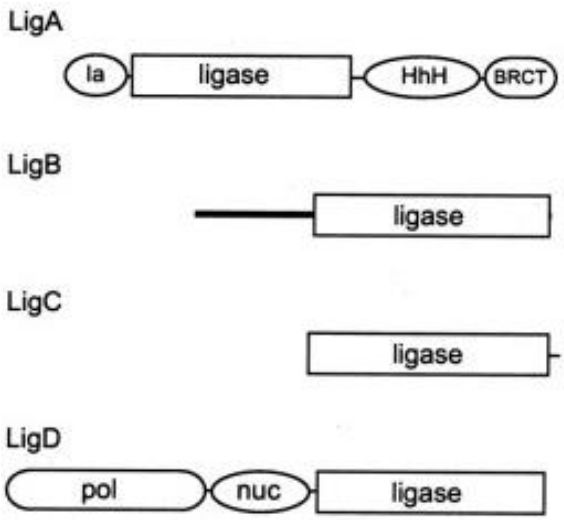


Figure 1.19. Multiple DNA ligases of *M. tuberculosis*. Here LigA, LigB, LigC and LigD polypeptides are shown in cartoon form with the N termini on the left and the C termini on the right. The core ligase catalytic domains are represented by rectangles. Image was adapted from (Gong et al., 2004).

1.8.2 Nucleases in bacterial repair

Nucleases are a diverse family of protein enzymes, found across all domains of life. Many of these nucleases are central to all DNA repair processes, particularly in their ability to remove structural anomalies, induced by exogenous or endogenous agents (Marti & Fleck, 2004). In DNA repair pathways, nucleases, by themselves or in enzyme complexes, play a role in the nucleolytic processing of DNA (Friedberg et al., 2005). This process is achieved through the cleavage of a phosphodiester bond between a deoxyribose and a phosphate residue, resulting in a 5'-terminal phosphate and a 3'-terminal hydroxyl group (Fernandez et al., 2011). In contrast to this process, AP lyase activities process DNA either by a β-elimination reaction producing a 3'-terminal phosphoglycerldehyde residue or by β-δ-elimination, which results in a 3'-phosphate end (Doetsch & Cunningham, 1990). Nucleases are classified as sugar-specific nucleases (DNA and RNA nuclease) or sugar non-specific nucleases (Rangarajan & Shankar, 2001). Nucleases can be further grouped, by their DNA targeting specificity (**Figure 1.20**), either requiring a free 5' or 3' end (exonucleases), or hydrolysing internal phosphate bonds, with no requirement for a free end (endonucleases) (Marti & Fleck, 2004). There is also a group of nucleases that cleave DNA flap structures, at or near the junction between single-stranded and double stranded region, so-called Flap endonucleases (FENs) (Marti & Fleck, 2004).

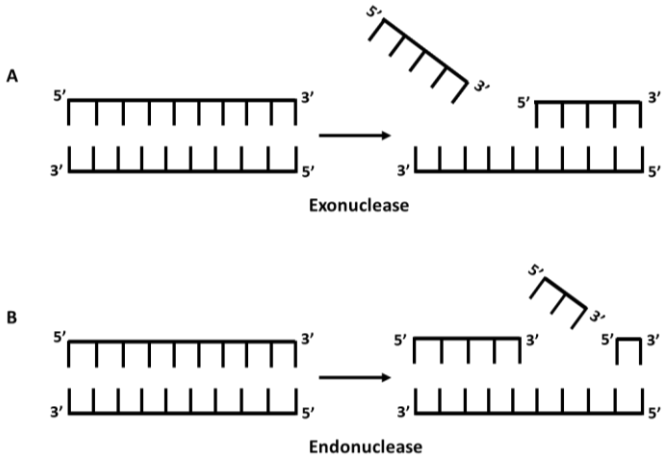


Figure 1.20. Schematic of the endonuclease and exonuclease activity exhibited by DNA binding nucleases.

DNA repair nucleases have been intensively characterised in eukaryotes and selected model Bacteria, such as *Escherichia coli* and *B. subtilis*. From these model organisms, various nucleases possessing different biological roles, catalytic activities, biochemical properties, and regulatory mechanisms have been described. Thanks to advancements in genomic technologies, homologs of some of these nucleases have been identified *in silico* across all domains of life. For example, within the genome of *P. putida* is a four gene cluster of predicted DNA repair proteins. Within this cluster are two predicted DNA repair nucleases, binuclear metallo-phosphoesterase (MPE), and a metallo-β-lactamase exonuclease. MPE protein was found to share homology to Mre11 nuclease protein, while metallo-β-lactamase exonuclease is predicted to be a homolog of SNM1/Apollo family of nucleases (Ejaz & Shuman, 2018).

The following table shows known nucleases involved in prokaryotic DNA repair pathways. As many proteins and unique pathways are still being discovered, this list of proteins and their pathways is not a complete picture (**Table 1.2**).

Table 1.2. Examples of DNA modifying nucleases and their associated DNA repair pathways.

Repair pathway	Prokaryote/Bacteriophage	Key references
Base excision repair **Abasic site processing**	DNA glycosylases AP endonuclease Endo VIII Endo V Endo /IV Exo III	(Kurthkoti et al., 2020)
Mismatch repair	MutH RecJ ExoI ExoX ExoVIII NucS/EndoMS (Archaea and ActinoBacteria)	(Jun et al., 2006) (Castañeda-García et al., 2017)
Nucleotide excision repair **5' processing** **3' processing** **Short patch repair**	 UvrC, UvrB UvrC Vsr	(Kisker et al., 2013)

Double strand break repair End processing	RecB, RecCD SbcD, SbcC RecJ Exo VIII, RecE ExoI, SbcB	(Nishino & Morikawa, 2002)
Holiday junction resolvase	RuvC RusA T4 endo VII T7 endo I	(Sharples et al., 1999)

1.9 DNA repair in extremophiles

Extremophiles are organisms that flourish in environmental conditions that most life forms find inhospitable, such as high/low temperature, pH, salinity, and pressure (Zhu et al., 2020). Extremophilic organisms can be categorised as acidophilic (optimal growth at low pH), alkaliphilic (optimal growth at high pH), halophilic (thriving well in high salt concentrations), thermophilic (optimal growth at high temperatures), psychrophilic (thriving well at low temperatures), desiccation or radiation resistant organisms and so on (**Figure 1.21**) (Dalmaso et al., 2015; Singh et al., 2019; Zhu et al., 2020). Extremophilic organisms inhabit environments that can cause severe DNA damage as described in **Section 1.5**. The following sections will describe three main extremophiles; radioresistant, psychrophiles, and thermophiles and the unique DNA repair enzymes and pathways they have evolved as a consequence of their environment.

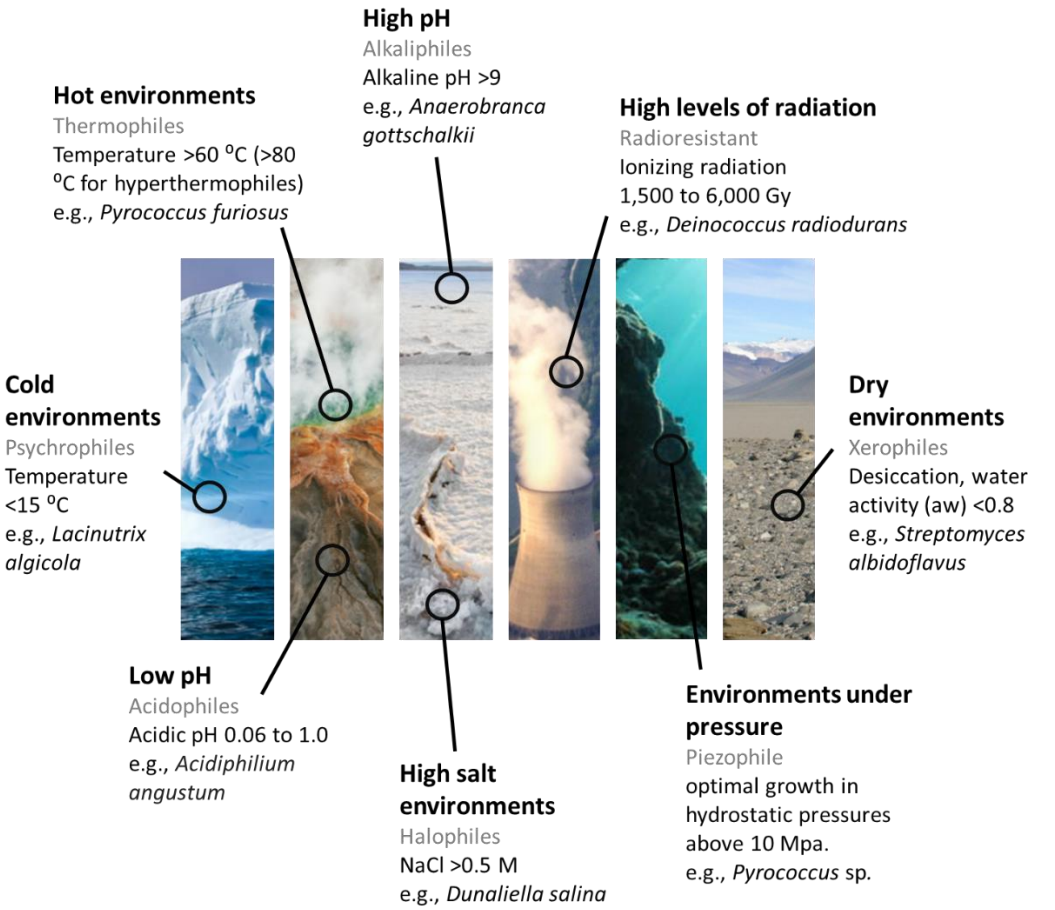


Figure 1.21. Schematic representation showing examples of extreme environments found on Earth and the type of extremophiles that inhabit each of these environments (Gupta et al., 2014; Schröder et al., 2020; Seckbach & Oren, 2005; Singh & Gabani, 2011).

1.9.1 Radiation and desiccation resistance

There is little likelihood that organisms have evolved specific pathways to protect themselves from the effects of high dose radiation, as there are no known naturally occurring environments where exposures exceed 400 mGy per year. (UNSCEAR, 2000). Instead, it appears that the damage caused by γ-irradiation shares similarities with damage caused by other stresses that bacteria have adapted to. For instance, desiccation leads to many DNA double-strand breaks (DSBs) in the genomes of *Deinococcus radiodurans* (Mattimore & Battista, 1996) and members of the Cyanobacterial genus *Chroococcidiopsis* (Billi et al., 2000). Both organisms can tolerate desiccation and are resistant to the potentially fatal effects of ionising radiation, which suggests that their radioresistance may be a result of their ability to tolerate strand breaks induced by desiccation (Cox & Battista, 2005).

Currently several distinct species of radiation-resistant bacterial species have been identified, across a broad range of phyla (**Figure 1.22**). The presence of numerous radioresistant species that are not closely related to each other, indicates that the molecular mechanisms responsible for protecting against the harmful effects of ionising radiation, have evolved separately in these organisms (Daly, 2012).

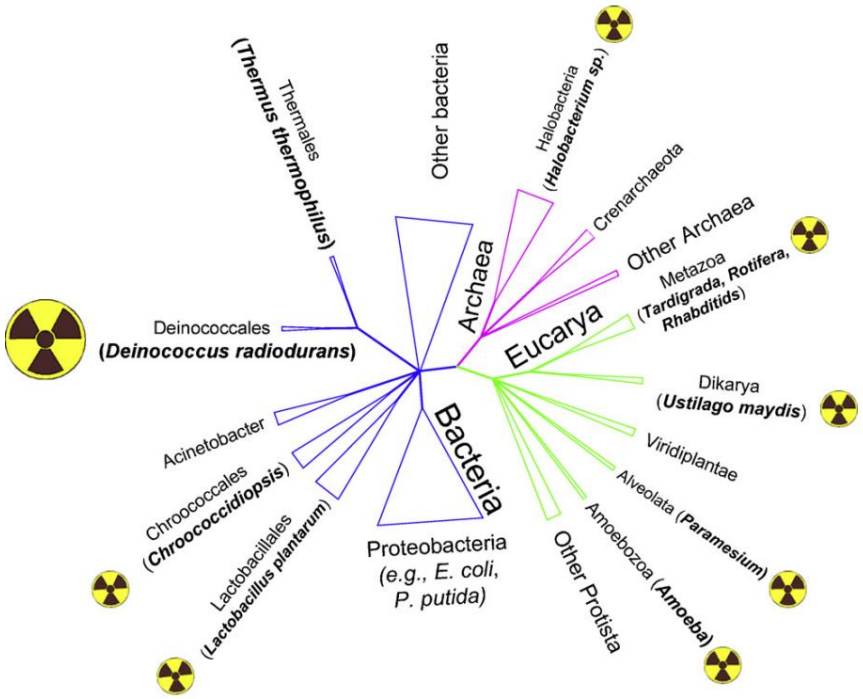


Figure 1.22. Phylogenetic distribution of radiation resistant organisms. Figure sourced from (Daly, 2012).

Radiation resistance appears to be governed by both passive and active (enzymatic) mechanisms. Passive mechanisms involve having multiple genome copies, a nucleoid organization that is highly condensed, and an accumulation of Mn (II) ions, which could hinder the production of reactive oxygen species. Enzymatic mechanisms for radiation resistance consist of both conventional DNA repair processes and novel functions (Cox & Battista, 2005).

Deinococcus radiodurans is one of the best studied radiation-resistant bacterial species and serves as a model organism for understanding radiation resistance (Liu et al., 2023). This bacterium has an extraordinary resistance to ionising radiation (IR), and can tolerate a radiation dose of 5 kGy during its stable

growth period, without evidence of mutation (Moseley & Mattingly, 1971). The strong radioresistance of *D. radiodurans* is attributed to several factors, including an efficient DNA repair capacity. IR can induce four different types of DNA damages: DNA double-strand breaks (Ujaoney et al., 2021), bulky lesion crosslinks, base mismatch, single-strand breaks and single base damages (Liu et al., 2023). To overcome these different DNA damages, *D. radiodurans* uses four DNA repair strategies: recombination repair, which includes the commonly-used HR mechanisms as well as more specialized pathways, nucleotide excision repair (NER), base excision repair (BER) and mismatch repair (MMR) (**Section 1.6**) (Liu et al., 2023; Misra et al., 2013). Several of these pathways include Deinococcus-specific proteins, directly involved in the repair of these particular DNA damages which are found in the *D. radiodurans* genome in addition to more conserved repair enzymes (Figure 1.23) (Liu et al., 2023).

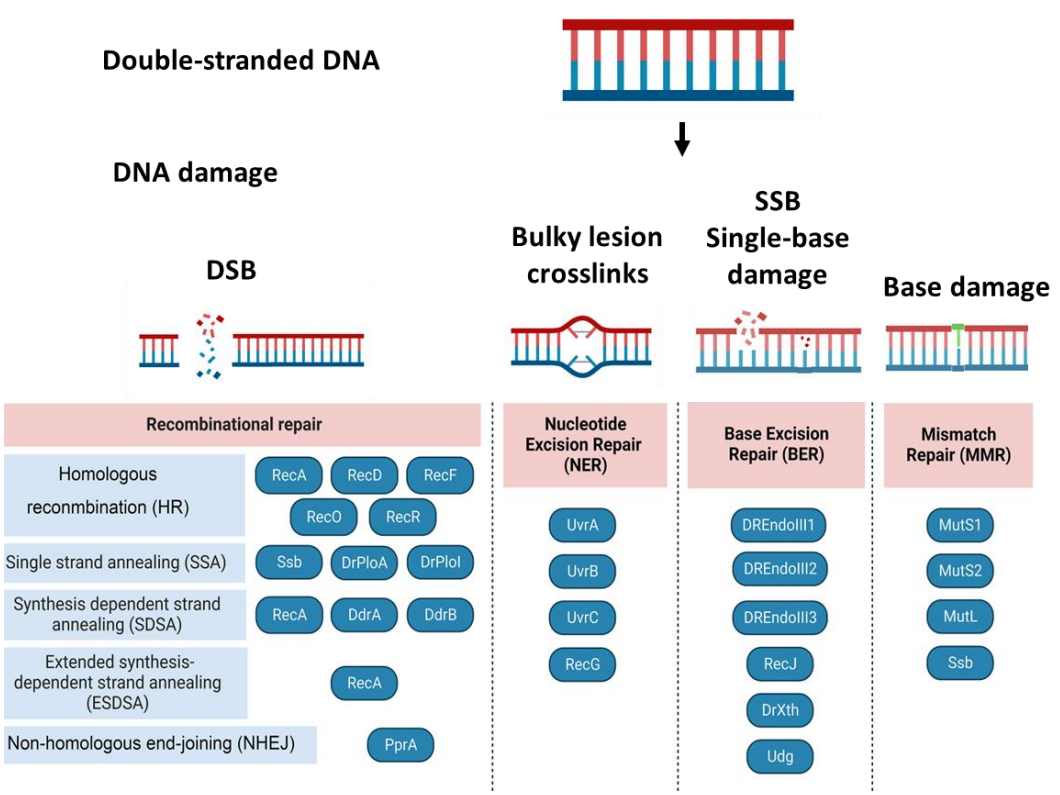


Figure 1.23. A figure showing the four types of DNA damage, due to IR and the corresponding repair pathways and proteins, from *D. radiodurans*. Figure adapted from (Liu et al., 2023).

Several Actinobacteria species exhibit extraordinary radiation resistance including *Kineococcus radiotolerans*, *Rubrobacter radiotolerans*, *Rubrobacter xylanophilus*, and *Kocuria rosea* (Bagwell et al., 2008). *K. radiotolerans* in

particular exhibits radiation resistance approaching that of *D. radiodurans* (Phillips et al., 2002). *D. radiodurans* and *K. radiotolerans* belong to different phyla, and it remains unclear whether both organisms attain radiation resistance through a common set of gene products (Bagwell et al., 2008).

Many of the genes involved in radiation-induced DNA repair by *D. radiodurans* appear to be absent in other radiation resistant bacteria. For example, *K. radiotolerans* lacks homologs of pprA, ddrA, ddrB, ddrC, ddrD genes suggesting it may have a unique genetic toolbox for radiation protection (Bagwell et al., 2008).

1.9.2 Extreme temperature tolerance

Extremely cold environments present many challenges for all organisms. These environments are mostly occupied by psychrophilic and psychrotolerant bacteria, which have various strategies to deal with this environment (Flores et al., 2009).

In the literature, there are a limited number of studies that have focused on DNA repair proteins in psychrophiles. However, advances in genomic techniques, such as metagenomics are beginning to uncover novel nucleic acid repair proteins from psychrophiles.

In proteome analysis of a *Pseudomonas* psychrophile, from the Arctic, proteins involved in DNA repair, including exodeoxyribonuclease III, were found to be upregulated at 4 ºC (Abraham et al., 2020). Proteomic evidence (Nunn et al., 2015) has revealed the marine psychrophile, *Colwellia psychrerythraea*, shows a higher abundance of proteins involved in DNA repair, after exposure to sub-zero temperatures (-10 ºC) for 8 weeks. They found that some of those proteins were associated with the stress response to DNA damage (DNA helicase II, exodeoxyribonuclease III, MutS and Nicotinamide adenine dinucleotide (NAD) dependent DNA ligase (Nunn et al., 2015).

Pseudomonas frederiksbergenis, a psychotropic bacterium isolated from a glacial stream in Sikkim Himalaya, can resist freezing temperatures, freeze thaw

cycles and radiation (Kumar et al., 2019). Genes present in plasmids of psychrophilic bacteria have been shown to play a role assisting them with cold tolerance. Such genes are involved in carbon storage, DNA repair and replication, transport of carbohydrates and amino acids. Here several DNA helicases were identified, that are associated with cold adaptation: RecQ, RecG, UvrD, RuvB, RuvA, Rep and DinG (Kumar et al., 2019).

Another study (Dziewit & Bartosik, 2014) found that plasmids belonging to cold-active Bacteria, contain numerous genes encoding enzymes involved in protection against cold and ultraviolet radiation, scavenging of reactive oxygen species, resistance to heavy metals and an abundance of novel DNA repair proteins. Bacteria inhabiting polar regions have to cope not only with cold, but also with increased solar UV radiation due to stratospheric ozone depletion (Madronich, 1994). Dziewit & Bartosik, 2014 identified two highly homologous genetic modules that confer tolerance to UV radiation- the rulAB and umuDC DNA repair operons, from plasmids of cold-active Bacteria. DNA repair systems are essential in mending the DNA damage which occurred during cold stress and are said to be involved in the survival mechanisms of psychrophiles at low temperatures (Abraham et al., 2020).

Thermophiles are organisms that thrive at relatively high temperatures, between 41 and 122 ºC. Many of these thermophilic organisms are Archaea, though are small group are made up of Fungi and Bacteria (Takai et al., 2008). These organisms have adapted to survive in extreme conditions such as hot springs, deep-sea hydrothermal vents, and geothermal sites (Bargagli et al., 2004; Ionescu et al., 2007; Jin et al., 2019). To cope with these conditions, thermophilic bacteria have evolved unique mechanisms to repair their DNA under-high temperature conditions (Grogan, 2000; Makarova et al., 2002; Sandigursky & Franklin, 1999). These repair systems involve a variety of enzymes and proteins, including DNA glycosylase, endonucleases, helicases, ligases, and DNA polymerases (Luo & Barany, 1996; Sandigursky & Franklin, 1999; Trivedi et al., 2005). DNA repair proteins are essential for the survival of thermophilic bacteria in these extreme environments.

The extremely thermophilic bacterial species *Thermus thermophilus*, has an optimal growth temperature above 65 ºC. Genome investigations of this species has revealed a high number of DNA repair functions, directly involved in strategies for survival at high temperatures, such a reverse gyrase (Brüggemann & Chen, 2006). This protein consists of a helicase and a type I topoisomerase and introduces positive supercoiling into circular DNA, thus preventing excess local unwinding of the double helix at high temperatures (Rodríguez & Stock, 2002). This protein is considered as a molecular marker for hyperthermophily, and is a unique feature of hyperthermophilic Bacteria and Archaea, which is absent from all other mesophilic genomes (Forterre, 2002). Also, within the genome of *T. thermophilus* is a DNA ligase. This ligase differs from mesophilic ATP-dependent ligases through its NAD-dependency, its temperature optimum, at 65 ºC instead of 37 ºC and its higher fidelity than T4 DNA ligase (Luo & Barany, 1996).

One of the main mechanisms of DNA repair in thermophilic bacteria is the HR pathway. Thermophiles have a set of enzymes involved in the HR pathway, including RecA and RadA. HR is a required mechanism for repairing DSBs that are caused by high-temperature conditions, such as heat shock and oxidative stress (Haldenby et al., 2009; McIlwraith et al., 2001; Reich et al., 2001).

Another process of DNA repair in thermophilic bacteria is the NER pathway. The NER mechanism in thermophilic Bacteria involves the UvrABC endonuclease complex, which identifies and cleaves the damaged DNA strand. The NER pathway in thermophilic Bacteria is similar to that found in mesophilic bacteria, but the enzymes involved are more heat-stable (Trivedi et al., 2005).

In addition to HR and NER, thermophilic Bacteria also possess a unique DNA repair enzyme involved in the base excision repair pathway. *Thermus aquaticus* endonuclease V (TaqEndoV), belonging to the AP endonuclease family, has been found to have a specialised mechanism of DNA binding and cleavage, that allows it to function optimally at high temperatures. This enzyme recognises and cleaves DNA at sites of damage caused by oxidative stress, such as 8-oxoguanine and thymine glycol lesions (Warner, 1983).

1.10 Biodiscovery of enzymes for new applications

Extremophiles, over the last few years, have gained significant interest due to their capacity to catalyze reactions and their potential for industrial applications under extreme conditions. Although extremozymes have been discovered for many decades, researchers are still focusing on two main approaches: genetically engineering existing enzymes to increase their potency, and screening for new enzymes from diverse sources that exhibit the necessary characteristics for industrial and biotechnological applications. This is because many commonly used commercial enzymes are unable to meet industrial requirements, including the ability to withstand varying pH levels, temperatures, and aeration conditions with high reproducibility. As a result, extremozymes are being increasingly recognized as a viable strategy for industrial processes and biorefining (Sarmiento et al., 2015).

Many enzymes derived from psychrophiles exhibit high levels of low-temperature activity and are heat labile. These characteristics provide three key advantages for the use of cold-active enzymes in biotechnology. Firstly, due to their high activity, a lower concentration of the enzyme is needed to achieve a desired level of activity. Consequently, this reduces the amount of expensive enzyme preparation required in each process. Secondly, their ability to function effectively at lower temperatures allows them to remain efficient without the need for heating during a process. For instance, a cold-active lactase derived from an Antarctic bacterium has been patented for its ability to hydrolyze lactose during milk storage at low temperatures (Hoyoux et al., 2001). Thirdly, their heat labile nature enables efficient inactivation through moderate heat input after a process is complete (Feller, 2013). For example, a heat-labile alkaline phosphatase sourced from an Antarctic bacterium is utilized in molecular biology for dephosphorylating DNA vectors (Wang et al., 2007). Additionally, two other psychrophilic enzymes are commercially available for molecular biology applications, capitalizing on their heat-labile property. Shrimp nuclease is used to remove carry-over contaminants in PCR mixtures, while heat-labile uracil-DNA N-glycosylase from Atlantic cod is used to eliminate DNA contaminants in sequential PCR reactions (Feller, 2013; Leiros et al., 2003). Following their application, these enzymes are deactivated through heat treatment.

Along with these other applications, studies have been looking into the use of extremophile enzymes as molecular biology tools with non-canonical nucleic acids. A study by (Betz et al., 2012) identified bacterial DNA polymerases that can replicate DNA-containing unnatural base pairs (UBPs). They show that Klen Taq polymerase replicates unnatural base pair by inducing a Watson-Crick geometry. Steven Benner and his colleagues (Benner et al., 2016) have successfully synthesized unnatural base pairs (UBPs) that exploit distinct patterns of hydrogen bonding. These UBPs are collectively known as "Hachimoji" DNA, representing an artificially expanded genetic system (AEGIS). In this system, the traditional nucleotides A, G, T, and C are expanded by the inclusion of two unnatural pyrimidine analogs: pseudocytidine (or isocytidine) denoted as S, and 6-amino-3-(2'-deoxyribofuranosyl)-5-nitro-1H-pyridin-2-one, referred to as Z. Additionally, the purine analogs isoguanine (B) and 2-amino-8-(1-β-d-2'-deoxyribofuranosyl) imidazo [1,2-a]-1,3,5-triazin-[8H]-4-one (P) serve as their size and hydrogen bond complementary partners (**Figure 1.24**).

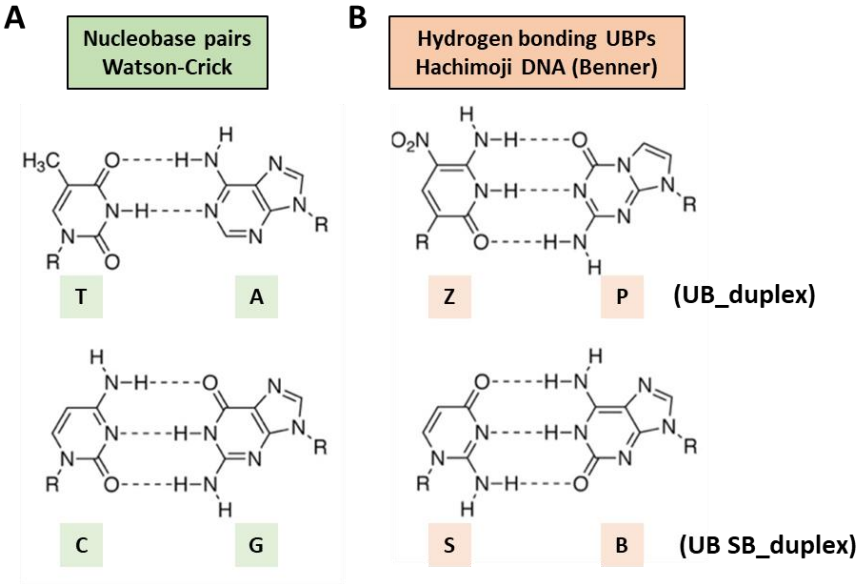


Figure 1.24. Chemical structures of nucleobase pairs. R indicates the point of covalent attachment to either deoxyribose or ribose in DNA or RNA, respectively. **A**) Watson-Crick nucleobase pairs. **B**) hydrogen-bonding UBPs used in Hachimoji DNA developed by Benner and co-workers (Benner et al., 2016). Figure adapted from (Ouaray et al., 2020).

Owing to the current findings on specialized DNA repair enzymes in extremophiles (**Section 1.9**), it is likely that other DNA repair and DNA

replication enzymes will also be tested with these non-canonical nucleic acids, to see if they retain their normal repair function when presented with these UBPs.

1.11 Protein identification from bacteria of the McMurdo Dry Valleys

The McMurdo Dry Valley (DVs) systems in Antarctica are among some of the most hostile environments on Earth, inhabited by extremophilic microbiota (Panda et al., 2022). Given that organisms inhabiting this environment are exposed to multiple environmental stressors, e.g., desiccation that can cause DNA damage, it is hypothesised that these organisms must possess extremely efficient DNA repair systems and enzymes to survive this environment (Rzoska-Smith et al., 2023). As these organisms cannot be cultured in a laboratory environment, insight into how these organisms survive comes from metagenomic studies: where all the DNA in a particular sample is extracted, sequenced, and assembled in order to obtain genetic information.

The New Zealand Terrestrial Antarctic Biocomplexity survey (nzTABS, https://ictar.aq/nztabs-science/) was part of a study to understand how the geochemistry, geology, climate, and biotic factors affected the diversity and complexity of organisms inhabiting the McMurdo DVs of Antarctica. Results from this study have been previously described by (Bottos et al., 2020; Lee et al., 2019). As part of this study soil DNA samples were collected from the Meirs, Marshall, and Garwood valleys of Antarctica, over two successive austral summers (January 2009 and 2010) (Error! Reference source not found.**, A**). These soil samples were sequenced at the Joint Genome Institute (JGI) and resulted in the assembly of 31 metagenomes, available through the JGI Genomes Online Database (GOLD). Details of DNA preparation, sequencing assembly and annotation can be found in our publication (Rzoska-Smith et al., 2023) and references therein. Research conducted on these soil samples has revealed that these soils support a diverse, mostly bacterial community that appears to be extremely reactive to change, suggesting the microbiota are both viable and adaptive (Cary et al., 2010; Tiao et al., 2012). Preliminary analysis of gene function in the Clusters of Orthologous Genes (COGs) system, from the DV

metagenomes, indicated that a higher content of DNA replication and repair genes was present in the DV metagenomes relative to other environments (up to 5.5 % in COG in replication and repair, vs, 3.8 % in *E. coli*) (Error! Reference source not found., **B**, **C**) (Rzoska-Smith et al., 2023).

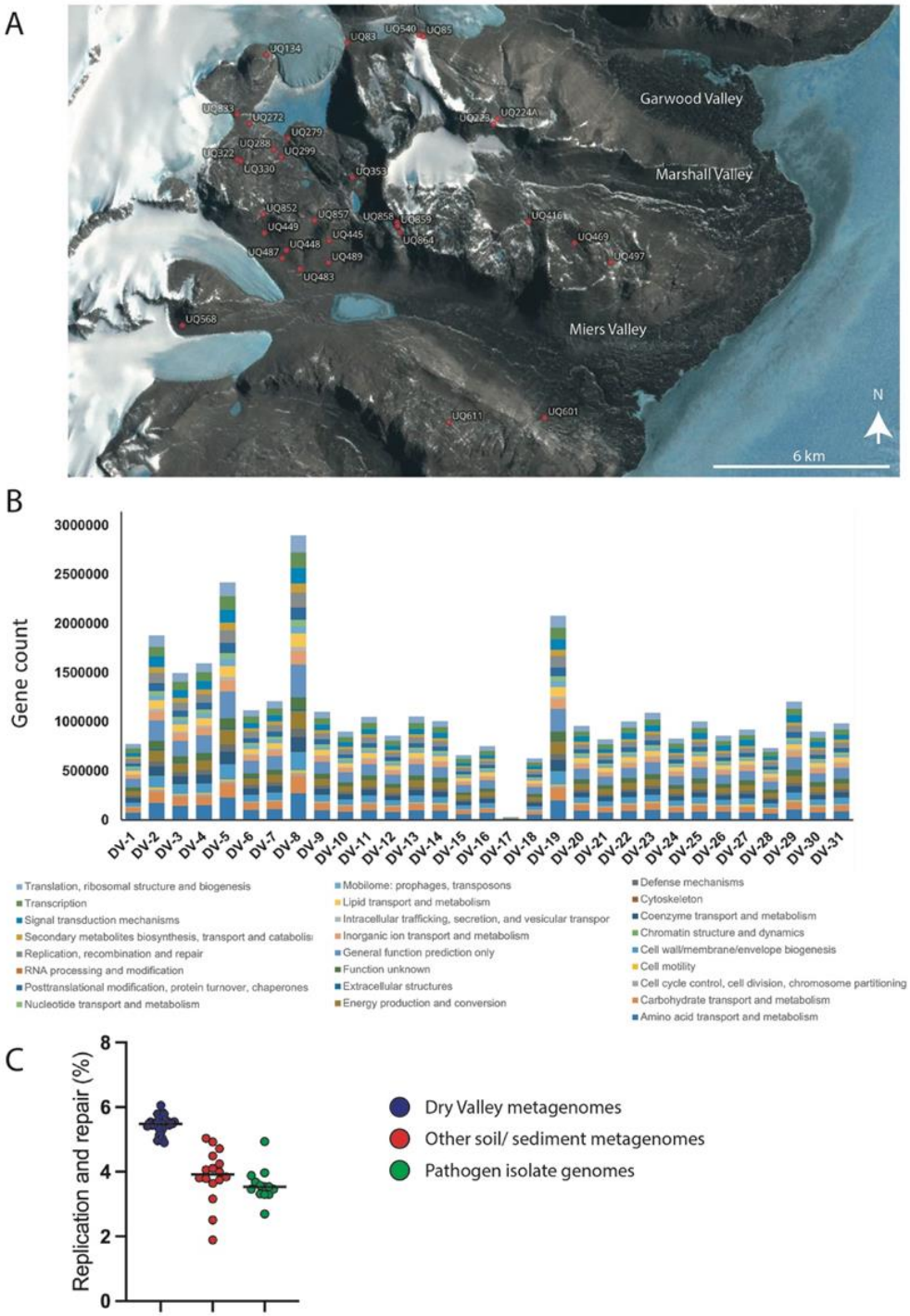


Figure 1.25. Dry Valley metagenomes sampling and sequencing. **A**) Map of sample sites for metagenome sequences. **B**) Counts of genes in COG categories for metagenome sequences from each site. **C**) Percentage of genes in the COG category 'DNA replication and repair' from Dry Valley metagenomes (blue), other soil/ sediment metagenomes (red) or genomes from isolated pathogens (green). Figure sourced from (Rzoska-Smith et al., 2023).

Using various bioinformatic techniques such as sequence similarity networks (SSN) (Atkinson et al., 2009) and programs from the HMMER suite (hmmsearch and hmmscan) (Potter et al., 2018), Dr Adele Williamson conducted an analysis of these metagenomes to identify novel groups of DNA modifying enzymes. HMM profiling was used to identify sequences from the DV metagenomes that matched certain Pfam domains (e.g., NucS, DNA_ligase_A_M, Hjc, UvdE etc). The probe domains used in this search were chosen on the basis that they comprise a large family with many on them involved in DNA repair or they belong to a recently described or specialist DNA repair pathway. Hits were retrieved for all domains used in this search, suggesting that the DV metagenomes have a diverse number of DNA repair enzymes. There was a large number of NucS homologs and ATP-dependent DNA ligases (Rzoska-Smith et al., 2023).

Following on from this, sequence similarity networks were constructed for each set of sequences identified from the HMM search (Rzoska-Smith et al., 2023). This bioinformatic tool is used to determine how related groups of genes are to each other and grouping them into clusters based on how similar they are to one another. This technique is often used to identify novels groups of DNA modifying enzymes (Zallot et al., 2018). The sequence similarity networks identified clusters that primarily contained sequences from metagenome representatives. The sequences in these clusters were further investigated to determine if these were potentially novel DNA repair genes. Sequences that were not unique, poorly aligned or lacking probable start or stops codons were discarded (Rzoska-Smith et al., 2023).

From the sequence similarity networks candidate genes were selected based on one or more of three conditions: the first being they belonged to a large contig with other DNA modifying enzymes, second was that they contained a unique domain insert and the third was that they had low sequence identity to other homologues in the database. From these conditions several candidate genes were selected for recombinant expression and preliminary characterisation. Preliminary expression trials were conducted on several types of DNA repair enzymes including helicase, nuclease, polymerase, and ligase genes. Soluble

expression was unachievable for many of the enzymes and only those that were able to be recombinantly expressed where further investigated further. The final candidates include: a NucS homolog (hereafter DV-Nuc3), and three LigB DNA ligases, with one of the ligases containing an N-terminal nuclease domain (hereafter DV-Lig2, DV-Lig5 and DV-1-1-Lig-Nuc) (Rzoska-Smith et al., 2023).

1.12 Research aims and objectives

1.12.1 Research statement

My research aims to characterise novel DNA repair pathways and proteins used by Antarctic psychrophiles and determine how these proteins recognise and correct DNA damage. These pathways and proteins will be compared to those found in model organisms, allowing us to interpret differences in terms of adaption to the environment. This research will explore the diversity and evolutionary origins of extremophilic DNA repair systems and determine how widespread these pathways are among different microbial communities at different sites in Antarctica, using bioinformatic methods on previously collected samples. To uncover novel bacterial DNA repair mechanisms, I will investigate DNA repair enzymes from Antarctic Dry Valley microbial metagenomes using a combination of *in silico* analysis, and *in vivo* characterisation of recombinantly produced proteins.

1.12.2 Hypothesis

Based on extreme DNA-damaging factors found in the natural environment of these psychrophiles, taxonomic distance from characterised model organisms and findings from preliminary analysis of metagenome libraries, it is hypothesised that some bacteria inhabiting the Antarctic Dry Valley systems possess novel DNA repair systems involving novel enzymes, mechanisms, or complex interactions.

1.12.3 Research aim

The overall aim of this thesis was to recombinantly produce these enzymes from the Dry Valley metagenomes, to characterise their structure and biological

functions, and to determine the molecular details of how they potentially recognise and correct damages.

1.12.3.1 Key questions

1. What DNA damages do these enzymes recognize? E.g., do they act on freeze-thaw induced breaks or UV- induced thymine dimers?
2. What do they do about the damage? E.g., do they cut out the damaged piece
3. How do they do it? What catalytic residues are essential?
4. What is the 3D structure of these enzymes?
5. What other proteins are required and how do they interact?
6. Can any of these be used for biotechnological applications, e.g., molecular biology tools or diagnostics?

1.12.3.2 Initial research objectives

- Recombinantly express and purify several candidate DNA repair proteins identified from the Dry Valley metagenomes, through the process of cloning, expressing, and purifying.

- Structurally characterise these candidate proteins using X-ray crystallography techniques.

- Determine the enzyme activity of the candidate proteins using gel-based activity assays.

- Recombinantly produce other DNA modifying proteins from the same gene cluster and determine interactions between protein components.

1.12.3.3 Revised research objectives

- Compare the structures of these proteins, to homologous proteins, using crystallisation techniques, or through the use of AlphaFold2 prediction models.

- Determine the range of DNA damages and cofactor usage that the Lig B-type enzymes (DV-Lig2 and DV-Lig5) are active on and compare this to previously characterised DNA ligases.

- If the ligase enzymes are active on several different DNA damages, it would be interesting to see if they are able to ligate DNA substrates that contain unnatural base pairs (UBPs). These DNA substrates with UBPs are known as ‘’Hachimoji’ DNA and have been previously tested against DNA polymerases (Section 1.10) and other DNA ligases by the Williamson group.

- Recombinantly express the entire DV-1-1-Lig-Nuc and use enzyme activity assays to determine if it exhibits nuclease and ligation abilities. Separately express each domain of DV-1-1-Lig-Nuc (DV-1-1-Nuc and DV-1-1-Lig) to determine their independent activities and design a mutant for DV-1-1-Nuc that inhibits catalytic activity or shows reduced activity.

- Determine what DNA damages the NucS type enzyme (DV-Nuc3) shows specificity towards in regards to its nuclease activity, as well as its cofactor and temperature requirements. Validate this activity by designing a mutant, to remove catalytic residues and generating an N-terminal truncation protein to see if this affects DNA binding.

1.12.4 Summary of research

This project describes the characterisation of four different proteins, from Antarctic Dry Valley metagenomes, that we hypothesise are involved in novel or specialised DNA repair pathways, that contribute to survival of their host bacteria. These include: three DNA ligases, two of which are LigB type ATP dependent DNA ligases (DV-Lig5, DV-Lig2), while the other is a putative multifunctional DNA ligase, with an N-terminal MBL Beta-Casp nuclease domain (DV-1-1-Lig-Nuc). The fourth enzyme has homology to the NucS proteins (DV-Nuc3).

1.12.5 Significance of the topic

My project focused on the DNA repair proteins from psychrophiles inhabiting the Dry Valley systems of Antarctica. The McMurdo Dry Valleys, belonging to the Antarctic continent are among one of the harshest environments on the planet. These valleys are exposed to high levels of UV radiation, multiple daily freeze thaws and very low moisture and nutrients.

Our increased understanding from the structure and activities of the DNA repair enzymes from the bacteria inhabiting the Dry Valleys has provided insight into how they survive these stresses. In particular, the presence of specialised nuclease-ligase dependent pathways suggests a dependence on stationary-phase repair mechanisms which remove damages and then rejoins DNA with minimal replication. This is consistent with previous observations that environmental bacteria inhabiting extreme environments are often slow-growing, adapted to low-nutrient conditions. Studying these DNA repair systems could provide insight into microbial survival in extra-terrestrial conditions, such as the Martian surface and could also provide leads for new biotechnological tools.

2 Chapter 2

Materials and methods

2.1 AlphaFold

Protein 3D structures were predicted using the AlphaFold prediction software accessible through Google ColabFold-v2.3.1. AlphaFold2 (from Deepmind) and AlphaFold2_advanced, use jackhmmer or mmseq2 as structural prediction tools. All protein structures, expect for DV-1-1-Lig-Nuc, were predicted using AlphaFold2. A monomeric model was used for all proteins and the number of prediction cycles was set to 20. All other parameters used original settings in the workbook. AlphaFold2_advanced was used for the structural prediction of DV-1-1-Lig-Nuc, as the protein sequence was too big to put through AlphaFold2. Predicted models were downloaded as PDB files and visualized using PyMOL v 4.6.0 (Schrödinger, 2020). For each predicted model, a predicted aligned error (PAE) plot and a predicted local distance difference test (pLDDT) plot was generated.

2.2 Cloning and DNA manipulations

Details of media, primers and PCR cycling conditions can be found in Appendices **A.2**, **A.4** and **A.6**.

2.2.1 Cloning of recombinant proteins

Gene sequences of DV-1-1-Lig-Nuc, DV-1-2-RecA and DV-1-3-DNA polymerase, were codon optimised for *E. coli* and ordered from Integrated DNA technologies IDT™ as gene blocks, with attB sites, flanking each gene. Ordered constructs included an N-terminal His-tag and cleavage recognition site for the tobacco mosaic virus protease (TEV-protease) for tag removal. Gateway cloning reactions were performed into pDONR221 entry vectors, using the BP clonase cloning kit from ThermoFisher, Scientific, following manufacturers protocol. Reactions were transformed into *E. coli* DH5α cells, following standard transformation procedures. Successful colonies, grown on LB agar plates, with

kanamycin as a selection marker, were used in colony PCR (**Section 2.2.3.3**), with M13 primers (**Appendix A.4**) to confirm correct insert size. Following confirmation of correct insert, colonies were used in starter cultures (50 µg/ml kanamycin, 5 ml LB) and incubated overnight at 37 ºC. Seeders cultures were used in plasmid prep reactions using the QIAprep Spin Miniprep Kit (Qiagen, Netherlands), following the manufacturer's protocol and purified plasmids were sent for sequencing to Massey Genome Services (MGS) to confirm correct gene insert. Plasmids with correct gene inserts were used in a Gateway cloning reaction, with pHMGWA and pDEST17 as entry vectors, using the LR clonase cloning kit from ThermoFisher, Scientific, following manufacturer's protocol. Reactions were transformed into *E. coli* DH5α and pLysS BL21 DE3 cells, following standard transformation procedures. Successful colonies, grown on LB agar plates, with ampicillin as a selection marker, were used in colony PCR, with T7 and or maltose binding protein (MBP) forward primers to confirm correct insert size. Plasmids, from *E. coli* DH5α colonies were extracted, following the procedure from above, and sent for sequencing at MGS to confirm the correct gene insert.

Clonal genes for DV-Lig5, DV-Lig2, DV-Nuc3, DV-Nuc3 mutant, DV-Nuc3 N-terminal truncation, DV-1-1-Nuc P1 mutant and DV-1-1-Lig-Nuc new start site, were ordered from Twist Biosciences in the pTwist-ENTR vector with codons optimised for *E. coli*. Ordered constructs included an N-terminal His-tag and cleavage recognition site for the tobacco mosaic virus protease (TEV protease) for tag removal. Genes were sub-cloned into the expression vectors pDEST17 (Invitrogen) and pHMGWA (GenBank #Eu680841) using the Gateway ™ LR reaction kit (Thermofisher) according to the manufacturer's instructions. Resulting expression constructs were transformed into chemically competent DH5α cells for propagation.

2.2.2 Gene construct design and cloning of DV-1-1-Lig and DV-1-1-Nuc

New gene constructs were designed for DV-1-1-Lig-Nuc, to produce the two domains separately. Primers were designed using Geneious Prime 2019.2.1,

with DV-1-1-Lig+Nuc gene, in pDONR221 plasmid, as a template. Primers were designed to include insert attBP sites, for gateway cloning and a Tev cleavage site into the forward primer sequences. A two-step PCR protocol was used with primer sequences given in **Appendix A.4**. In the first round the pDONR221 plasmid, containing DV-1-1-Nuc-Lig gene insert, was used as a template to amplify both DV-1-1-Nuc construct (DV-1-1Nuc Forward and DV-1-1Nuc Back) and the DV-1-1-Lig construct (DV-1-1Lig Forward and DV-1-1Lig Back). PCR products were visualised on a 1% agarose gel and bands, corresponding to the correct PCR product size (1.7 bp for DV-1-1Lig and 1.3 bp for DV-1-1-Nuc), were excised and purified using the 200 QIAquick ®, Gel extraction kit (QIAGEN) according to the manufacturer's instructions. DNA purity and concentration was estimated by measuring A260/A230 nm and A260/A280 nm ratios using a NanoDrop®ND-1000 Spectrophotometer (NanoDrop Technologies, USA).

The second PCR step added attBP sites to DV-1-1Nuc (DV-1-1 Forward PCR 2 and DV-1-1Nuc Back) and DV-1-1-Lig (DV-1-1 Forward 2 and DV-1-1Lig Back) generating products suitable for Gateway cloning. The PCR products were cloned into the Gateway donor vector, pDONR221 (Invitrogen), using the GatewayTM BP reaction kit (Thermofisher) and transformed into chemically competent DH5α cells, then further sub-cloned into expression vectors as described above.

Following insoluble protein expression of the DV-1-1-Nuc domain, three new constructs were designed, with each new construct having a new start, end (or both) site. (**Appendix A.5**). Primers were designed using Geneious Prime 2019.2.1, with the DV-1-1-Nuc gene, in pDONR221 plasmid, as a template. Primers were designed as above. A two-step PCR protocol was used for constructs with new start sites (DV-1-1Nuc P1, with DV-1-1Nuc Forward 2 and DV-1-1Nuc Back 2) and (DV-1-1Nuc P2, with DV-1-1Nuc Forward 2 and DV-1-1Nuc Back), while a one-step PCR protocol was used for the construct with the original start site (DV-1-1Nuc P3, with DV-1-1 Forward PCR 2 and DV-1-1Nuc Back). PCR was performed and analysed as above. PCR products were cloned

into the Gateway vector, pDONR221, using the Gateway™ BP reaction kit and transformed into expression vectors as described above.

2.2.3 Polymerase Chain Reaction (PCR)

2.2.3.1 Primers

Primers for Gateway cloning were designed using Geneious Prime 2019.2.1, following the manufacturer's instructions. Primers were supplied by IDT (USA). Primers were supplied purified with standard desalting techniques and were reconstituted in 1x TE to a concentration of 100 µM with working stock concentrations prepared at 10 µM in MQ H2O. Primers for PCR can be found in **Appendix A.4**.

2.2.3.2 PCR for amplification of gene constructs

PCR with HotFire Pol (Solis Biodyne) was used for the amplification of new gene constructs. For each set of primers, a gradient PCR was carried out to determine the optimal T_m for the PCR reaction. Five °C above and below the calculated T_m of the primers were selected as the upper and lower limits of the gradient. A T_m that resulted in a product of the expected size was selected for use in PCR for amplification. PCR reactions were carried out in 25 µl volumes using the following concentrations. 1x HotFire Pol blend, 0.3 µM for each F and B primer, 5-50 ng/µl of template DNA and up to 20 µl of H_2O. Reactions conditions for PCR are detailed in **Appendix A.6.** PCR products were viewed on 1 % TAE (40 mM TRIS-acetate, 20 mM EDTA) agarose gel containing 1 x SYBR Safe stain. PCR product size was determined by comparison to Invitrogen 1 Kb Plus DNA ladder (Thermo Fisher Scientific, USA). Products of the expected size were excised from an agarose gel and purified.

2.2.3.3 Colony PCR

Colony PCR was used to test positive transformants for the correct insertion of vector insert DNA. Positive transformants colonies were resuspended in a dilute antibiotic solution and used as DNA template for PCR using MBP F or T7 and M13 primers. PCR for amplification of genes inside pDONR221 plasmids

required M13 forward (F) and back (B) primers, and an annealing T_m of 55 ºC for all reactions. While PCR on pHMGWA and pDEST17 required T7 B and F primers, with MBP F also used at times for pHMGWA plasmid, all with an annealing T_m of 52.5 ºC. PCR reactions were carried out in 25 µl volumes using the same reaction concentrations as above, with 1 µl of resuspended colony template. PCR products were visualised on 1 % agarose gel, as above.

2.2.4 Agarose Gel Electrophoresis

DNA fragments were separated using agarose gel electrophoresis. The percent of agarose used in the gel depended on the size of DNA; as a general rule samples in the range of 400-1000 bp were run on a 1 % agarose gel in 1x TAE buffer (40 mM Tris-acetate, 20 mM EDTA). Samples were mixed with 5 x DNA loading dye (25 % glycerol, 0.2 % bromophenol blue) prior to loading onto the gel. Agarose gels were stained with 1 x SYBR Safe™ DNA gel stain (Invitrogen, USA). Gels were visualized on an iBright imager (Invitrogen) using the fluorescent stained nucleic acid gels setting and images captured. Band sizes were determined by comparison with the 1kb-Plus DNA ladder (Invitrogen, USA).

2.2.5 Extraction of recombinant DNA plasmids from *E. coli*

Recombinant DNA plasmids from *E. coli* were extracted from overnight cultures. 5 ml LB and antibiotic mixture was inoculated with the glycerol stock for each protein and grown overnight at 37 °C, 180 rpm. Overnight cultures were centrifuged at 4500 g, for 10 minutes, at 4 °C, and plasmid extracted using the QIAprep Spin Miniprep Kit (Qiagen, Germany) according to the manufacturer's instructions. Vector DNA was eluted in 50 µl of elution buffer.

2.2.6 DNA Quantification

DNA was quantified using a NanoDrop (Thermo Fisher Scientific, USA). This measures absorbance of DNA at 260 nm, quantifying the amount of DNA.

2.2.7 DNA sequencing of plasmids

To confirm the sequence identity of each recombinant protein the plasmid was purified, as described above, and sequenced between T7 and/or MBP promotor sequences. Purified plasmid was diluted to 250-625 ng with 4 pmol of each T7 forward or MBP forward, and T7 reverse primer and sequenced using Sanger sequencing (Massey Genome Service, New Zealand). Returned sequences were trimmed, aligned, and mapped to reference sequence using Geneious Prime (Geneious Prime 2019.2.1 (https://www.geneious.com), Biomatters Ltd, New Zealand) to ensure target DNA sequence was accurate and free of mutations before expression.

2.3 Expression of recombinant proteins

Media used for expression of recombinant proteins can be found in **Appendix A.2**.

2.3.1 Starter cultures

Starter cultures were prepared from positive transformants, by inoculating a single colony from a transformation plate into 10 ml of LB broth supplemented with 50 µg/ml kanamycin or ampicillin and incubated at 37 °C, shaking at 200 rpm, overnight. Glycerol stocks for long term storage of transformed *E. coli* Bacterial strains (Origami (DE3), pLysS BL21 (DE3), arctic express and BL21 (DE3)) were made by the addition of 0.5 ml of overnight culture to 0.5 ml of sterile 50 % (v/v) glycerol. Glycerol stocks were stored at -80 ºC.

2.3.2 Small scale expression growth trials

To determine optimal production conditions, expression from each plasmid was tested in the *E. coli* expression strains pLysS BL21 (DE3) (Novagen), Origami (DE3) (Novagen) and ArcticExpress (DE3) (Aligent) at 15 and 25 ºC. For small-scale expression trials, 5 ml seeder cultures in 50 ml tubes were grown overnight with LB media, 1 µl of glycerol stock, 100 µg/ml ampicillin at 37 ºC with shaking at 180 rpm. Following this, 1 ml of cultures was added to 50 ml conical flasks, with TB media, 100 µg/ml ampicillin at 37 ºC with shaking at 180 rpm. Upon reaching an OD_{600} between 0.3 and 0.4 the temperature was adjusted,

and cells were equilibrated for 30 minutes before addition of 0.5 mM isopropylthio-βD-galactosidase (IPTG) to induce expression. Cells were harvested after 18 hours by centrifugation, resuspended in 5 ml lysis buffer (50 mM Tris pH 8.0, 750 mM NaCl, 1mM MgCl2, 5% glycerol), and lysed by sonification on ice. Insoluble material was pelleted by centrifugation and the soluble fraction was incubated with 20 µl of nickel beads (Cytiva) for 15 minutes and recovered by centrifugation followed by two washes with lysis buffer. Protein-bound nickel beads, insoluble fractions and soluble fractions were electrophoresed on 12 % SDS-PAGE gels with successful expression being indicated by a strong band at the expected molecular weight in the nickel-bead fraction and, in some cases, the soluble sample. Expression as inclusion bodies was indicated as a band, of the correct molecular weight, in the insoluble samples.

2.3.3 Large scale expression growth

Optimal conditions (**Appendix A.9**) were used for large scale cultivation for purification. Large scale expression cultures were prepared by inoculating 10 ml of starter culture (prepared as per **Section 2.3.1**) into 1 L baffled conical flasks, with TB broth, supplemented with 50 µg/ml ampicillin. Cultures were incubated at 37 ºC, 180 rpm, until the culture reached an OD_{600} between 0.3 and 0.4. Upon reaching the desired OD_{600}, the temperature was adjusted to the appropriate temperature as identified from the small-scale cultures, and after a 30-minute equilibration period, protein expression was induced by the addition of IPTG (0.5 mM final concentration). Cultures were grown for 18 hours and harvested by centrifugation, with the supernatant being discarded and the pellet was stored at -80 ºC.

2.4 Protein purification and identification of target protein

Target proteins were isolated and purified using immobilized metal affinity chromatography (IMAC), maltose binding protein (MBP) affinity chromatography and gel filtration chromatography. Components for buffers can be found in **Appendix A.7**.

2.4.1 Cell lysis

Frozen protein cell pellets were thawed at room temperature. Cell pellets were resuspended by vortexing and shaking, in 25 mls of lysis buffer. Resuspended cells were lysed by sonication (QSonica, 12-amp, 3-minute total processing time, 1 second on, 1 second off pulses), in a Bioscience cool block. Cell debris was removed by centrifugation at 9072 g, for 1 hour at 4 ºC to isolate protein containing supernatant. The supernatant was filtered through 1.2, 0.45 and 0.2 µM filters (Minisart syringe filters; Sartorious AG, Germany).

2.4.2 Immobilised metal affinity chromatography

IMAC purification was carried out using either an ÄKTA Prime fast protein liquid chromatography (FPLC) system, at 4 ºC (GE Life Sciences) or an NGC FPLC system (BioRad), depending on protein stability. The filtered supernatant was either manually loaded, loaded using a 50 ml Superloop™ (Cytiva), or loaded using a sample pump, connected to the NCG system, onto a His-Trap™ HP 5 ml column (GE Life Sciences, New Zealand) that was pre-equilibrated in buffer A (50 mM Tris pH 8.0, 750 mM NaCl, 5% glycerol, 10 mM imidazole). Proteins with no affinity to NiNTA resin came off the column in the flow through. Weakly bound non-target proteins were eluted from the column with a solution comprising 4 % buffer B (50 mM Tris pH 8.0, 750 mM NaCl, 5% glycerol, 500 mM imidazole) : 96 % buffer A at a flow rate of 1 $ml.min^{-1}$. Some target proteins were often bound to chaperone proteins and were further washed with 4 % wash buffer (50 mM Tris pH 8.0, 800 mM NaCl, 40 mM imidazole, 30 % glycerol). Protein isolation from the IMAC column was achieved using an ÄKTA Prime, or purifier FPLC system using an isocratic gradient from 4-100 % elution over a 80-100 ml gradient.

2.4.3 MBP affinity chromatography

MBP-tagged protein samples that eluted off the IMAC column, with many contaminating *E. coli* proteins, were further purified with maltose binding protein (MBP) affinity chromatography. Fractions, containing target protein, from IMAC purification, were pooled and up concentrated to 5 mls in a 20 ml Vivaspin

concentrator (10 kDa molecular weight cut off; Sartorius AG, Germany) at 3,600 rpm, at 4 ºC. Up concentrated protein was buffer exchanged, into MBP binding buffer (20 mM Tris-HCl, 200 mM NaCl, 1 mM EDTA, pH 7.4), using a HiPrep 26/10 Desalting column (Cytiva), connected to an ÄKTA Prime FPLC system or an NGC FPLC system. Proteins eluted off column as a single peak before salt peak. Fractions containing target protein were pooled and loaded onto the MBPTrap™ HP Column, pre-equilibrated with MBP binding buffer. Proteins with no affinity to amylose, came off the column in the flow through. Weakly bound non-target proteins were eluted from the column with a solution comprising 4 % MBP elution buffer (20 mM Tris-HCl, 200 mM NaCl, 1 mM EDTA, 10 mM maltose, pH 7.4) : 96 % MBP binding buffer at a flow rate of 1 ml.min-1. Protein isolation from column was achieved using an isocratic gradient, with MBP elution buffer, from 4-100 % elution over a 80-100 ml gradient.

2.4.4 MBP tag removal with TEV protease

Fractions containing target protein from either IMAC purification or MBP affinity chromatography, were pooled, up concentrated and buffer exchanged, into buffer C (50 mM Tris pH 8.0, 200 mM NaCl, 1mM DTT, 5% glycerol), using a de-salting column, connected to either an ÄKTA Prime FPLC system or an NGC FPLC system. Fractions from the protein peak were pooled and incubated overnight at 4 ºC with 1 mg of TEV-protease, which had been produced in-house according to published protocols (Tropea et al., 2009). Cleaved proteins were subjected to reverse IMAC by re-application to the His-Trap column. Cleaved proteins either eluted in the flow through, or after addition of 4 % buffer B. To increase purity and remove high molecular-weight aggregates, the flow through which contained the untagged target protein was up-concentrated to less than 5 ml volume and loaded onto either a Superdex200 16/600 (S200) or Superdex75 16/600 (S75) column. In cases where the tag was not removed by TEV cleavage, fractions from the initial gradient elution in Buffer B were pooled and used directly in gel filtration chromatography.

2.4.5 Gel filtration chromatography

Fractions containing target protein from a reverse IMAC purification, were pooled and concentrated to <1 ml volume in a 20 ml Vivaspin concentrator. Concentrated protein was filtered to 0.2 µM and injected onto a Superdex200 16/600 (S200) or Superdex75 16/600 (S75) column (Cytiva) pre-equilibrated with buffer C. Protein was separated and eluted with buffer C at a flow rate of 0.5 ml.min-1 and collected in 1 ml aliquots. Fractions containing protein were identified by following 280 nM wavelength trace. Final purified protein was up concentrated to 0.5 -5 mg ml-1, mixed 50:50 v/v with glycerol 243 and stored at -80 °C.

2.4.6 Analytical size exclusion

The analytical size exclusion column (ENrich™ SEC 650 10 x 300 column, Bio-Rad Laboratories, USA) was first equilibrated with the protein specific size exclusion buffer, then calibrated with 1 mg/ml Blue Dextran and Gel filtration standard (Bio-Rad Laboratories, USA) that were reconstituted following manufactures recommendations. 250 µl of purified recombinant protein was run through the analytical size exclusion column, and the elution volume of protein was used against the standard curve to determine protein size and oligomeric arrangement.

Blue Dextran and Gel filtration standards were used to calculate the void volume and a standard curve of protein size respectively. A calibration curve was then determined for the column by calculation of the Kav (gel phase distribution coefficient) values for the calibration kit proteins using the equation: Kav = (Ve - Vo) / (Vc - Vo) (Vo = column void volume, Ve = elution volume, Vc = geometric column volume). The calibration curve was made by plotting Kav against the log molecular weight and an equation determined to calculate the molecular weight of sample proteins.

For DV-1-1-Lig protein the K_{av} (gel phase distribution coefficient) was determined using the following equation:

$K_{av} = (Ve - Vo) / (Vc – Vo)$

Where

Ve = elution volume

Vo = column void volume = 9.78 ml

Vc = geometric column volume = 24 ml

The K_{av} value for the protein was then substituted into the equation derived from the curve to determine the MW of DV-1-1-Lig protein. The equation from the calibration curve (y = -0.09x + 1.2751) was rearranged to calculate the MW of DV-1-1-Lig protein.

2.4.7 SDS-Polyacrylamide Gel Electrophoresis (SDS-PAGE) protein analysis

PAGE gels were used routinely to assess purity and quality of recombinant expressed proteins. SDS-PAGE gels were cast in a Hoefer gel casting system. SDS-PAGE gels consisted of a 5 % acrylamide stacking gel overlaid on a 12 % acrylamide resolving gel. All SDS-PAGE gels were made up with 30 % acrylamide with an acrylamide:bisacrylamide ratio of 37.5:1 (Bio-Rad Laboratories, USA) and included 0.1 % filtered SDS, and were polymerised by the addition of 0.05 % (w/v) ammonium persulphate (APS) and 0.5 % (v/v) TEMED.

Protein samples were mixed in a 3:1 ratio with 4 x SDS loading buffer (250 mM Tris HCl pH 6.8, 20 % glycerol, 4 % SDS, 10 % β-mercaptoethanol, 0.025 % (w/v) bromophenol blue) and denatured for 5 min at 95 °C prior to loading onto the gel. Gels were run with 1 x SDS-PAGE running buffer (25 mM Tris, 250 mM glycine, 0.1 % (w/v) SDS) at constant 80 V until the samples entered the resolving gel, then at 150 V until the dye front reached the end of the gel. For protein MW estimation, 10 µl of Precision Plus Protein Standard (Bio-Rad Laboratories, USA) was run alongside each gel.

Coomassie Blue Stain for Protein Gel Electrophoresis Gels were stained by colloidal Coomassie staining using the quick stain method (Wong et al., 2000) with Fairbanks staining solutions A-D. Gels were microwaved for 30 s in 50 ml

Fairbanks staining solution A (0.05 % Coomassie, 25 % isopropanol, 10 % acetic acid), cooled to room temperature while shaking gently. Stain A was removed, and 50 ml of Fairbanks staining solution D (10 % acetic acid) was added and left for 1 hr while shaking gently.

2.4.8 Measurement of protein concentration

Protein concentration was measured using NanoDrop (Thermo Fisher Scientific, USA). The Nanodrop measures absorbance at 280 nm and the accompanying software calculates protein concentration using the Beer-Lambert equation: $A=\varepsilon.c.l$.

The theoretical molar extinction coefficients were calculated using the online tool ProtParam (http://web.expasy.org/protparam), by providing the amino acid sequence.

2.5 LC-MS/MS

MS3 Solutions Limited, at Ruakura Research Centre, performed LC-MS/MS on trypsin digested DV1-1-Nuclease protein band. DV1-1-Nuc protein was run on a 12 % SDS-PAGE and stained with Blue Safe Protein Stain from Thermo Scientific™. The protein band was excised from the gel and processed using a standard in-gel tryptic digestion method. Tryptic peptides were extracted and detected by LC-MS/MS using an Orbitrap Q-Exactive Hybrid mass spectrometer. Spectral data was searched against a custom-made consisting of *E. coli* protein sequences obtained from UniProtKB/Swiss-Prot and DV1-1-Nuc protein sequence.

2.6 Protein Crystallisation

2.6.1 Initial Crystallisation screens

Crystallisation trials were carried out with proteins, purified in 50 mM Tris buffer pH 8, 200 mM NaCl, 10 % glycerol and 2 mM DTT, using the following Hampton Research crystallisation screens (Hampton Research, USA): Crystal (HR2-130), PEG Rx (HR2-086), Salt Rx (HR2-136), Natrix 2 (HR2-116) and

Index (HR2-134). One hundred microlitres of each screen condition were pipetted into four 96-well Intelli-plates (Hampton Research, USA). A Mosquito® crystallisation robot (TTP LabTech Ltd, UK) was used to dispense protein and reservoir solutions (100 nl each) into the crystallisation well, and the plates were sealed with ClearsealTM film (Hampton Research, USA). Screens were left at 18 ºC.

Screens were also performed for DV-1-1-Lig purified in the same buffer conditions as above, with the addition of a 21 bp DNA oligo, with a symmetrical nick (OMC), as previously described by (Nair et al., 2007). The singly nicked 26 bp duplex was formed by annealing three strands: a 13-mer 5'PO_4 DNA strand (5'-pCACTATCGGAATG3'), a 13-mer 3'OH strand (5'ACAATTGCGACCOMe C3′), and a complementary 26-mer DNA template strand (5' CATTCCGATAGT GGGGTCGCAATTGT-3′). An equal amount of each strand was mixed and heated to 85 ºC, in a buffer containing 10 mM Tris (pH 8.0), 50 mM NaCl, and 1 mM EDTA, then cooled to room temperature overnight and stored at -20 ºC. Purified protein (8-15 µM) and OMC oligo were mixed at a 1:1:2 ratio, with the addition of 5 mM EDTA. Sample was incubated for 30 minutes on ice, followed by use in crystallisation screens, as described above.

2.6.2 Fine screens by Hanging drop vapor diffusion

When promising crystallisation conditions were identified from the initial crystallisation screens, fine screens were performed to optimise the condition for protein crystal growth.

DV-1-1-Lig protein bound to OMC DNA oligo, formed small crystals in robot screens, with natrix conditions and was also used in fine screens, as described below.

For fine screens, a hanging drop vapour diffusion technique was used. For hanging drops, the tops of the wells of a Crystalquick 24-well plate (Greiner Bio-one, Germany) were greased with glisseal grease and 500 µl of each fine screen condition was added to the wells. One microlitre of DV-1-1-Lig (6 µM), mixed at

a 1:1:2 ratio with OMC oligo was added to the reservoir solution, at a 1:1 ratio and pipetted onto a 22 mm cover slip (siliconised glass, square) (Hampton Research, USA). The cover slip was then inverted and placed over the reservoir solution and pressed down gently to seal. Screens were left at 18 ºC.

Table 2.1. Natrix2 conditions used in fine screens with DV-1-1-Lig, mixed with DNA substrate OMC.

Reagent	Composition
H1	0.08 M Sodium chloride, 0.02 M Barium chloride dihydrate, 0.04 M Sodium cacodylate trihydrate pH 7.0, 45% v/v (+/-)-2-Methyl-2,4-pentanediol, 0.012 M Spermine tetrahydrochloride
H2	0.08 M Potassium chloride, 0.02 M Barium chloride dihydrate, 0.04 M Sodium cacodylate trihydrate pH 7.0, 40% v/v (+/-)-2-Methyl-2,4-pentanediol, 0.012 M Spermine tetrahydrochloride
H3	0.08 M Potassium chloride, 0.02 M Barium chloride dihydrate, 0.04 M Sodium cacodylate trihydrate pH 6.0, 40% v/v (+/-)-2-Methyl-2,4-pentanediol, 0.012 M Spermine tetrahydrochloride
H5	0.1 M Potassium chloride, 0.05 M Sodium cacodylate trihydrate pH 6.0, 16% w/v Polyethylene glycol 1,000, 0.0005 M Spermine
F3	0.08 M Sodium chloride, 0.02 M Magnesium chloride hexahydrate, 0.04 M Sodium cacodylate trihydrate pH 6.0, 35% v/v (+/-)-2-Methyl-2,4-pentanediol, 0.012 M Spermine tetrahydrochloride
F4	0.08 M Strontium chloride hexahydrate, 0.04 M Sodium cacodylate trihydrate pH 6.0, 35% v/v (+/-)-2-Methyl-2,4-pentanediol, 0.012 M Spermine tetrahydrochloride
F5	0.08 M Potassium chloride, 0.02 M Barium chloride dihydrate, 0.04 M Sodium cacodylate trihydrate pH 7.0, 40% v/v (+/-)-2-Methyl-2,4-pentanediol, 0.012 M Spermine tetrahydrochloride
G7	0.08 M Strontium chloride hexahydrate, 0.02 M Magnesium chloride hexahydrate, 0.04 M Sodium cacodylate trihydrate pH 7.0, 20% v/v (+/-)-2-Methyl-2,4-pentanediol, 0.012 M Spermine tetrahydrochloride

For streak seeding, hanging crystallisation drops were set up as above and left to equilibrate for 2-3 hours before seeding. A dry cat's whisker cleaned with 70 % ethanol was streaked through a source drop containing crystals and then brushed through the new pre-equilibrated drop. The cover slip was then inverted and placed over the reservoir solution and pressed down gently to seal.

2.6.2.1 Testing of Crystals by X-Ray Diffraction

Prior to testing for diffraction, crystals were removed from their drop using a nylon cryoloop (Hampton) and moved into a cryo protectant solution which was made up of the crystallisation solution plus 5-20 % glycerol, before being snap

frozen in liquid nitrogen. X-ray diffraction data collected at the Australian synchrotron on beamline MX1, equipped with a ADSC Quantum 210r detector (Area Detector Systems Corporation, USA) with λ=0.9537 Å.

2.7 Circular Dichroism spectroscopy

Proteins were dialyzed into CD buffer (10 mM potassium phosphate pH 8.0, 100 mM sodium fluoride) at 4 ºC overnight. Protein concentrations were between 1 and 2 uM and were checked by nanodrop after clarification by centrifugation immediately prior to CD measurement. Spectra were measured on a Jasco 815 circular dichroism spectrophotometer located at the Biomolecular Interactions Centre (University of Canterbury, Christchurch New Zealand). Wavelength scans were collected using a 1 mm pathlength quartz cuvette with constant temperature control via the cell holder coupled to a Peltier device. Wavelengths from 190-260 nm were scanned at 0.2 nm intervals with 0.2 sec averaging, scanning speed of 20 nm/min and bandwidth set to 1nm. A series of three scans were collected and averaged. Data was processed using the instrument software to subtract the buffer baseline and smoothed. Data was truncated to remove portions of the scan where the high tension voltage exceeded 700 V. Denaturation curves were measured using equivalent instrument settings, but at a constant wavelength of 222 nm. Temperature control was provided by the Peltier controller and was set at a 5 min pre-incubation of either 5 or 20 ºC, followed by a temperature ramp of 2 ºC per min from 20 to 90 ºC. Three melt curves were measured and averaged.

Data from protein circular dichroism (CD) spectra was analysed using the single spectrum analysis from BeStSel database (Micsonai et al., 2018). This analysis is used for secondary structure determination, distinguishing parallel β-sheets and β-sheets of different twists, and fold recognition.

2.8 Differential Scanning Fluorimetry

Thermal stability measurements of proteins were carried out by differential scanning fluorimetry (DSF), using SYPRO Orange (Life Technologies, USA) as previously described (Ericsson et al., 2006). Protein and SYPRO Orange concentrations were held at 1-4 mg/ml and 50 x respectively. Each reaction was

run in triplicate along with blank reactions, containing no protein. Unfolding over a temperature range from 25 to 98 ºC was measured, with continuous monitoring of fluorescence (excitation 470, emission 550 nm) using a RotoGene Q thermocycler (Qiagen). For testing protein stability in different pH conditions, the pH was adjusted by dilution into Britton-Robinson universal buffers over a range of 5.0 to 9.5. For metal stability testing magnesium, manganese and zinc were added to the reactions at a final concentration of 10 mM. The data were plotted as the first derivative in Graphed prism and the T_m, i.e., the midpoint of the transition was taken as the highest point. DSF thermal melts were presented in GraphPad Prism 9.0 (GraphPad Software, USA), with T_m values indicated by dotted lines or plotted up against different variables.

2.9 DNA binding and nuclease activity assays

Enzymatic activity on DNA ligase substrates, DNA nuclease substrates, DNA-damage substrates, and flapped/splayed substrates (**Appendix A.10.2**) were analysed by denaturing gel electrophoresis, while binding experiments were analysed by native PAGE. Substrates were generated from oligonucleotides (IDT) with synthetically incorporated DNA damages and fluorescent labels, which are listed in **Appendix A.10.1**. These were annealed in the combinations given in **Appendix A.10.2** as described previously (Sharma et al., 2020) using final concentration of 80 nM for the 5' FAM-labelled probe strand and 112 nM for the unlabeled strand (s) (damage, nuclease and double-strand break ligase substrates) or 400 nM unlabeled strands (nicked ligase substrates).

All assays were carried out in in 50 mM Tris pH 8.0, 50 mM NaCl, 10 mM DTT buffer solution. Standard enzyme activity assays also included 10 mM magnesium, 10 mM manganese or 10 mM zinc, unless otherwise specified. While for binding experiments this was replaced with 5 mM EDTA. For DNA ligase assays 1 mM of nucleotide cofactor (ATP, ADP, GTP or NAD) was used, unless otherwise specified. Reactions were initiated by addition of protein and were incubated at the temperatures and time, as indicated in figure captions.

For the DNA ligase nucleotide cofactor preference experiments, enzyme was pre-incubated with unlabeled nicked DNA substrate (**Appendix A.10.2**) for 2

hours at 20 ºC to ensure any enzyme purified in the pre-adenylated state was turned over. After this time, labeled DNA substrate and 1 mM of cofactor (ATP, ADP, GTP or NAD) was added to the reaction and incubated for indicated times in the figure captions.

Activity assays for damage, flapped/splayed, nuclease and ligase substrates were quenched by addition of Quench Buffer to give a final concentration of 25 % formamide, 20 mM EDTA, 0.05 % bromophenol blue and heated to 95 ºC for 5 minutes, before electrophoresis on 20 % denaturing TBE urea gels (20 % acrylamide/Bis-acrylamide 29:1, 7 M urea, 1x TBE). For binding assays (EMSA), native loading buffer was added (20 mM EDTA, 0.05 % bromophenol blue, 5 % glycerol, EDTA, 0.05 % bromophenol blue, 5 % glycerol) before the samples were electrophoresed on a 10 % TBE gel (10% acrylamide/Bis-acrylamide 29:1, 1x TBE). Gels were visualized on an iBright imager (Invitrogen) using the fluorescein setting. Results of ligation or nuclease activity (only for single cut products) were quantified using Image J software (Rasband, 2011) and results were displayed as graphs using GraphPad Prism, version 8 for Windows.

3 Chapter 3

DNA ligases

3.1 Introduction

As stated in **Section 1.8.1**, DNA ligases play a vital role in the diverse processes of DNA replication, recombination, and repair across all domains of life. It was initially thought that bacteria only possessed NAD-dependent DNA ligases, however it has been discovered over the last several years that some bacteria also contain several types of ATP dependent DNA ligases (Pergolizzi et al., 2016; Shuman & Lima, 2004; Williamson et al., 2016).

All bacterial species have at least one NAD-dependent DNA ligase (Lig A) and some species also contain a second NAD-dependent DNA ligase, referred to as Lig B (Sriskanda et al., 2001; Sriskanda & Shuman, 2001), not to be confused with the ATP-dependent Lig B DNA ligases. The catalytic core of bacterial NAD-dependent ligases contains an adenylation (nucleotidyltransferases (NTase)) (AD) domain and an oligonucleotide /oligosaccharide binding (OB)-fold domain. This core is flanked by a short N-terminal domain (Ia) and three C-terminal domains: a tetracysteine domain that binds a single zinc atom (ZnF), a helix-hairpin-helix domain (HhH) and a BRCT domain, which gets its name after the C-terminus of the breast cancer gene product BRACA1 (Sriskanda & Shuman, 2001; Wilkinson et al., 2001). Lig B from *E. coli*, shares some of the same domains, to that of Lig A: the catalytic AD and OB core domains, as well as the Ia and HhH domains. However, Lig B lacks the C-terminal BRCT domain and two of the four Zn-binding cystines from the ZnF domain (Sriskanda & Shuman, 2001) (**Figure 3.1**).

In addition to the NAD-dependent DNA ligases, some bacterial genomes contain multiple genes for DNA ligases that are predicted to use ATP as their cofactor. These ATP-dependent DNA ligases display a wide diversity among bacteria, with some species possessing as many as five ligase encoding genes, while others, sometimes closely related species harbour none (Williamson & Leiros, 2020). One of the best-studied examples is *M. tuberculosis* which has three ATP-dependent DNA ligases Lig B, Lig C and Lig D. All three ATP-

dependent ligases of *M. tuberculosis* are capable of nick sealing, although their catalytic efficiency is lower than that of Lig A. Neither Lig B, Lig C or Lig D are essential for mycobacterial growth (Gong et al., 2004). Such ATP-dependent sub-classes of ligases have dedicated repair functions in bacteria and often play a role in alternative stationary-phase pathways in species that have spore forming or dormant life-phases (Pergolizzi et al., 2016; Williamson et al., 2016).

Currently, four distinct classifications of bacterial ATP-dependent DNA ligase homologs have been described based on their structural and functional characteristics (**Figure 3.1**) (Williamson & Leiros, 2020). The first belongs to the large, multi-domain DNA ligase, known as Lig D, which possesses multiple enzymatic functions. These multiple functions are owing to the different structural domains of Lig D that are: the polymerase domain for the addition of nucleotides, the phosphoesterase/nuclease domain that converts 3' phosphate groups to hydroxyl groups and the ligase domain to seal nicks in the DNA. The ligase domain, consists of an N-terminal AD domain and a C-terminal OB domain (Amare et al., 2021). There are different arrangements of the domains that make up Lig D ligases, across different species of bacteria and some Lig D ligases are missing the accessory polymerase and phosphoesterase/nuclease domains altogether (Williamson & Leiros, 2020). The group of Lig B ATP dependent DNA ligases, closely resemble archaeal replicative ligases with a common globular α-helical DNA binding (DB) domain preceding the AD domain and OB-fold domain (Williamson & Leiros, 2020). The last two groups, belong to the minimal ATP-dependent DNA ligases, Lig C and Lig E. Both Lig C and Lig E lack any accessory domains and do not contain an N-terminal DB domain. Lig E proteins display a periplasmic localisation sequence (PLS) (Williamson & Leiros, 2020). Unlike the Lig C and Lig D ligases, both Lig B and Lig E display high rates of nick sealing in the absence of any accessory enzymes (Gong et al., 2004).

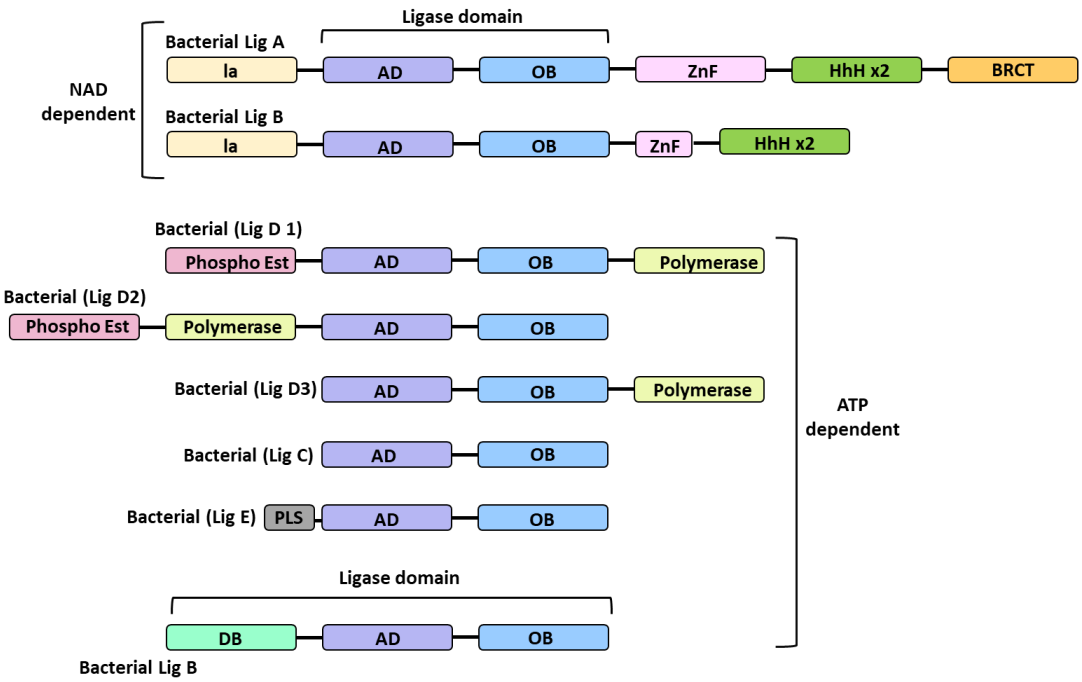


Figure 3.1. Schematic of domain arrangements in major classes of bacterial DNA ligases characterized to date: NAD-dependent DNA ligases (Lig A, Lig B), ATP-dependent DNA ligases with N-terminal helical DB domains and a common OB domain type (LigB), Lig-D type non-homologous end joining proteins with auxiliary polymerase and phosphoesterase domains, Lig C with no auxiliary domains and bacterial Lig E proteins that have a periplasmic localization sequence (PLS). Figure adapted from (Williamson & Leiros, 2020).

Lig B DNA ligases, are the most extensive class of bacterial ATP-dependent DNA ligases, having been discovered in species of *Mycobacterium* and *Pseudomonas*, as well as Cyanobacteria such as *Prochlorococcus marinus* (Berg et al., 2019; Ejaz & Shuman, 2018; Gong et al., 2004; Williamson et al., 2016). Members of this group have a highly modular architecture consisting of a unique arrangement of two or more discrete domains including, an N-terminal DNA binding (DB) domain, an adenylation (nucleotidyltransferases (NTase)) (AD) domain and an oligonucleotide/oligosaccharide binding (OB)-fold domain. The DB domain has been implicated in nicked DNA recognition in *M. tuberculosis* (Gong et al., 2004). The AD and C-terminal OB domains comprise a catalytic core unit that is common to most members of the ATP-dependent DNA ligase family. The common catalytic core unit contains six conserved sequence motifs (I, III, IIIa, IV, V, and VI) that links this family of related nucleotidyltransferases (Nishida et al., 2006). The RxDK motif (motif VI), which is essential for ATP hydrolysis, is in the OB domain (Nishida et al., 2006).

Lig B ligases are often found in gene clusters with a novel Lhr-helicase, binuclear metallophosphoesterase (MPE) and a putative exonuclease. The Lhr helicase and MPE proteins, from this gene cluster have been previously characterized in *P. putida*. Here, they described Lhr helicase as an ssDNA dependent ATPase, an ATP-dependent 3'-to 5' single-stranded DNA translocase and an ATP-dependent 3' to 5' helicase. While MPE is described as a manganese-dependent phosphodiesterase and DNA endonuclease. The ligase and exo nuclease components of this operon have not been biochemically characterised and their functions have been characterized *in silico*. The ligase is a predicted ATP-dependent DNA ligase, consisting of three domains, homologous to the equivalent domains of human DNA ligase I (Pascal et al., 2004). The putative nuclease has been classified as a putative nuclease of the metallo-β-lactamase (MBL) family (Ejaz & Shuman, 2018). Although the precise biological substrate and order of activity have not yet been determined, it is likely that these enzymes represent yet another distinct repair pathway in bacteria (**Figure 3.2**) (Williamson & Leiros, 2020).

The order of genes in this operon, has been previously classified as a class I cluster, as described by (Ejaz & Shuman, 2018). Variations of this gene cluster have been identified in other bacteria and class systems have been used to label the different arrangements, as seen in **Figure 3.2**. The different gene clusters, vary by additions of proteins unrelated to nucleic acid enzymology, inversion of the exo nuclease-ligase cassette and loss of the exo nuclease-ligase cassette entirely.

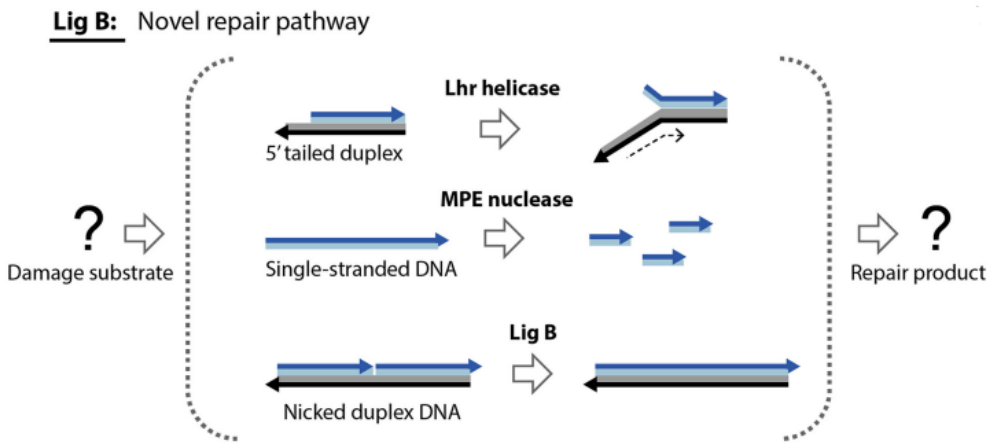


Figure 3.2. Probable repair pathway involving components of an operon including the ATP-dependent DNA ligase, Lig B. Figure adapted from (Williamson & Leiros, 2020).

Sequence similarity networks (SSNs) are used to identify relationships among protein sequences. In SSNs the most related proteins are grouped together in clusters. Tools like EFI-EST (Enzyme Function Initiative's Enzyme Similarity Tool) and cytoscape can be used to generate SSNs (Gerlt et al., 2015). The generation of a SSN involves two steps. First a set of sequences to analyse is chosen and an all by all BLAST search is performed to determine the similarity of sequences in the database, as a consideration of their relatedness. The second step involves filtering the sequences into clusters, based on a similarity threshold. When visualising an SSN, protein sequences are represented as nodes and the lines connecting two nodes is an edge. This edge is formed if the BLAST pairwise similarity score, between the connecting nodes is above a user defined threshold. As this value increases, the number of clusters increase, and only highly similar proteins will be grouped together (Oberg et al., 2023). Overall SSN analysis is an effective way to classify groups of proteins and identify unexplored sequences that may exhibit novel functionality (Atkinson et al., 2009).

A recent study (Williamson & Leiros, 2020) looked into the sequence diversity of DNA ligases available in public databases and used the Enzyme Function Initiative's Enzyme Similarity Tool (EFI-EST) to generate sequence similarity networks (SSNs) for protein sequences including the catalytic AD of either the ATP-dependent-ligases (Pfam 01068), or the NAD dependent-ligases (Pfam 01653).

Here they identified four different clusters of ATP-dependent ligases in bacteria (**Figure 3.3**). Of interest are the two different clusters of Lig B type DNA ligases. Cluster #1 is taxonomically diverse and includes all archaeal replicative proteins in the dataset, with the majority of the eukaryotic ligase I enzymes and around half the bacterial Lig B representatives. These ligases share a common central DB-AD-OB domain arrangement, and all characterised members of these groups possess independent ligase activities that do not require additional scaffolding proteins. The co-clustering arrangement of the bacterial Lig B proteins, with the majority from *Actinobacteria, Acidobacteria* and *Chloroflexi,* may support the theory that these accessory ligases were horizontally acquired from Archaea (Williamson et al., 2016). There is a second Lig B cluster #4, that

consists mostly of ligases from *Proteobacteria*. This suggests that not all bacterial Lig B ligases come from the same ancestor and may have arisen from multiple acquisition events (Williamson et al., 2016).

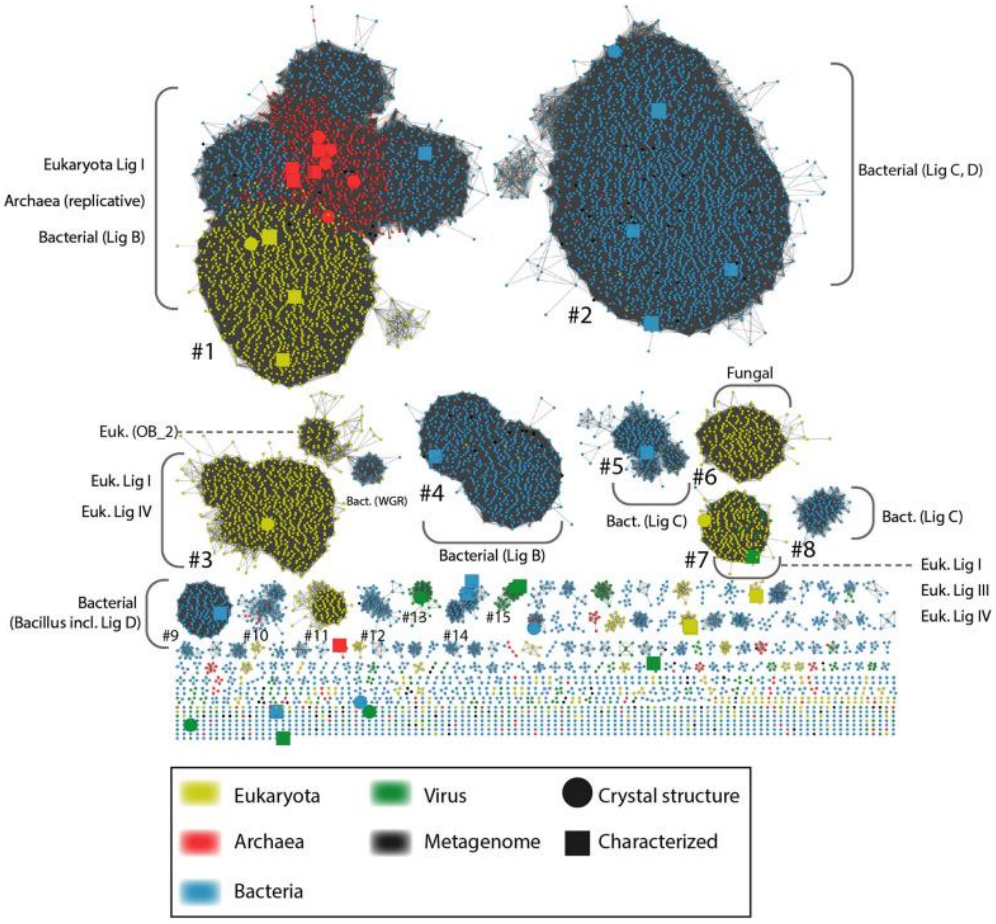


Figure 3.3. Sequence similarity network (SSN) of ATP-dependent DNA ligases coloured by super kingdom. Figure adapted from (Williamson & Leiros, 2020).

3.2 Results

3.2.1 *In silico* characterisation and homology modelling of Lig-B homologs

The Dry Valley metagenomes encode a plethora of DNA ligase proteins identified by the methods described in (Rzoska-Smith et al., 2023) and **Section 1.11**. Preliminary analysis of the Dry Valley metagenomes found a number of these DNA ligase proteins with an N-terminal DNA-binding domain of the LigB class DNA_ligase_A_M (PF04675). Sequence similarity networks conducted on these ligases, showed that they grouped with other ligases from UniRef, when the similarity was set to a 50 % threshold (**Figure 3.4**) (Rzoska-Smith et al., 2023).

Several candidate Lig B DNA ligases, from the Dry Valley metagenomes, were selected from each of three main clusters that formed from the sequence similarity networks. These ligases were selected from the basis that they belonged to a gene contig with other DNA modifying proteins in close proximity, or they contained an interesting domain (Rzoska-Smith et al., 2023). From the five candidate DNA ligases identified, three of these ligases were selected for *in silico* and biochemical characterisation, due to time constructs on the project. The three selected DNA ligases, identified from sequence similarity networks, are indicated by green circles in **Figure 3.4**. The three ligases were named DV-Lig2, DV-Lig5 and DV-1-1-Lig-Nuc based on them belonging to metagenomes from the Dry Valleys (DV) of Antarctica and either the order of cloning (DV-Lig2, DV-Lig5) or their position in a gene cluster (DV-1-1-Lig-Nuc).

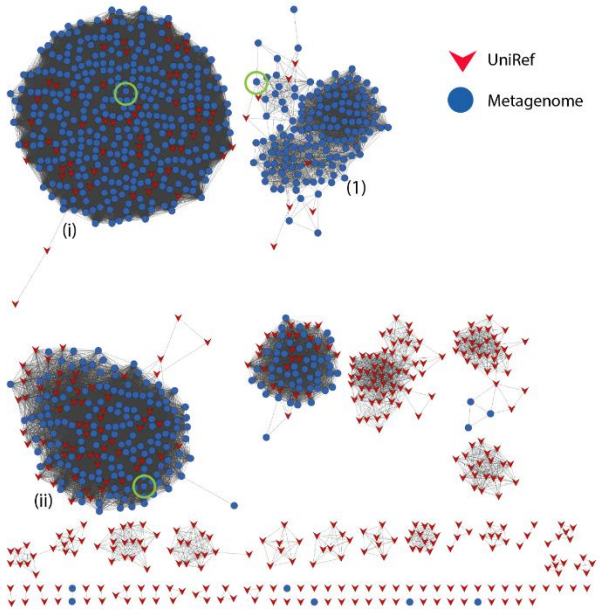


Figure 3.4. Sequence similarity network of metagenome hits to LigB-type DNA ligases at 50% identity edge threshold. Sequence similarity networks were constructed for each set of sequences identified by hmmsearch using the EFI-EST server. Domain compositions include the catalytic DNA_ligase_A_M domain together with the N-terminal DNA binding domain DNA_ligase_A_N. Dry-Valley metagenome nodes are coloured blue, UniRef50 nodes are indicated in red. The sequences used in further analysis on Clusters (1), (i) and (ii) are indicated by a green circle. Figure adapted from (Rzoska-Smith et al., 2023).

DV-Lig2 and DV-Lig5 are both typical 'ligase-only' Lig B enzymes from each of the major SSN groups and are both found in gene clusters with other putative nucleic acid binding proteins. Of note, a nuclease-ligase fusion protein, DV-1-1-Lig-Nuc with similarity to the nuclease ligase fusion protein from

Opitutus terrae was also identified in the Dry Valley metagenomes and is further discussed in **Chapter 4**.

DV-Lig2, from the predicted *Stenotrophomonas rhizophila* lineage, is part of a co-oriented four gene cluster comprising a putative exonuclease of the metallo-β-lactamase enzyme family, an ATP-dependent DNA ligase, a DNA helicase Lhr, and a member of the binuclear metallo-phosphoesterase (MPE) enzyme family (**Figure 3.5**). The order of genes in this cluster, belong to the class I cluster as described by (Ejaz & Shuman, 2018).

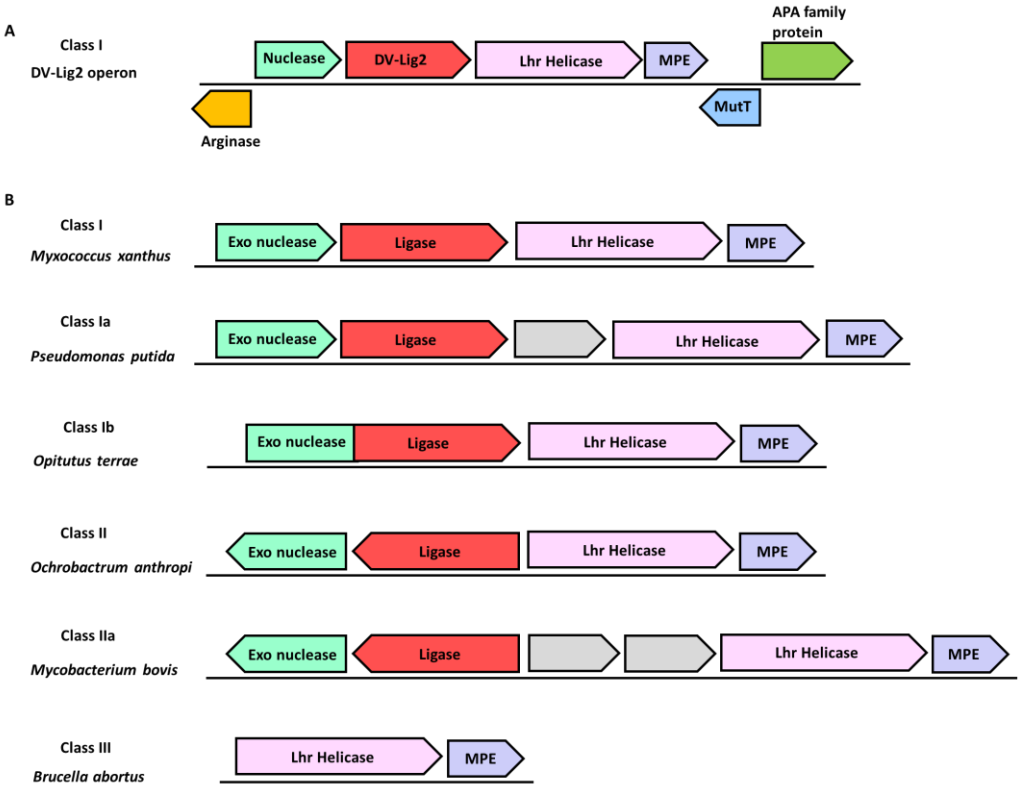


Figure 3.5. Genetic clustering of Lig B type DNA ligases in gene clusters with three other putative nucleic acid repair enzymes. **A**) DV-lig2 Lig B ligase is in a class I type operon arrangement, where a nuclease (green), ligase (red), helicase (pink) and MPE (purple) are transcribed together. **B**) Different bacterial species, representing the different classes of operon arrangements for these four main enzymes. The order and orientations of the clustered genes may vary among taxa and are classified accordingly in the figure. A bacterial species exemplifying each class is indicated on the left. Where additional, functionally unrelated ORFs are inserted into the gene cluster, these are denoted by gray arrows. The cases in which only Lhr-Core and MPE comprise a two-gene cluster are designated as class III in the schematic. Figure is adapted from (Ejaz & Shuman, 2018).

AlphaFold2 predictions (as discussed in **Section 2.1**) were generated for each of the four DNA repair genes from the DV-operon and from *P. putida* operon. Here proteins from the DV-operon were super imposed onto the matching

protein from *P. putida* operon. There is structural homology observed between proteins from the DV-operon and their protein complement from the operon of *P. putida*. RMSD values are indicated for each super-imposition in **Figure 3.6**. The greatest similarity was seen between MPE and exo nuclease proteins, shown by low RMSD values.

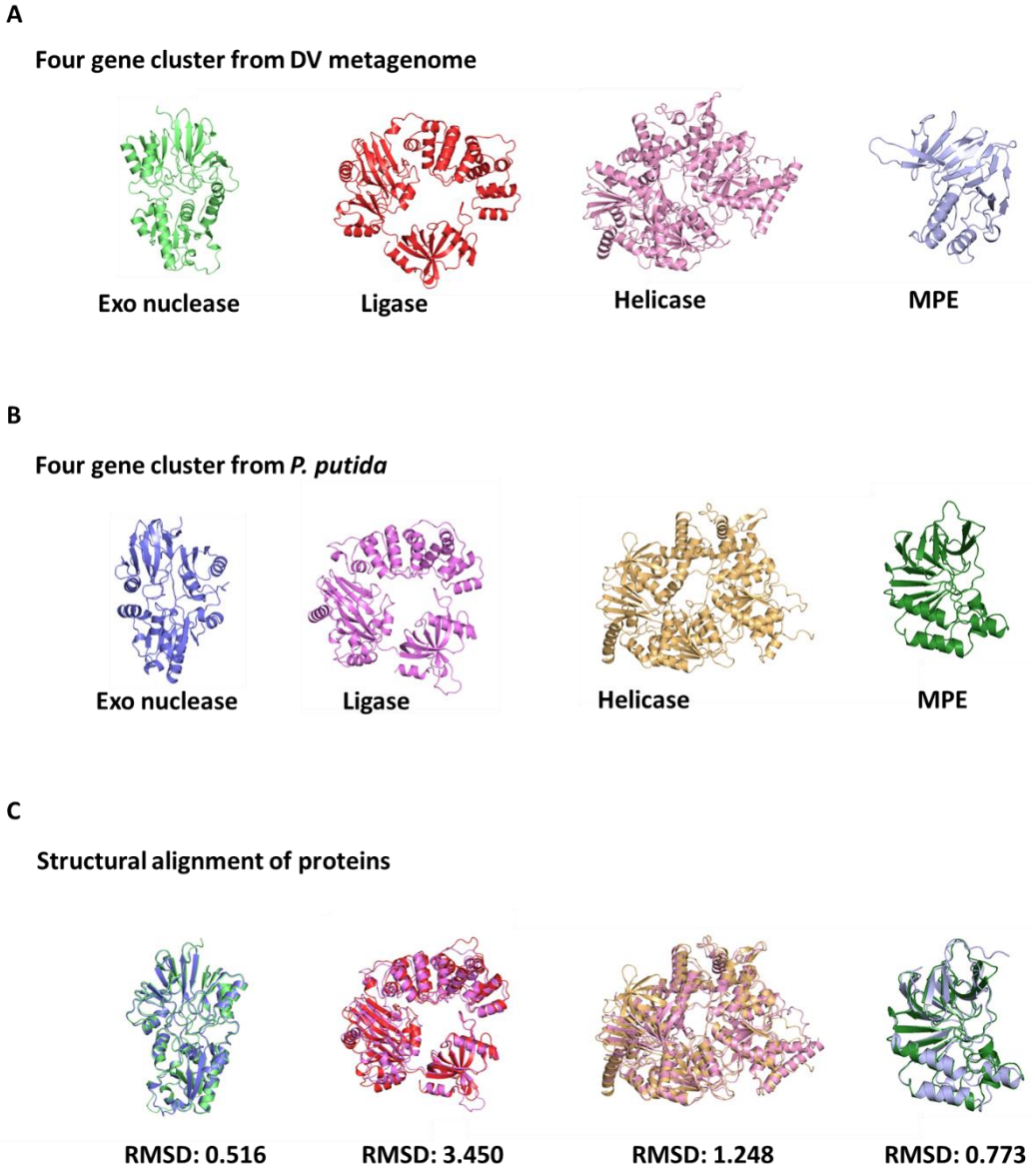


Figure 3.6. AlphaFold2 predicted structures of four DNA repair proteins from DV-metagenome and *P. putida*. **A**) AlphaFold 2predicted structures of exo nuclease, ligase, helicase and MPE proteins from a four gene cluster operon, from DV-metagenome. **B**) AlphaFold2 predicted structures of exo nuclease, ligase, helicase and MPE proteins from a four gene cluster operon, from *P. putida* genome. **C**) Super-imposition of the four DNA repair proteins from DV-Metagenome and *P. putida* genome, with RMSD values shown below. Protein predicted models were generated by AlphaFold2, from Google Colab, version v2.3.1 (John Jumper, 2021; Varadi et al., 2022) and all structures were presented in PyMOL (Schrödinger, 2020).

DV-Lig5 is found in a gene contig with other nucleic acid processing enzymes (DNA primase, DNA ligase), as well as several hypothetical proteins (**Figure 3.7**). Examination of other gene contigs containing LigB homologs from the DV-metagenomes or other bacterial genomes, found that this arrangement of genes up-stream and down-stream of DV-Lig5 was only observed in the DV-metagenome UQ272 (data not shown). DNA primases are enzymes that catalyse the synthesis of short oligonucleotides that act as a primer for DNA polymerase (Lao-Sirieix et al., 2005). The DNA primase identified here, has sequence similarity to the polymerase domain, of Lig D DNA ligase from *M. tuberculosis* (Shuman & Glickman, 2007). Along with DV-Lig5 DNA ligase, there is also another ligase present in gene contig, which shares sequence similarity to the ligase Lig C (Namba & Makino, 2022). Lig C enzymes are minimal ligases that are composed of only the NTase and OB domains and have no auxiliary flanking domains. Lig C in mycobacteria have been implicated in a minor pathway of Ku-dependent NHEJ (Shuman & Glickman, 2007).

On either side of DV-Lig5, are two hypothetical proteins, with unknown functions. The hypothetical protein located to the left of DV-Lig5 belongs to a protein family of unknown function. An Interpro scan of this protein characterises this protein as a conserved hypothetical protein. Members of this family are widely distributed bacterial proteins, about 230 residues in length. All members have a motif RxxRDxRFxxx[DN]KxxY (Mulder & Apweiler, 2007). An Interpro scan of the hypothetical protein, to the right of DV-Lig5 was not able to generate any results on this protein. AlphaFold predicted models were generated for each hypothetical protein and these predicted models were used in a PDB search, to identify any similar protein homologs. The results of this search connected these hypothetical proteins to other hypothetical proteins from different microorganisms, with unknown functions. The sequences for these hypothetical proteins were also used in the Phyre^2 database (Kelley et al., 2015). From this search, the hypothetical protein to the left of DV-Lig 5 showed some similarity to a wide array of proteins, with different functions: oxidoreductase, hydrolase, and methyl transferase. The hypothetical protein to the right of DV-Lig5 showed very low similarity to any function defined proteins, with most of these proteins playing roles in cell adhesion and the immune system.

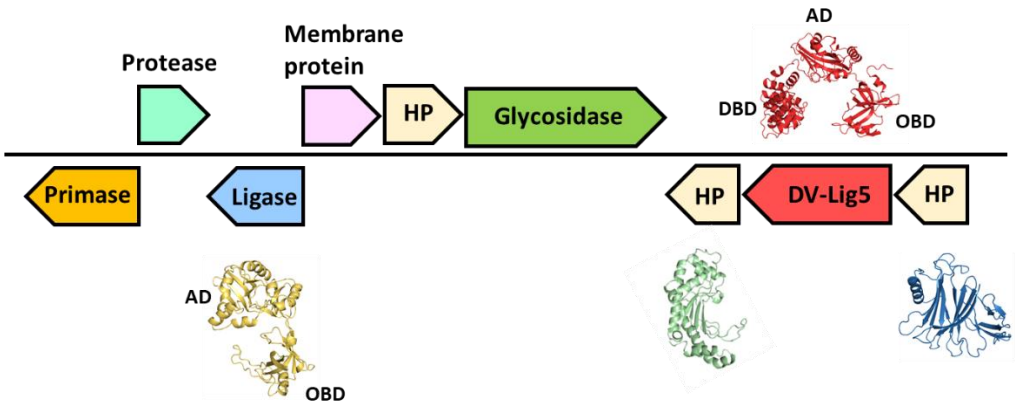


Figure 3.7. Location of DV-Lig5 within the DV-genome UQ272. Bacterial lineage: Bacteria; Actinobacteria. Other nucleic acid binding proteins are upstream from DV-Lig5. AlphaFold predicted models show structures for DV-Lig5, two hypothetical proteins and a minimal DNA ligase, with an AD and an OBD. Protein predicted models were generated by AlphaFold2, from Google Colab, version v2.3.1 (John Jumper, 2021; Varadi et al., 2022) and all structures were presented in PyMOL (Schrödinger, 2020).

AlphaFold2 was used to predict the 3D structures of DV-Lig5 and DV-Lig2 as described in **Section 2.1**. The predicted local distance difference test (pLDDT) shows that a majority of the structure, for both proteins, gives a high confidence score for each amino acid, however there are some regions with low confidence, particularly regions with high flexibility e.g., linkers between domains (**Appendix C.1**). Examination of the Predicted Aligned Error (PAE) plots, for both structures, shows low error within domains, and higher error in the relative positions of each domain (**Appendix C.1**). This reflects the flexibility and movement of the domains and is consistent with what we know about the structures of DNA ligases.

To get an overview of the placement of secondary structural elements and compare these with DNA ligases of known structure, topology maps were generated for the AlphaFold models. Consistent with the structural and sequence alignments, these indicate that secondary structural elements are positioned equivalently with known DNA ligase structures (**Figure 3.8**). Topology maps generated for DV-Lig5 and DV-Lig2 show that these proteins share a similar arrangement of secondary structure, although DV-Lig2 has three additional helices in the DNA binding domain, in comparison to DV-Lig5.

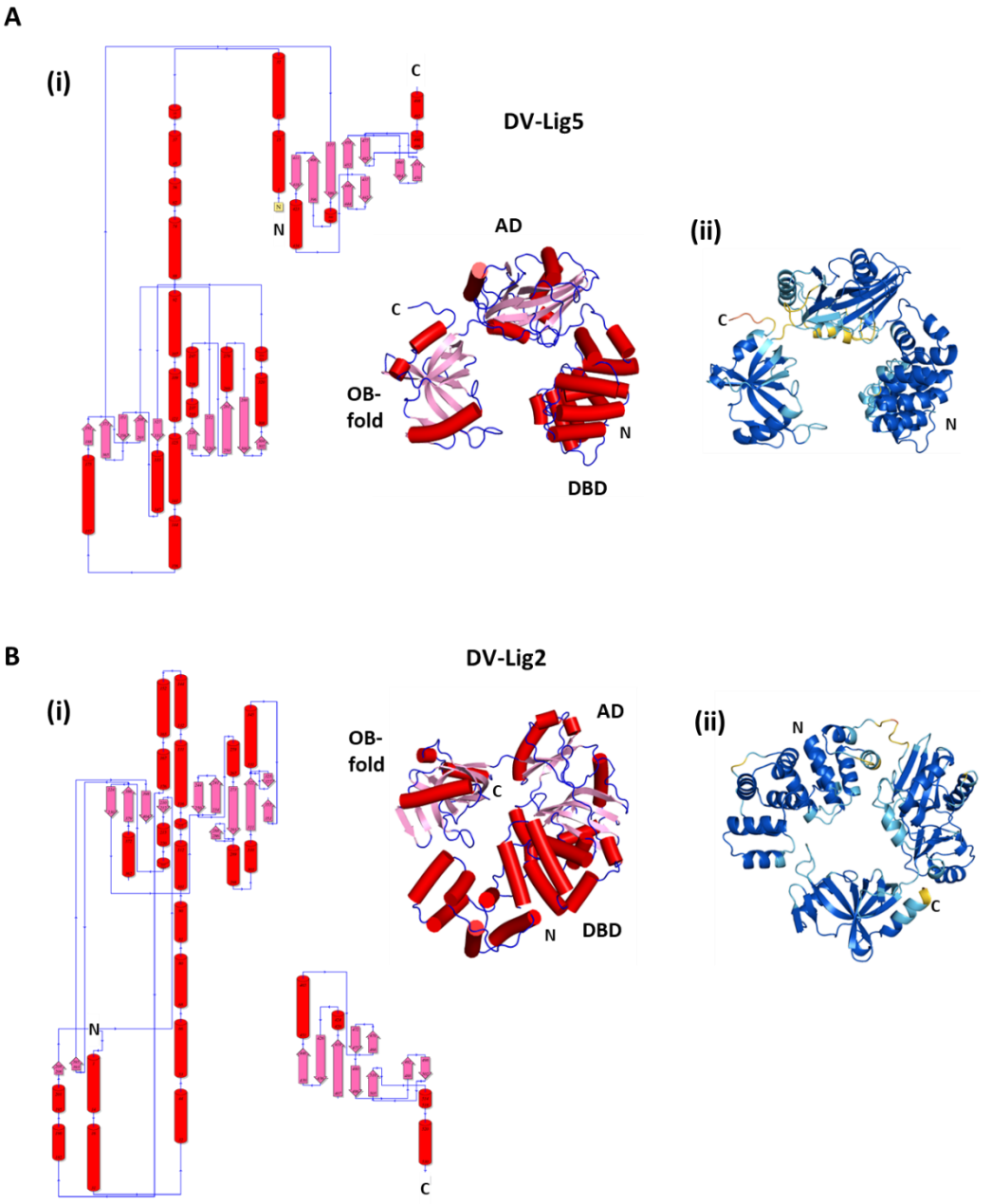
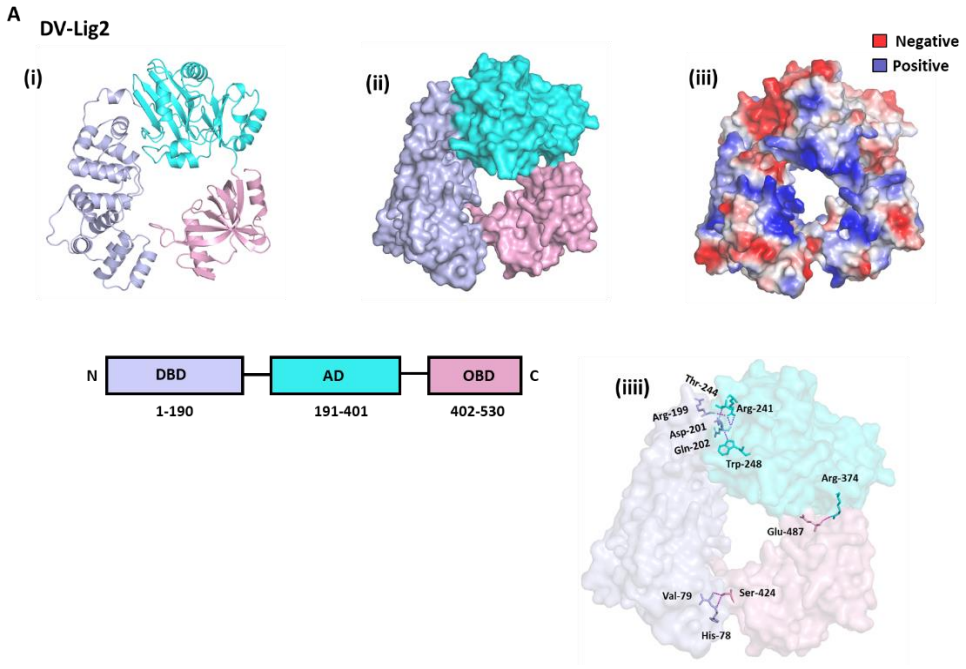


Figure 3.8. Structural arrangements of DV-Lig5 and DV-Lig2 proteins**. A, B** (i) Topology maps showing arrangement of secondary structural elements for DV-Lig5 and DV-Lig2 and AlphaFold predicted models, colour coded in reference to secondary structure, helices in red, strands in pink and loops in blue. **A, B** (ii) AlphaFold predicted structural models of DV-Lig5 and DV-Lig2. AlphaFold models are often shown with high confidence residues coloured blue, and lower confidence in yellow, orange and red. Protein topology maps were generated in PDBsum (Laskowski, 2022) using AlphaFold predicted structural models for DV-Lig5 and DV-Lig2 (John Jumper, 2021; Varadi et al., 2022).

Examination of the 3D structures of DV-Lig5 and DV-Lig2 and InterPro scan analysis of their sequences, confirms that both proteins are made up of three main domains that are common to Lig B DNA ligases; a DNA binding domain (DBD) (IPR036599), an adenylation domain (AD) (Cd07901) and an OB-fold domain (OBD) (IPR012340) (**Figure 3.9**, **A**, **B**). Protein contact potential

generated for both DV-Lig2 and DV-Lig5 shows a positive pocket within the central channel of each protein where the DNA is expected to bind.

Interdomain interactions of DV-Lig2 and DV-Lig5 were investigated in PyMOL (**Figure 3.9**, **iiii**). DV-Lig5 has remarkably few contacts between the three protein domains. The only interactions observed are between the DB doamin and the AD domain, where Asp-303 (AD) is interacting with the side chain from Leu-178 (DB) and Asp-306 (AD) and Arg-150 (DB) forms a salt bridge. This salt bridge interaction may be important in stabilising the closed confirmation around the DNA. DV-Lig2, on the other hand, contains many contacts between all three domains. Val-79 and His-78 (DB) form contacts with Ser-424 (OB). A salt bridge is formed between Glu-487 (OB) and Arg-374 (AD). Between the DB domain and the AD domain there are several contacts formed between residues: Arg-241 (AD) and Gln-202 (DB), Trp-248 (AD) and Asp-201 (AD), Asp-201 (DB) and Thr-244 (AD), Arg-199 (DB) and Arg-241 (AD).



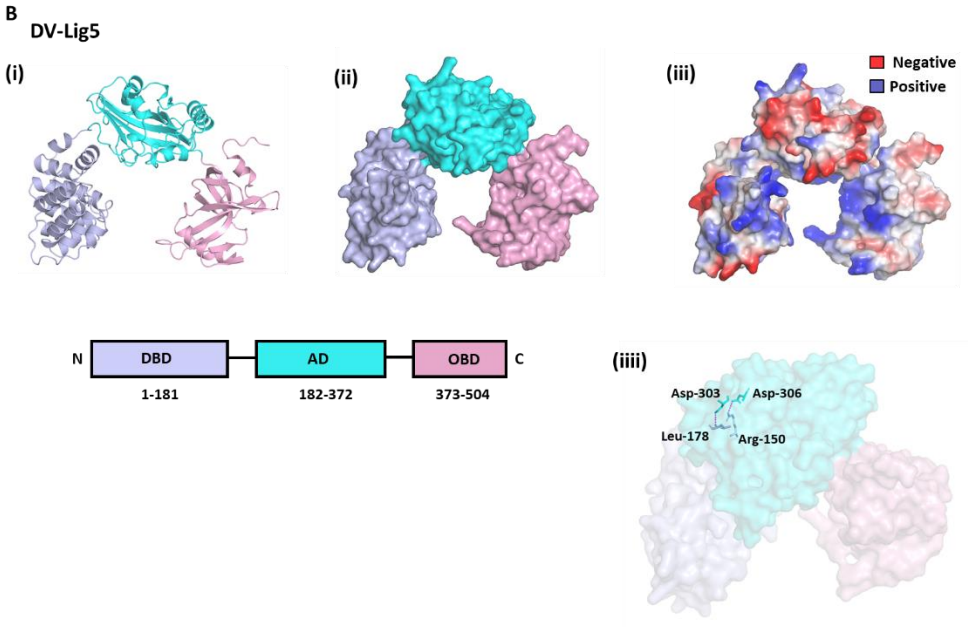


Figure 3.9. AlphaFold predicted models for DV-Lig2 and DV-Lig5. **A**, **B**) (i) Arrangement of DBD (purple), AD (blue) and OBD (pink), with domains coloured accordingly. (ii) Arrangement of domains shown as a surface. (iii) Protein contact potential, with positive contact potential coloured blue and negative contact potential coloured red. (iiii) Interdomain interactions, with interacting residues labelled on figure and coloured according to their domain. DV-Lig2 and DV-Lig5 predicted models were generated by AlphaFold2, from Google Colab, version v2.3.1 (John Jumper, 2021; Varadi et al., 2022) and all structures were presented in PyMOL (Schrödinger, 2020).

The DB, AD and OB domains from DV-Lig2 and DV-Lig5 as well as several other Lig B-type ligases were separately super imposed onto the crystal structure of human DNA ligase I bound to DNA (1X9N) (**Figure 3.10**, **A**). As expected, the AD and OB domains have high structural similarity between these proteins, however the DB domains differ in number of helices and their positioning; for example, DV-Lig2 contains additional helices, near the N-terminal portion of the domain relative to DV-Lig-5 (**Figure 3.10**, **C**).

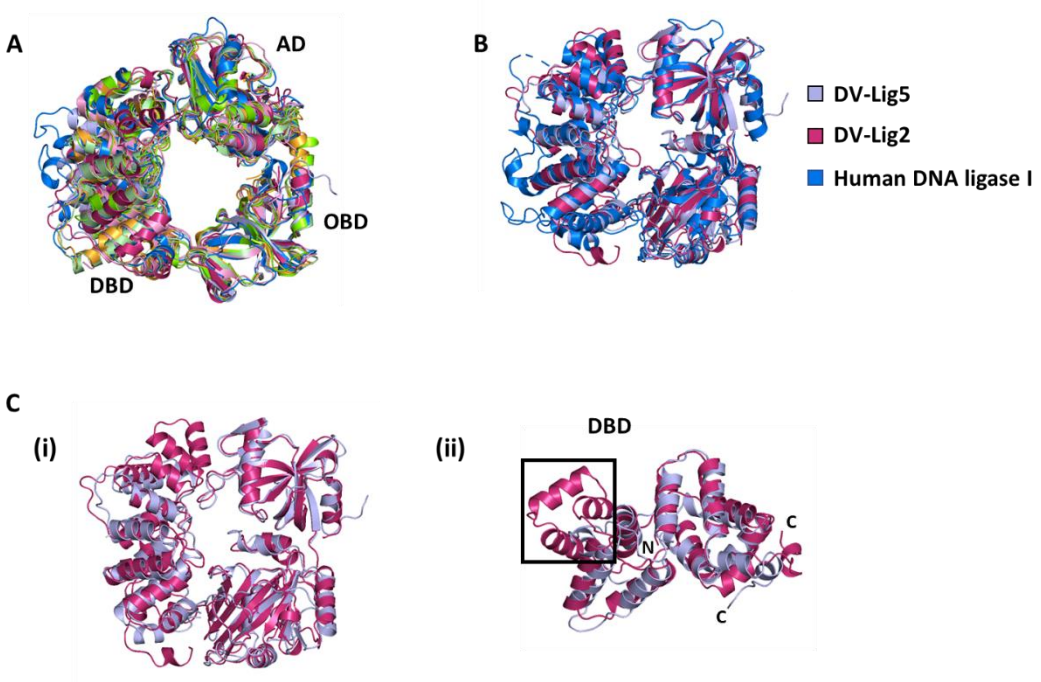
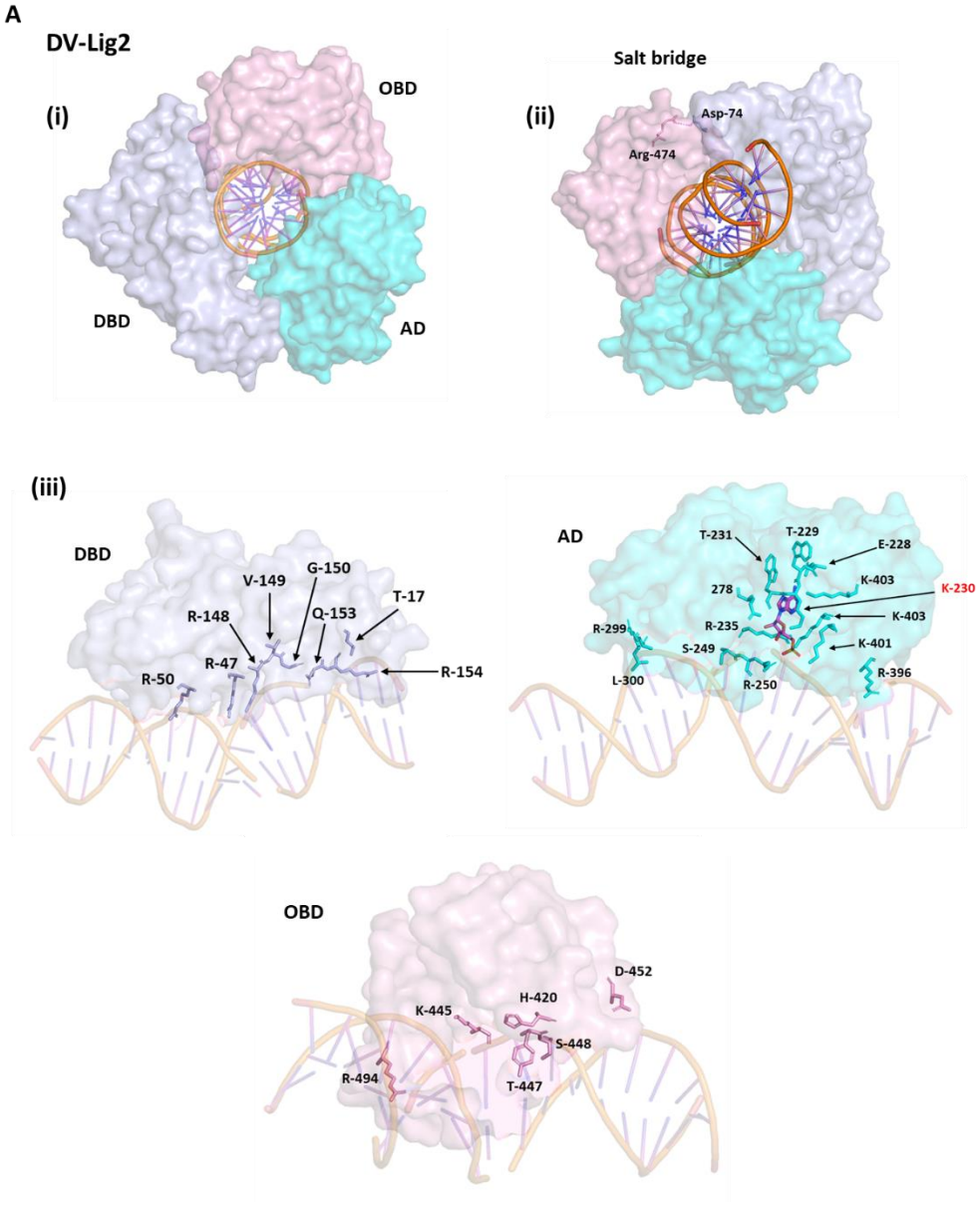


Figure 3.10. Structural alignments of DNA ligases with DV-Lig2 and DV-Lig5. **A**) Domains of DNA ligase crystal structures (*Archaeoglobus fulgidus, Pyrococcus furiosus, Thermococcus sibiricus* and *Saccharolobus solfataricus*) and predicted structures of DV-Lig2 and DV-Lig5 super imposed onto human DNA ligase I (1X9N). **B**) DV-Lig5 and DV-Lig2 domains super imposed onto human DNA ligase I. RMSD values were generated separately for each domain (DV-Lig2: DB, 3.143, AD, 1.171, OB, 1.257. DV-Lig5: DB, 1.982, AD, 1.16, OB, 1.016). **C**) (i) Domains from DV-Lig2 super imposed onto domains of DV-Lig5. RMSD values were generated for each domain (DB, 2.363, AD, 1.169, OB, 1.068). (ii) Analysis of DBDs from DV-Lig2 and DV-Lig5 super imposed. Additional helices in DV-Lig2 are indicated withing back box. DV-Lig2 and DV-Lig5 predicted models were generated by AlphaFold2, from Google Colab, version v2.3.1 (John Jumper, 2021; Varadi et al., 2022) and all structures were presented in PyMOL (Schrödinger, 2020).

To identify what potential residues were making polar contacts with the DNA, the predicted DV-Lig2 and DV-Lig5 structures were super imposed onto the crystal structure of human DNA ligase I bound to 5' adenylated nicked DNA. In both DV-Lig2 and DV-Lig5 models, all three domains are predicted to contact the DNA strands via positive and polar uncharged residues.

In DV-Lig2 and DV-Lig5 most residues are forming polar contacts with the phosphate backbone of the DNA duplex, with a select few forming polar contacts with the ribose sugar unit, or the nucleobase from different nucleotides of the DNA, or the AMP group. In DV-Lig2, from the OB domain, Arg-494 is forming polar contacts with a single nucleobase, Tyr-447 is forming polar contacts with 3 nucleobase groups, from different nucleotides and Ser-448 is forming a polar contact with a ribose sugar group. From the DB domain, Arg-50 is interacting with the phosphate backbone and ribose sugar of a single nucleotide

and within the AD domain, Arg-299 is forming polar contacts with a ribose sugar on a single nucleotide (**Figure 3.11**, **A**) In DV-Lig5, from the OB domain, Arg-477 is forming polar contacts with a ribose sugar and the phosphate backbone from different nucleotides, Lys-420 is forming polar contacts with a nucleobase and a ribose sugar, from different nucleotides, Arg-392 is forming polar contacts with a nucleobase and the phosphate backbone from different nucleotides, and Gln-462 is forming polar contacts with the ribose sugar and phosphate backbone of the same nucleotide. From the DB domain, Arg-52 is forming polar contacts with a nucleobase, and two ribose sugar groups from different nucleotides and within the AD domain, Gln-271 is forming a polar contact with the ribose sugar unit of a single nucleotide (**Figure 3.11**, **B**) These residues are conserved among many DNA ligases, including human DNA ligase I. Residues from the OB and AD domains make the most contact with DNA. Both catalytic lysines from DV-Lig5 (K-208) and DV- Lig2 (K-230) have polar contacts, with the bound AMP group, as excepted. In its predicted DNA-bound conformation, DV-Lig2 forms a potential salt bridge, between Arg-478 in the DB domain and Asp-74 in the OB domain. In DV-Lig5, there are no predicted contacts between the DB domain and the OB domain, and the structure does not fully encircle the DNA duplex.


A
DV-Lig2
(i)
OBD
DBD
AD
Salt bridge
(ii)
Asp-74
Arg-474
(iii)
DBD
G-150
V-149
T-17
R-148
Q-153
R-47
R-50
R-154
AD
T-231
T-229
E-228
K-403
K-230
278
K-403
R-235
R-299
S-249
K-401
L-300
R-250
R-396
OBD
D-452
H-420
K-445
S-448
R-494
T-447

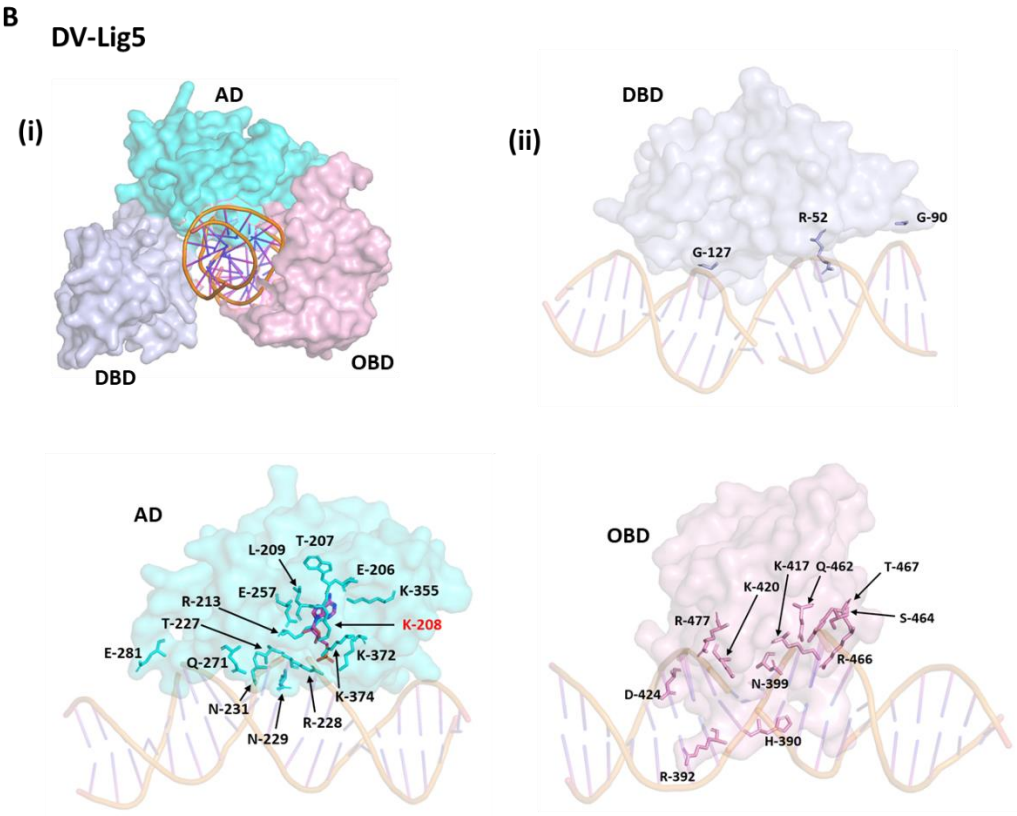


Figure 3.11. AlphaFold predicted models of DV-Lig2 and DV-Lig5 superimposed onto h-LigI bound to nicked DNA duplex (IX9N). **A**, **B** (i) Domains of DV-Lig2 and DV-Lig5 shown as a surface around a DNA duplex. **A** (ii) DV-Lig2 forms a salt bridge between Arg-474, from the OB domain and Asp-74, from the DB domain. **A** (iii), **B** (ii) Domains from DV-Lig2 and DV-Lig5 form polar contacts with nicked DNA, with the catalytic lysine residues, from the AD domains, indicated on the figure by red text. DV-Lig2 and DV-Lig5 predicted models were generated by AlphaFold2, from Google Colab, version v2.3.1 (John Jumper, 2021; Varadi et al., 2022) and all structures were presented in PyMOL (Schrödinger, 2020).

ATP-dependent DNA ligases have limited areas of sequence homology, with the KxDGxR motif (shown as the first motif in **Appendix C.2**.) being the most conserved. This motif contains the active site lysine for various nucleotidyl transfer enzymes, including DNA and RNA ligases. When sequence alignments are restricted to more specific sets, greater homology can be observed (Doherty et al., 1996; Tomkinson et al., 1991). Sequence alignments of DV-Lig5, and DV-Lig2, against homologous polypeptides, (**Appendix C.2**), shows many sequence similarities between the proteins, which all contain the N-terminal DNA binding domain, an adenylation domain and the C-terminal OB-fold domain. Location of these three domains is based off the polypeptide sequence from DV-Lig5. The catalytic domain is defined by a set of six peptide motifs, indicated within black boxes in (**Appendix C.2)** that comprise the ATP binding pocket (Gong et al., 2004). The most highly conserved sequences are located primarily in the

adenylation region of the protein, with the majority of inserts occurring in the extremities of the N and C terminals.

3.2.2 Recombinant production of DV-Lig5

3.2.2.1 Small scale expression testing

The gene for DV-Lig5 was cloned into pDEST17 and pHMGWA expression plasmids and transformed into BL21 (DE3) pLysS, Arctic express and Origami (DE3) *E. coli* expression strains, as described in **Section 2.2.1**. The best protein expression was seen in the Origami strain (**Figure 3.12**). There was a lot of protein expressed from the pDEST17 plasmids, however this was in the insoluble fraction. Protein was also expressed in the pHMGWA plasmids, with 15 ºC giving the best soluble protein expression.

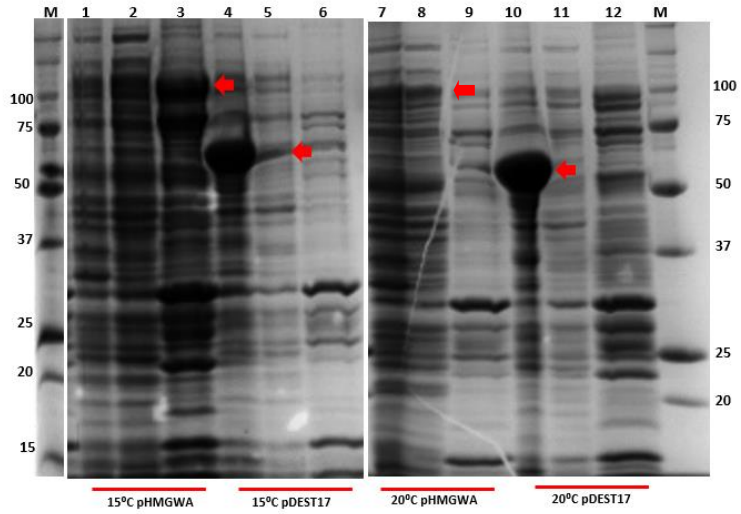


Figure 3.12. SDS PAGE of small-scale protein expression results for DV-Lig5 expressed in Origami (DE3) *E. coli*. Lanes 1, 4, 7 and 10 represent insoluble protein, lanes 2, 5, 8 and 11 represent soluble protein and lanes 3, 6, 9 and 12 represent soluble protein bound to Ni beads. Red arrows indicate expression of of DV-Lig5 protein, at the expected size for MBP tagged protein (99.7 kDa) and His-tagged protein (59.3 kDa). A precision plus protein ladder was used as a molecular weight marker (M). Gels were stained with Coomassie Blue stain, using the quick stain method (Wong et al., 2000) and protein bands were visualised and captured on the iBright™ CL750 Imaging System, Invitrogen™.

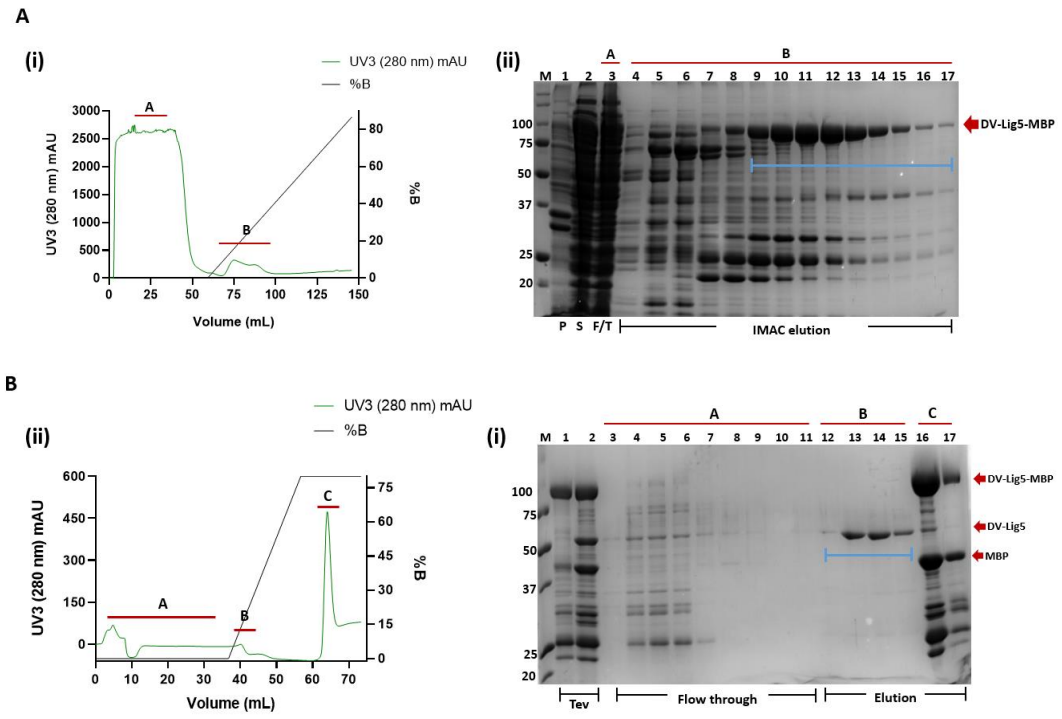
3.2.2.2 Large scale purifications

Expression of MBP-tagged DV-Lig5 protein, in *E. coli* (DE3) Origami was scaled up following methods from **Section 2.3**. A three-step purification via IMAC, reverse IMAC and gel filtration chromatography (**Section 2.4**), produced

soluble, active protein, suitable for characterisation experiments. The chromatograms and corresponding SDS-PAGE gels depict the purification, column load and flow through fractions.

The first IMAC purification gave fractions of soluble MBP-tagged DV-Lig5, that eluted off the column with the addition of 20 % buffer B, with the addition of several contaminating *E. coli* proteins (**Figure 3.13**, **A**).

An overnight incubation with TEV protease and a subsequent reverse IMAC purification, resulted in a successful removal of the MBP tag from DV-Lig5 protein, with the protein eluting off the IMAC column with the addition of 10 % buffer B. Fractions containing DV-Lig5 protein were highly pure (**Figure 3.13**, **B**) and a subsequent gel filtration purification resulted in a single peak that contained sufficient pure DV-Lig5 protein for further characterization (**Figure 3.13**, **B**).



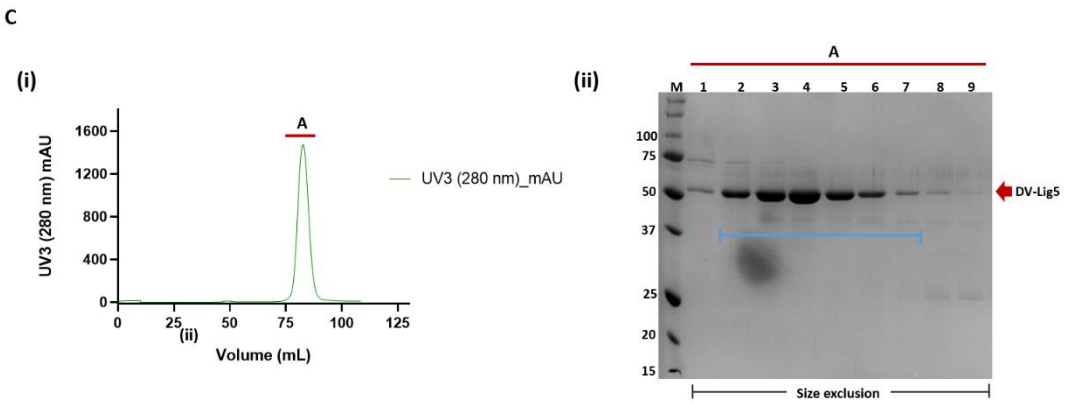


Figure 3.13. IMAC and gel filtration chromatograms (i) and SDS PAGE gels for production of DV-Lig5MBP protein from *E. coli* (DE3) Origami (ii). **A**) IMAC purification of DV-Lig5$_{MBP}$. (i) Peak A represents flow through during IMAC purification, peak B represents fractions of proteins that eluted during the elution step of the IMAC purification, including DV-Lig5-$_{MBP}$ protein (99.7 kDa). (ii) Lanes 1-3 represent insoluble (P), soluble (S) and flowthrough (F/T) samples. Lanes 4-17 represent fractions containing proteins that eluted off the column during the elution step, with the addition of buffer B. The blue bar indicates fractions that were pooled and incubated overnight with TEV protease, followed by a reverse IMAC purification. **B**) Reverse IMAC purification of DV-Lig5. (i) Peak A represents flow through during reverse IMAC purification, peak B represents fractions that contain de-tagged DV-Lig5 protein. Peak C represents proteins that eluted during the elution step of the Reverse IMAC purification. (ii) Lanes 1-2 are pooled IMAC fractions before the addition of TEV (1) and fractions after an overnight incubation with TEV (2), Lanes 3-11 are fractions from the flowthrough during the reverse IMAC purification. Lanes 12-15 are fractions that contain de-tagged DV-Lig5 protein, that eluted off the column with the addition of 10 % buffer B. Lanes 16-17 contain protein fractions eluted during the imidazole gradient step, with the addition of 60% buffer B. The blue bar indicates fractions that were pooled, up concentrated, and further purified by gel filtration. **C**) Gel filtration purification of DV-Lig5. (i) Peak A represents where DV-Lig5 protein eluted off the gel filtration column. (ii) Lanes 1-9 represent the following proteins present in fractions from peak A (i). The blue bar indicates fractions that were pooled and up concentrated and stored for further use. A precision plus protein ladder was used as a molecular weight marker (M). Gels were stained with Coomassie Blue stain, using the quick stain method (Wong et al., 2000) and protein bands were visualised and captured on the iBright™ CL750 Imaging System, Invitrogen™. Chromatogram graphs were designed in GraphPad Prism, version 9.0.0.

3.2.3 Protein folding and stability of DV-Lig5

The folded structure of DV-Lig5 protein was analysed using circular dichroism (CD). Secondary structure predictions from CD spectra and PDBsum analysis of the DV-lig5 AlphaFold predicted model were compared (**Figure 3.14**) Here there are some discrepancies between helical contributions, with CD predictions giving over 50 % for helices, compared to AlphaFold giving 30 % for helices. It is likely that some of the helices are included in other contributions from the AlphaFold predictions. These results provide additional confidence in the accuracy of the AlphaFold predicted structure and confirm that the purified protein is folded.

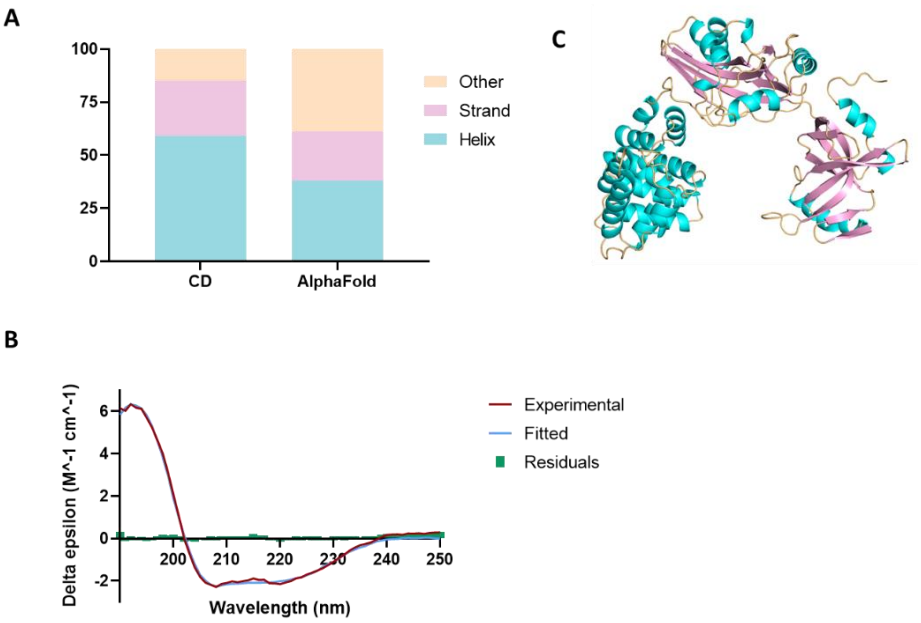


Figure 3.14. Circular dichroism (CD) and AlphaFold secondary structural composition of DV-Lig5 protein. **A**) quantification of secondary structural predictions from CD and AlphaFold prediction model, using PDBsum analysis (Laskowski, 2022). **B**) Single spectrum analysis of CD spectra, using BeStSel database (Micsonai et al., 2018). **C**) AlphaFold 3D structural prediction of DV-Lig5, coloured based on secondary structure (Helix in blue, strand in pink and other orange. (John Jumper, 2021). Graphs were designed using Prism version 8 (GraphPadSoftware). Wavelength range (190-250 nm) and scale factor (1). RMSD value (0.1474). NRMSD value (0.04824).

To determine the transition temperature of the loss of secondary structure, the CD signal was monitored at 222 nm as a function of temperature. **Figure 3.15, A**, shows the thermal melt curve generated for DV-Lig5 protein, from CD. Here the CD intensity drastically decreases from 65 to 75 ºC and the melting temperature (T_m) was calculated to be 63 ºC. Differential scanning fluorimetry (DSF) using SYPRO orange, as described in **Section 2.8**, was also used to analyze protein unfolding and determine the T_m of DV-Lig5 protein. **Figure 3.15**, **B**, shows the thermal melt curves generated for DV-Lig5 protein, from DSF, at different protein concentrations. Here the fluorescence signal increases from 35 ºC and drops off just before 60 ºC generating single peaks for each protein concentration. The average T_m, across all protein concentrations was calculated to be 45 ºC, which is much lower than the T_m calculated from CD thermal melt analysis. This discrepancy in T_m between the two different methods probably reflects that tertiary structure, which is monitored by DSF, is lost before secondary structure, and also suggests that the SYPRO dye is slightly denaturing.

Further DSF experiments were performed with DV-Lig5 protein, to investigate protein stability with the addition of metal ions, ATP and nick DNA substrate. The T_m from these thermal melts were calculated and T_m values were plotted against increasing metal ion or ATP concentrations. **Figure 3.15**, **C** shows results of thermal melts with varying concentrations of magnesium or manganese. The highest T_m was seen in samples containing 5 mM magnesium or manganese, while higher concentrations of metal ions in the samples lowered the T_m. Overall DV-Lig5 protein was more stable with the addition of magnesium compared to manganese, seen by higher T_m values across all concentrations. **Figure 3.15**, **D** shows results of thermal melts with varying concentrations of ATP cofactor. There was no obvious pattern of an optimum ATP concentration for protein stability. However, samples with no or high concentrations of ATP, resulted in lower Tm values. The DSF experiments, wherein nicked DNA substrate was introduced into protein-containing samples, yielded consistent melting temperatures (T_m) that showed minimal difference when compared to samples lacking DNA substrate. See **Appendix C.6**.

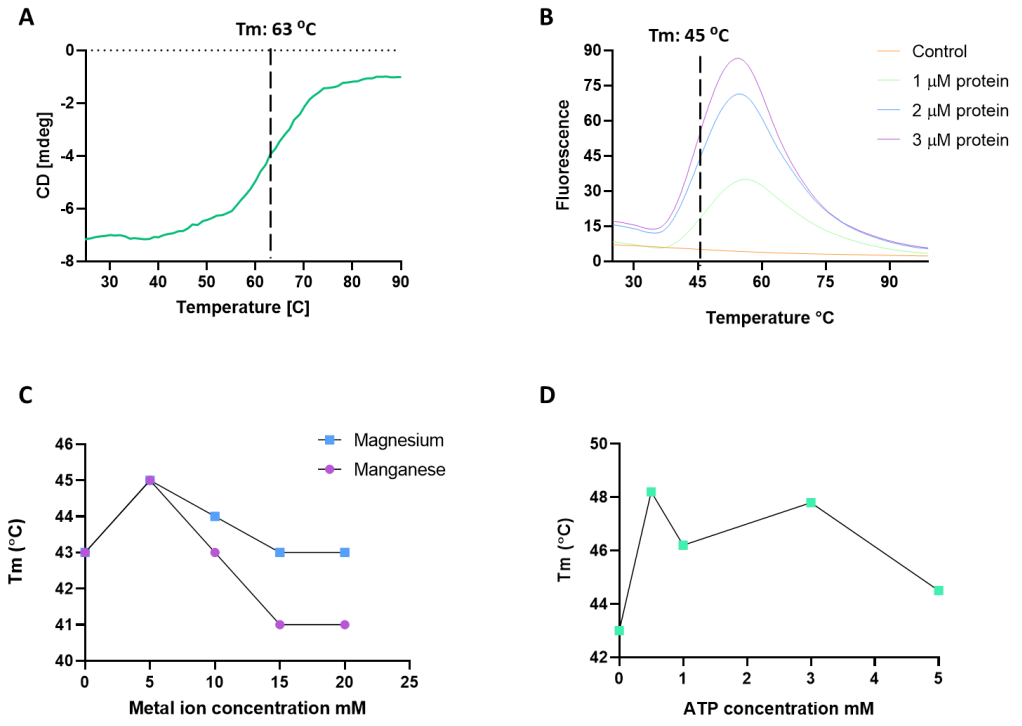
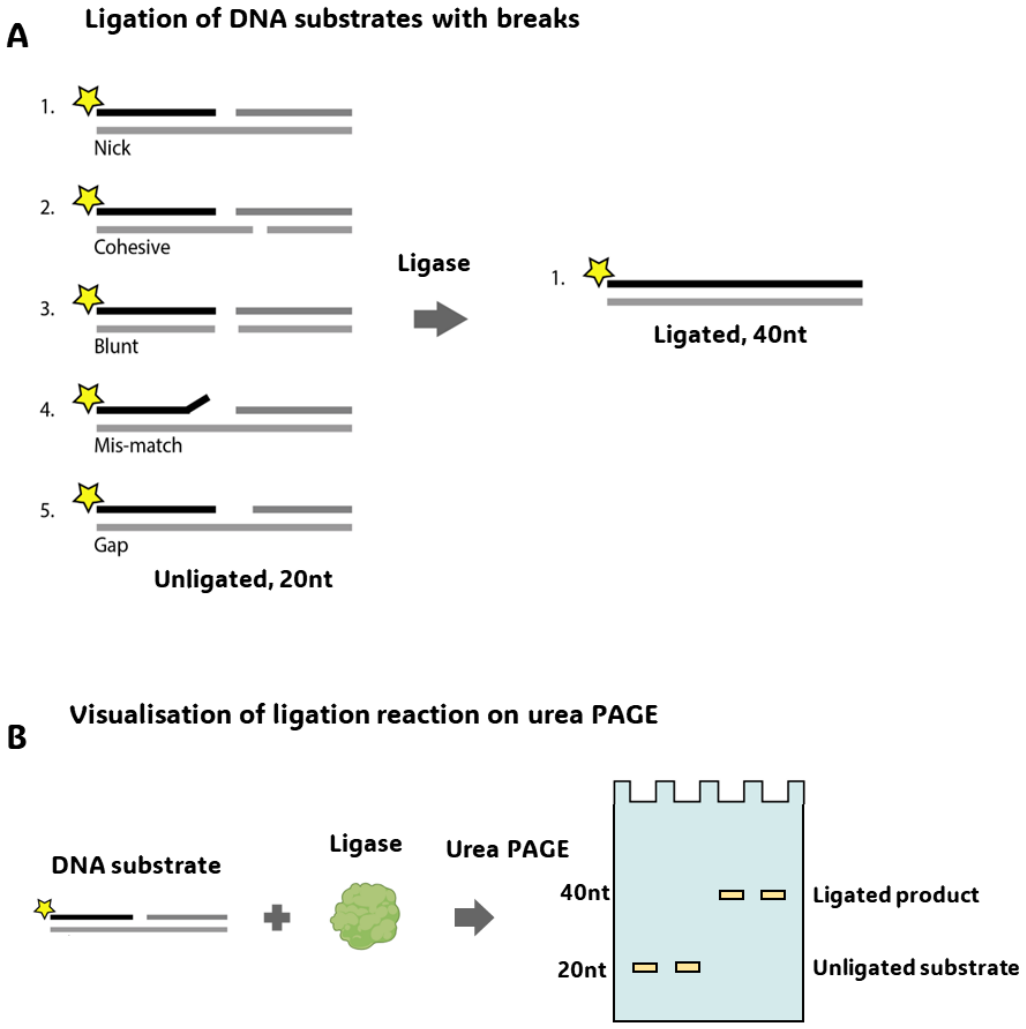


Figure 3.15. Results from thermal melts of DV-Lig5 protein, using CD and DSF. **A**) CD thermal melt data of DV-Lig5 protein, at 222 nm. T_m values were determined from the midpoint in the unfolding equilibrium and are indicated on the graph, by a dotted line. **B**) DSF, with SYPRO orange, showing melt curves of DV-Lig5 at three different protein concentrations (1, 2 & 3 µM). Reactions were carried out in replicates of three. T_m values were determined from the midpoint in the unfolding equilibrium and are indicated on the graph, by a

dotted line. **C**) First derivative T_m plots, from DSF, with SYPRO orange, with T_m values derived from the first derivative midpoint of each peak, at different metal ion concentrations (0-20 mM). **D**) First derivative T_m plots, from DSF, with SYPRO orange, with T_m values derived from the first derivative midpoint of each peak, at different ATP concentrations (0-5 mM). Protein was at a final concentration of 2 µM. Graphs were generated using GraphPad Prism version 8 (GraphPadSoftware).

3.2.4 Biochemical characterisation of DV-Lig5

The following section details the binding ability of DV-Lig5 to DNA substrates in electrophoretic mobility shift assays (EMSAs), and ligation activity on nicked DNA substrates, under different conditions. Additional activity assays show ligation ability on different DNA substrates, using gel-based activity assays.



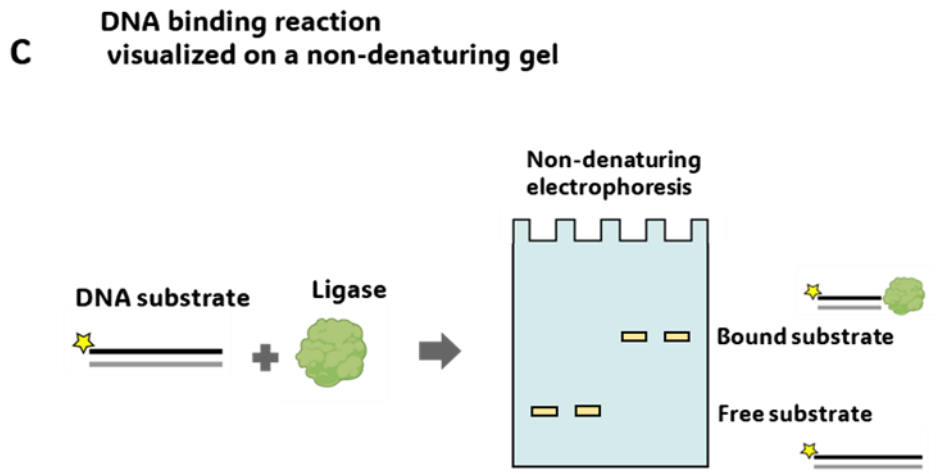


Figure 3.16. Schematic of enzyme assays for ligase activity and binding on DNA substrates. Stars represent labelling with the 6-carboxyfluorescein at the 5' terminus (5'FAM). Labelled strands are indicated by a black line while unlabelled portions of substrate duplexes are not visible during analysis are indicated by grey lines. **A**) Design of DNA substrates with different types of breaks and example of ligated nick DNA substrate by ligase enzyme, showing size difference of product after ligation (40nt). **B**) Analysis of assay products by urea PAGE indicating a size-shift based on ligation of product (yellow boxes). **C**) Analysis of DNA binding by ligase enzyme, visualised on non-denaturing TBE gels. DNA substrates bound to ligase protein run slower through the gel and are above substrates that are not bound by the ligase (yellow boxes).

3.2.4.1 DNA binding by DV-Lig5

The DNA binding ability of DV-Lig5 was tested with nicked DNA substrate, with results of binding visualised on a non-denaturing gel. Different concentrations of DV-Lig5 protein were used to determine the optimal concentration of protein required for binding.

Protein bound to DNA is indicated by an increase in band size, from the control band that doesn't contain any protein. Visualisation of the binding assay of DV-Lig5 and nicked DNA substrate, saw an increase in band size in reactions that contained DV-Lig5 protein. DV-Lig5 can bind to nicked DNA substrates from 5.5 µM down to 1.5 µM, with the best binding affinity observed from the highest concentration of protein (**Figure 3.17**).

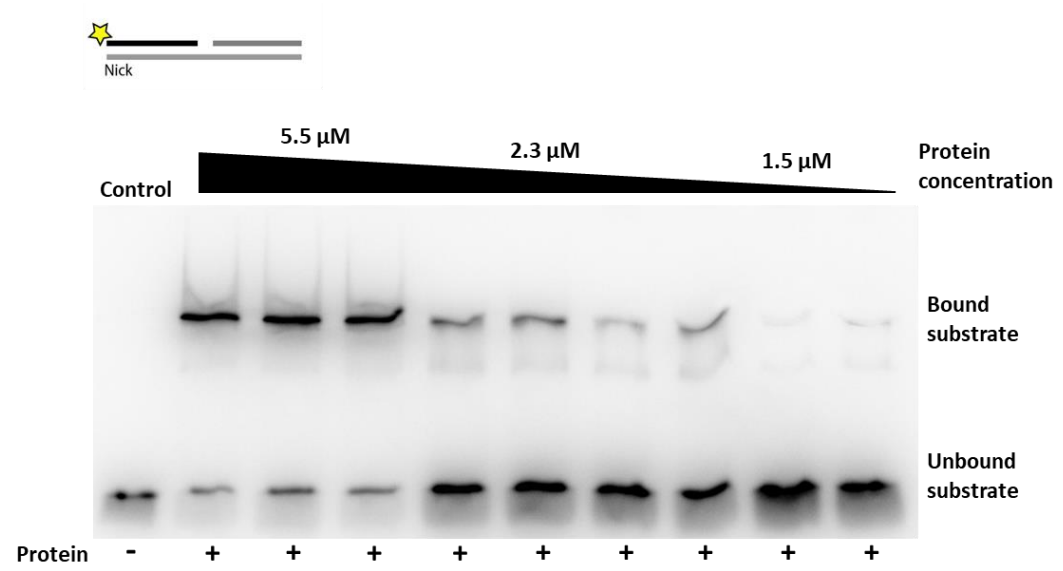


Figure 3.17. Electrophoretic mobility shift assay (EMSA) showing the binding ability of DV-Lig5 protein, to nicked DNA substrate, at different protein concentrations. DV-Lig5 protein was incubated with nicked DNA substrate, for 30 minutes, at 25^{o}, with 1 mM final ATP concentration and 40 mM EDTA. Three different protein concentrations were used (5.5 µM, 2.3 µM & 1.5 µM). Reactions were run out on a 10 % native TBE gel, with native loading dye in the reactions. Samples containing protein, were run out in replicates of three, for the different protein concentrations. The control lanes contain nicked DNA substrate with 1 mM, final ATP, EDTA, but no protein. Nicked DNA substrates are fluorescently labeled and were visualized on the iBright™ CL750 Imaging System, Invitrogen™.

3.2.4.2 Protein concentration optimisation for DV-Lig5 assays

A protein concentration gradient was carried out to determine the best concentration to work with in future activity assays. The results below (**Figure 3.18**) show that DV-Lig5 protein can ligate nicked DNA down to a final protein concentration of 0.1 µM. The best ligation was observed with a protein final concentration of 2 µM, giving close to 80 % ligation of substrate. Further concentration assays were carried out with higher protein concentrations. Higher DV-Lig5 protein concentrations also increases contaminating *E. coli* nucleases, which started to degrade the nicked substrate, skewing the final ligation results (Data not shown), therefore, a final protein concentration of 2 µM, was chosen for future activity assays.

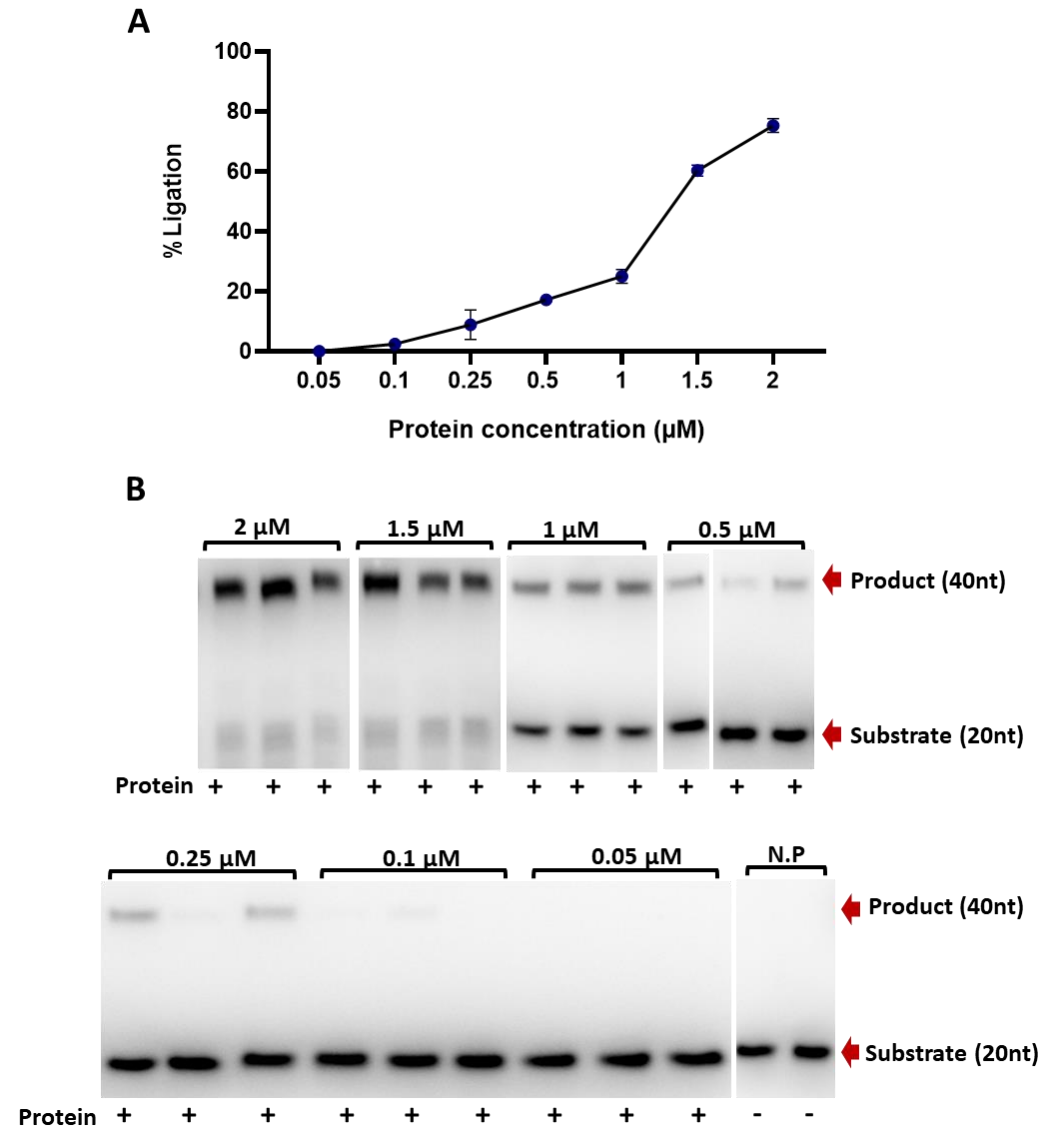
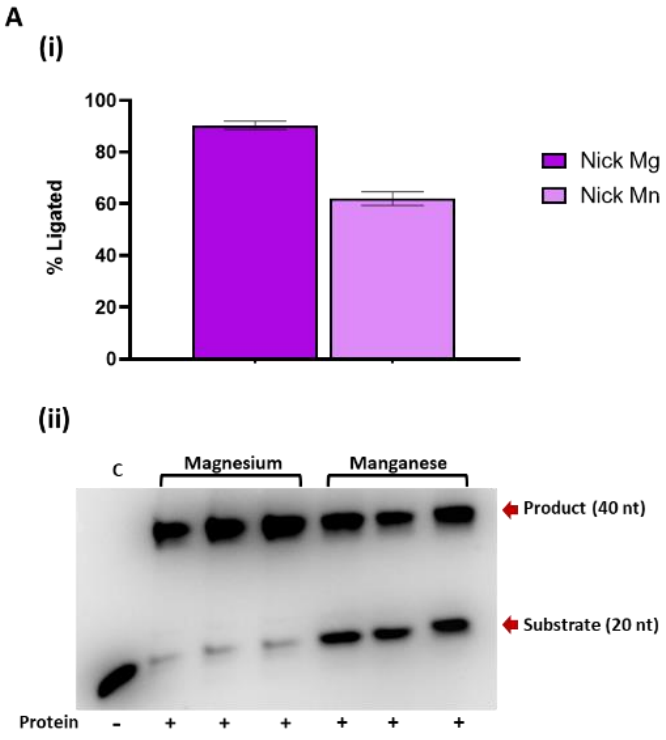


Figure 3.18. Shows ligation of nicked DNA substrate at different concentrations of DV-Lig5 protein. **A**) quantification of ligation by DV-Lig5 on nicked DNA, with different protein concentrations. Plots on the graph represent averages of each concentration. Standard deviation error bars are included. **B**) TBE urea gel showing results of DV-Lig5 protein concentration gradient. Addition of protein to the reaction is indicated by a plus symbol (+). Controls reactions are indicated by (N.P) and don't contain protein (-). Product (40 nt) and substrate (20 nt) are indicated by red arrows. Activity against each substrate was carried out in replicates of three. Reactions were carried out for 2 hours, at 25 ^{o}C, with varying final protein concentrations (2, 1.5, 1, 0.5, 0.25, 0.1, & 0.05 µM), 1 mM final ATP concentration and 10 mM final concentration of magnesium ions. Results of activity assays were visualized using iBright™ CL750 Imaging System, Invitrogen™. The graph was generated using GraphPad Prism version 8 (GraphPadSoftware).

3.2.4.3 Metal ion preference of DV-Lig5

The ligation ability of DV-Lig5 on nicked DNA substrate with different metal ions, was tested with activity assays, with resulting reactions visualized on TBE urea PAGEs and quantitative graphs. The following **Figure 3.19**, **A** shows comparative results of ligation on nicked DNA with magnesium (Mg) or

manganese (Mn). The addition of magnesium to the reaction, resulted in more ligated product (90 %) compared to reactions with manganese (60 %). In **Figure 3.19**, **B**, different concentrations of magnesium were added to the reaction, to determine the optimal metal ion concentration for ligation of nicked DNA substrate. Results of this experiment show that ligation of nicked DNA only occurs with the addition of a metal ion and can be inhibited with the addition of EDTA, to reactions containing metal ions. Ligation improves as the concentration of magnesium ion increases, up to 30 mM, where the ligation starts to plateau. Therefore, for further activity assays a final magnesium concentration of 10 mM would be used. A ligation assay was carried out with manganese at varying concentrations, however a high level of nuclease contamination was observed with higher manganese concentrations, making it difficult to quantify results (data not shown).



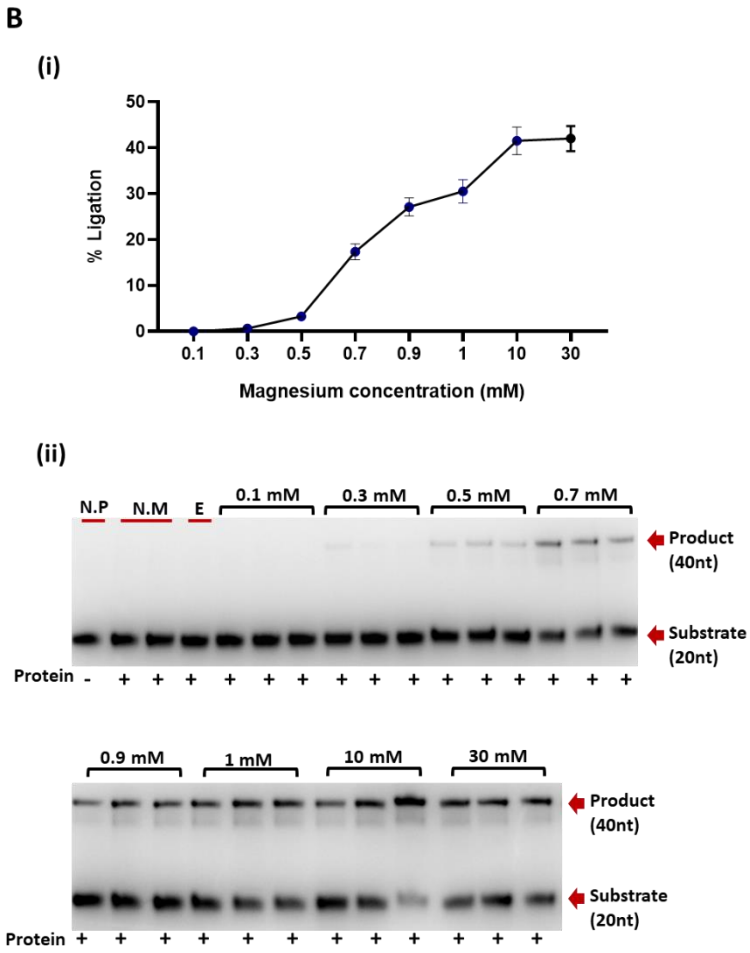


Figure 3.19. Ligation of nicked DNA substrate, by DV-Lig5 protein, with magnesium (Mg) or magnesium. **A**) (i) a graph representing the quantitative summary of ligation by DV-Lig5, with either magnesium (Mg) or manganese (Mn) as a cofactor. Bar graphs represent average ligation percentage, for reactions with Mg or Mn. Standard deviation error bars are included. (ii) TBE urea PAGE showing results of ligation with Mg or Mn, in replicates of 3. Addition of protein to the reaction is indicated by a plus symbol (+). Control reactions are indicated by (C), that don't contain protein (-). Product (40 nt) and substrate (20 nt) are indicated by red arrows. Reactions were carried out for 3 hours, at 25 ^{o}C, with 1 mM final ATP concentration and 10 mM final metal ion concentrations. **B**) (i) quantification of ligation by DV-Lig5, with varying final magnesium ion concentrations. Points on the graph represent the average ligation percentage for each magnesium concentration. Standard deviation error bars are included. (ii) TBE urea PAGE showing results of ligation with varying Mg ion concentrations, in replicates of three. Addition of protein to the reaction is indicated by a plus symbol (+). Control reactions all contain nicked DNA and vary by no protein (- N.P), protein with no metal ion (N.M) and protein, with 30 mM Mg and 40 mM EDTA (E). Product and substrate are indicated as described above. Reactions were carried out for 3 hours, at 25 ^{o}C, with 2 µM final protein concentration, 1 mM final ATP concentration and varying Mg ion final concentrations (0.1, 0.3, 0.5, 0.7, 0.9, 1, 10 & 30 mM). Results of activity assays were visualized using iBright™ CL750 Imaging System, Invitrogen™. Graphs were generated using GraphPad Prism version 8 (GraphPadSoftware).

3.2.4.4 Nucleotide cofactor specify of DV-Lig5

Gel based activity assays were used to test the ligation ability of DV-Lig5 with several nucleotide co-factors (ATP, ADP, GTP and NAD), to determine what co-factors are required for ligation of DNA breaks. The resulting reactions were visualised on TBE urea PAGE gels.

Analysis of results show DV-Lig5 can ligate nicked DNA substrate, without the addition of one of these cofactors, which is likely due to the protein being pre-adenylated during expression in *E. coli*. Attempts were made to deplete this covalently bound cofactor, as described in **Section 2.9**, however as complete removal was unsuccessful this background level of ligation was considered when evaluating further ligation with the additional cofactors. The percentage of ligation increased with the addition with the addition of ATP and ADP, with both cofactors giving similar increased percentages of ligation. There was some additional ligation with the addition of GTP, however these results were somewhat inconclusive, due to the presence of background ligation. This suggests that DV-Lig5 is indeed an ATP dependent ligase, with the ability to use ADP or GTP as back up cofactors, when required. The addition of NAD to the reaction showed a lower ligation percentage, compared even to the reactions with no cofactor and it is likely that the addition of NAD had an inhibitory effect on ligation of nicked substrate (**Figure 3.20**).

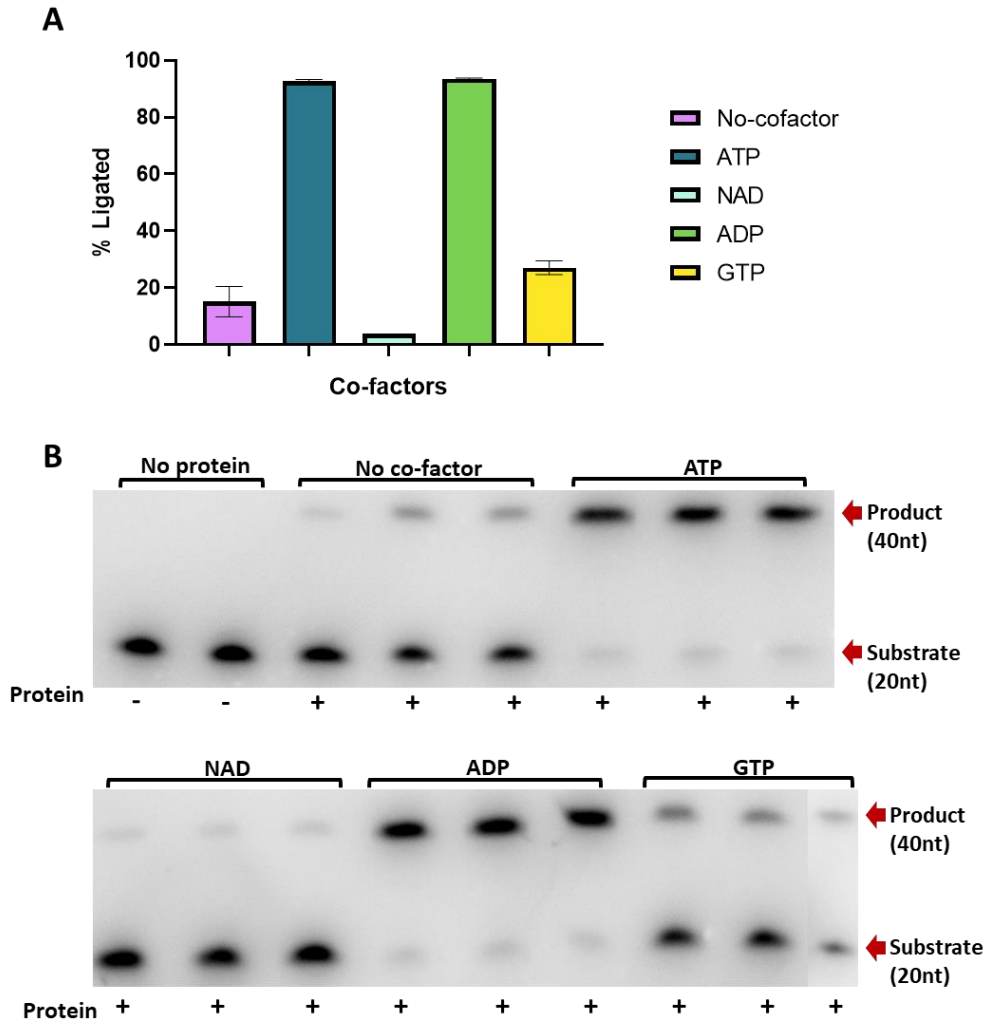


Figure 3.20. Ligation of nicked DNA substrate, by DV-Lig5 protein, with different cofactors. **A**) quantification of ligation by DV-Lig5 on nicked DNA, with different cofactors (ATP, ADP, GTP & NAD). Points on the graph represent averages of each concentration. Standard deviation error bars are included. **B**) TBE urea PAGE showing results of ligation by DV-Lig5, with and without the addition of different cofactors. Addition of protein to the reaction is indicated by a plus symbol (+). Controls reactions were used that don't contain protein (-) (No protein) or don't contain cofactor (No cofactor). Product (40 nt) and substrate (20 nt) are indicated by red arrows. Activity against each substrate was carried out in replicates of three. Reactions were pre-incubated for 2 hours at 25 ^{o}C with unlabeled nicked DNA substrate, 2 µM protein and 5 mM magnesium ion, followed by another 2-hour incubation with the addition of labelled nicked DNA substrate, 5 mM magnesium, and different cofactors at 1 mM final concentration. Results of activity assays were visualized using iBright™ CL750 Imaging System, Invitrogen™. The graph was generated using GraphPad Prism version 8 (GraphPadSoftware).

3.2.4.5 Temperature dependence of DV-Lig5

To determine the optimal temperature for DV-Lig5 protein activity assays, a temperature gradient assay from -40 ºC to 50 ºC was performed. These extreme low temperatures (-40 ºC and -20 ºC) were included as the protein comes from an organism that experiences such low temperatures in the Antarctic Dry Valleys. DV-Lig5 can ligate nicked DNA substrate at very low temperatures, however the

best ligation is seen at temperatures above 10 ºC. Analysis of temperatures above 30 ºC, showed a high level of DNA substrate degradation, most likely from contaminating *E. coli* proteins in the sample and were excluded from this figure (**Figure 3.21**).

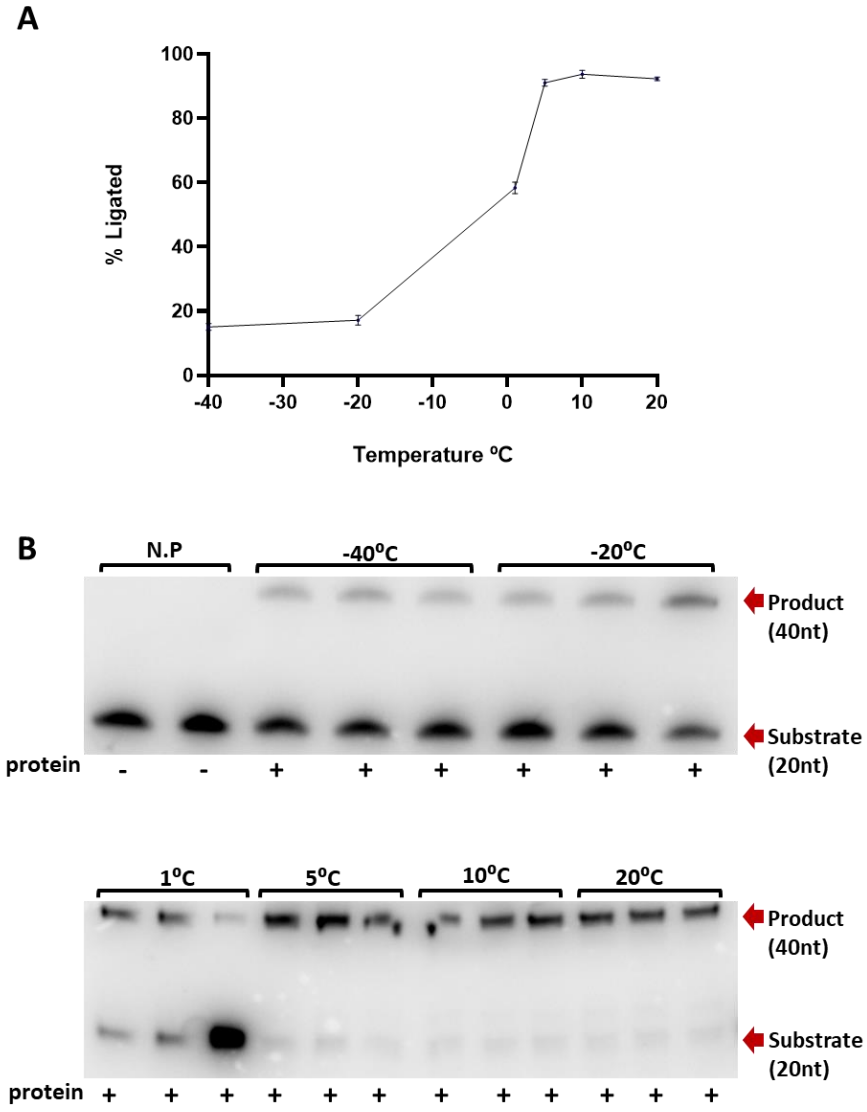


Figure 3.21. Ligation of nicked DNA substrate, by DV-Lig5 protein, at varying temperatures. **A**) a graph representing the quantitative summary of ligation by DV-Lig5 on nicked DNA, with different reaction temperatures (-40, -20, 1, 5, 10, 20 ^{o}C). Plots on the graph represent averages of each reaction temperature. **B**) TBE urea PAGE showing results of ligation by DV-Lig5, at different temperatures. Addition of protein to the reaction is indicated by a plus symbol (+). Controls reactions (N.P) don't contain any protein (-). Product (40 nt) and substrate (20 nt) are indicated by red arrows. Activity against each reaction temperature was carried out in replicates of three. Reactions were carried out for 5 hours, at varying temperatures, with 2 µM final protein concentration, 1 mM final ATP concentration and 10 mM final magnesium ion concentration. Results of activity assays were visualized using iBright™ CL750 Imaging System, Invitrogen™. The graph was generated using GraphPad Prism version 8 (GraphPadSoftware).

3.2.4.6 DNA substrate specificity of DV-Lig5

Ligation ability of DV-lig5 protein was tested against a range of different DNA substrates (nick, cohesive, blunt, mismatch and gapped). DV-Lig5 protein can ligate nicked and mismatch DNA substrates, with a better rate of ligation occurring with nicked DNA. No activity was observed with cohesive, blunt or gapped DNA substrates (**Figure 3.22**).

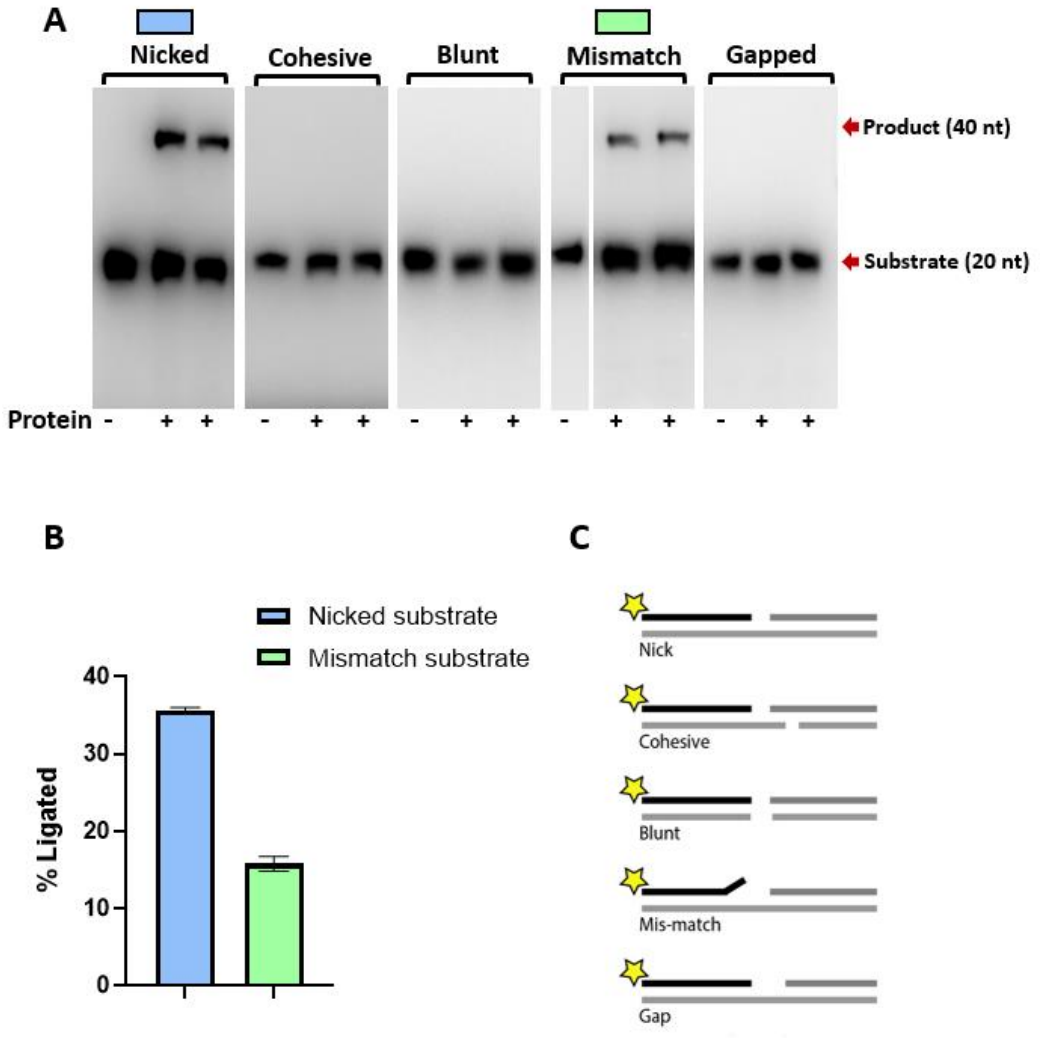


Figure 3.22. Results of ligation on different DNA substrates. **A**) TBE urea PAGE showing results of ligation, by DV-Lig5 on 5 different DNA substrates. Substrate (20 nt) and product (40 nt) are indicated by red arrows. Addition of protein to reaction is indicated by a plus symbol (+), controls (-) don't contain any protein. Reactions were run in replicates of 2. **B**) quantification of ligation by DV-Lig5 on nicked and mismatch DNA substrates. Standard deviation error bars are included. **C**) schematic of DNA substrates used in reactions. Star indicates fluorescent label. Reactions were carried out for 8 hours, at 20^{o}C, with 2 µM final protein concentration, 1 mM final concentration of ATP and 10 mM final concentration of magnesium. Results of activity assays were visualized using iBright™ CL750 Imaging System, Invitrogen™. The graph was generated using GraphPad Prism version 8 (GraphPadSoftware).

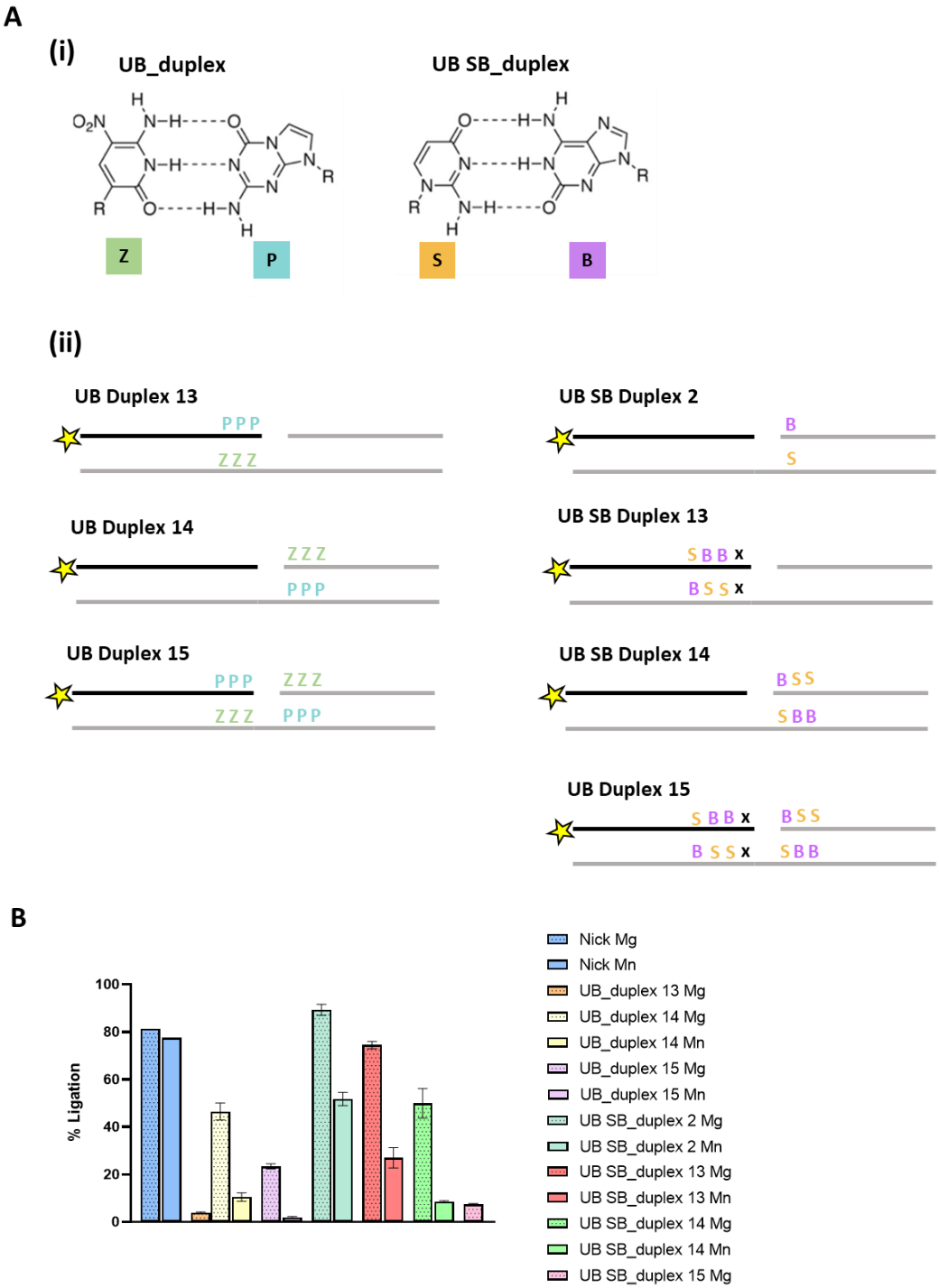
3.2.4.7 Ligation of non-canonical DNA substrates by DV-Lig5

The aim of this experiment what to determine whether DV-Lig5 could successfully ligate nicked DNA substrates containing unnatural base pairs (UBPs) and how the type and placement of these UBPs, in the DNA substrate, would affect the efficiency of ligation.

The UBPs used in this experiment contain two unnatural pyrimidine analogs denoted as S and Z and their complementary partners the purine analogs B and P (**Figure 3.22 A, i**) (Further details in **Section 1.10**). DNA substrates were designed to contain P and Z UPBs (UB_duplex) or S and B UBPs (UB SB_duplex). These UBPs were positioned in different variations and numbers, on either side of the nick in the DNA substrate. Overall, seven different non-canonical DNA substrates (UB_duplex 13, UB_duplex 14, UB_duplex 15, UB SB_duplex 2, UB SB_duplex 13, UB SB_duplex 14 and UB SB_duplex 15) (**Figure 3.22 A, ii**) were used in gel-based activity assays with DV-Lig5. Nick DNA substrate, with non-modified base pairs was used as a positive control.

Analysis of results from the ligation assays showed that DV-Lig5 can ligate successfully ligate nicked DNA substrates containing UBPs. DV-Lig5 was more successful at ligating DNA substrates containing S and B UBPs over P and Z UBPs. DV-Lig5 was particularly efficient at ligating UB SB_duplex2, which only contains one set of the S and B UBP. DV-Lig5 was also able to ligate three other non-canonical DNA substrates at or above 50 % ligation (UB duplex 14, UB SB_duplex 13 and UB SB_duplex 14). Both UB duplex 14 and UB SB duplex 14 contain UBPs on the 5’ side of the nick, same as UB SB duplex 2. UB SB duplex 13, however contains UBPs on the 3’ side of the nick. These results show that DV-Lig5 can ligate non-canonical DNA substrates with UBPs either side of the nick but has a preference for the UBP being on the 5’ side of the nick. Some ligation was also observed on UB_duplex 13, UB_duplex 15 and UB SB_duplex 15, but less than 25 % of the substrate was ligated at the end of the experiment. UB_duplex 13 contains UBPs on the 3’ side of the nick, while both UB_duplex 15 and UB SB_duplex 15 contain UBPs on both sides of the nick. These results suggest that DV-Lig5 is less efficient at ligating multiple UBPs when they are present on both sides of the nick. Activity was observed with both magnesium and

manganese as metal ion cofactors, however there was a trend to higher activity with magnesium. Interestingly, with the addition of magnesium ion to the reaction, DV-Lig 5 could ligate UB SB_duplex 2 and UB SB_duplex 13, at a similar efficiency to that of nick DNA substrate, with natural bases (**Figure 3.22, B, C**).



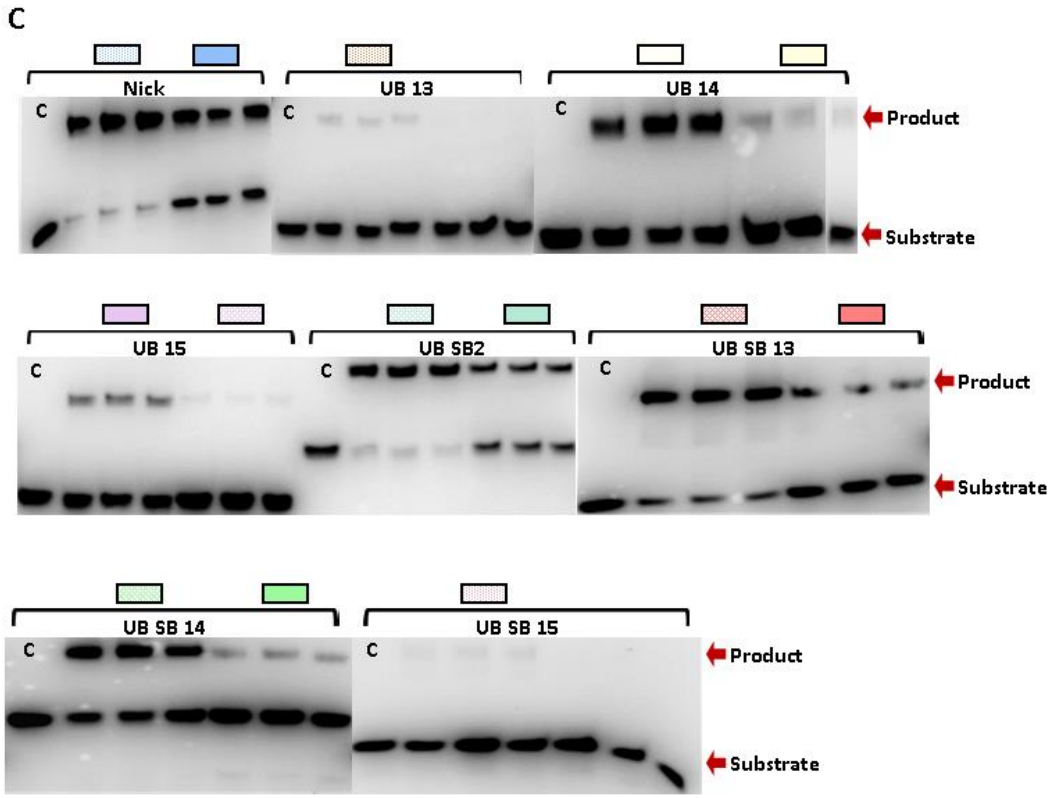


Figure 3.23. Represents the ligation ability of DV-Lig5 on a range of substrates with 3-6 non-canonical expanded base-pair substrates, with either magnesium (Mg) or manganese (Mn) as the divalent metal cofactor. **A**) **i** represents chemical modification of DNA to generate UB and SB DNA duplexes. **ii** represents the seven non-canonical DNA substrates, containing P and Z or S and B UBPs. X on the figures represents natural DNA bases. Stars represent labelling with the 6-carboxyfluorescein at the 5' terminus (5'FAM). **B**) represents the quantitative summary of ligation by DV-Lig5 on nicked DNA and seven different non-canonical substrates. Error bars represent standard deviation. **C**) represents the results of these ligation activity assays shown on urea PAGE gels. Nick DNA substrate, with non-modified bases, is indicated by a blue box. Controls are represented by C and contain no protein. Activity against each substrate was carried out in replicates of three. Product and substrate bands are indicated on the gel, by red arrows. Reactions were carried out for 2 hours, at 25 ^{o}C, with 2 µM final protein concentration, 1mM final concentration of ATP and 10 mM final concentration of metals. Results of activity assays were visualized using iBright™ CL750 Imaging System, Invitrogen™. The graph was generated using GraphPad Prism version 8 (GraphPadSoftware).

3.2.5 Recombinant production of DV-Lig2

3.2.5.1 Small scale expression testing

The gene sequence for DV-Lig2 was cloned into pDEST17 and pHMGWA expression plasmids and transformed into BL21 pLysS, Arctic express and Origami (DE3) *E. coli* expression strains, as described in **Section 2.2.1.** Small scale expression trials were performed using the three *E. coli* strains, at 15 and 20 ^{o}C and results of these trials were run on SDS PAGEs, as described in **Section 2.4.7**. The best protein expression was seen in the Origami strain at 15 ^{o}C.

There was a lot of protein expressed in the pDEST17 plasmids (His-tagged), however, the protein was mostly insoluble. Protein was also expressed in the pHMGWA plasmids (MBP-tagged), giving the best soluble protein expression **(Figure 3.24**).

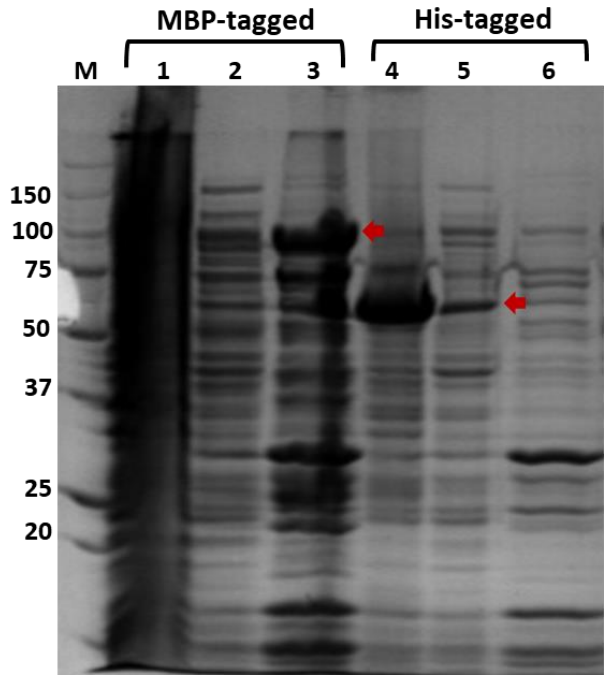


Figure 3.24. SDS PAGE of small scale protein expression results for DV-Lig2. Lanes 1 & 4 represent insoluble protein, lanes 2 & 5 represent soluble protein and lanes 3 & 6 represent soluble protein bound to Ni beads. Red arrows indicate expression of of DV-Lig2 protein, at the expected size for MBP tagged protein (104 kDa) and His-tagged protein (63.5 kDa). A precion plus protein ladder was used as a molecular weight marker (M). The gel was stained with Coomassie Blue stain, using the quick stain method (Wong et al., 2000) and protein bands were visualised and captured on the iBright™ CL750 Imaging System, Invitrogen™.

3.2.5.2 Large scale protein purifications

Following on from results of soluble protein expression of His-tagged and MBP-tagged DV-Lig2 protein, in *E. coli* (DE3) Origami, in small scale screens, protein expression cultures were scaled up following methods from **Section 2.3.3**.

IMAC purification of His-tagged DV-Lig2 protein resulted in no soluble protein expression, as the protein remained in the insoluble form in the pellet sample (data not shown). No further purification steps were taken with His-tagged DV-Lig2 protein. However, an IMAC purification of MBP-tagged DV-Lig2

protein did result in soluble protein expression and protein was further purified using reverse IMAC and gel filtration chromatography.

DV-Lig2 eluted off the IMAC column, with the addition of 40 mM buffer B. Pooled fractions containing DV-Lig2 were de-salted and incubated overnight with TEV protease. Reverse IMAC showed 50 % cleavage of the MBP tag, however tagged and un-tagged DV-Lig2 eluted off the reverse IMAC column in the same fractions. Fractions were pooled, up concentrated, and further purified using gel filtration chromatography. Here tagged and un-tagged DV-Lig2 eluted off the column as a single peak (**Figure 3.25**).

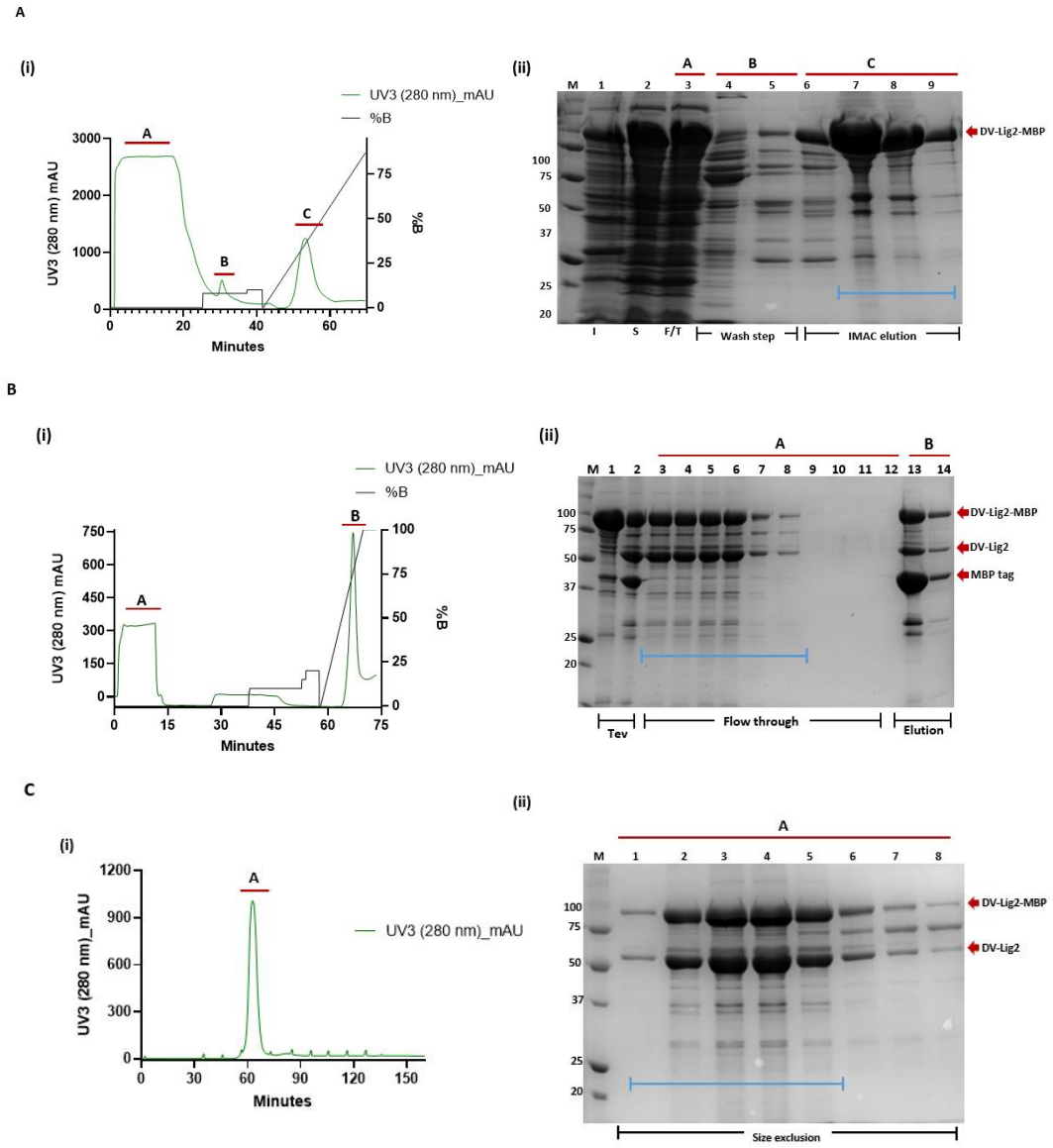


Figure 3.25. IMAC and gel filtration chromatograms (i) and SDS PAGE gels for production of DV-Lig2$_{MBP}$ protein from *E. coli* (DE3) Origami (ii). **A**) IMAC purification of DV-Lig2$_{MBP}$. (i) Peak A represents flow through during IMAC purification, peak B represents fractions of proteins that eluted during the wash step. Peak C

represents fractions of proteins that eluted during the elution step of the IMAC purification, including DV-Lig2-$_{MBP}$ protein (104 kDa). The blue bar indicates fractions that were pooled and incubated overnight with TEV protease, followed by a reverse IMAC purification. **B**) reverse IMAC purification of DV-Lig2. (i) Peak A represents flow through during Reverse IMAC purification, peak B represents fractions of proteins that eluted during the elution step of the Reverse IMAC purification. (ii) Lanes 1-2 are pooled IMAC fractions before the addition of TEV (1) and fractions after an overnight incubation with TEV (2), Lanes 3-12 are fractions from the flowthrough during the reverse IMAC purification. Lanes 13-14 are fractions eluted during the imidazole gradient step. The blue bar indicates fractions that were pooled, up concentrated, and further purified by gel filtration chromatography. **C**) gel filtration purification of DV-Lig2. (i) Peak A represents where DV-Lig2-MBP tagged protein (104 kDa) and DV-Lig2 un-tagged protein (60.1 kDa) eluted off the size exclusion column. (ii) Lanes 1-8 represent the following proteins present in fractions from peak A (i). The blue bar indicates fractions that were pooled and up concentrated and stored for further use. Gels were stained with Coomassie Blue stain, using the quick stain method (Wong et al., 2000) and protein bands were visualised and captured on the iBright™ CL750 Imaging System, Invitrogen™. Chromatogram graphs were designed in GraphPad Prism, version 9.0.0.

Several attempts were made to purify DV-Lig2 un-tagged protein by itself, unfortunately tagged and un-tagged protein always eluted in the same fractions. It was decided to continue ahead with characterisation studies, using tagged and un-tagged protein sample.

3.2.6 Protein folding and secondary structure analysis

No CD data was collected for DV-Lig2 protein, as the protein sample was a mixed proportion of tagged and untagged. Protein folding and secondary structure analysis was based off PDBsum analysis of the DV-Lig2 AlphaFold predicted model and DSF thermal melts.

Secondary structure analysis of the predicted model shows a higher percentage of α-helices over β-strands, and a high percentage of other contributing to the secondary structure of DV-Lig2. DV-Lig2 thermal melt experiments using DSF estimate the T_m of DV-Lig2 to be around 39 ºC. The generated melt curves suggest folded protein (**Figure 3.26**).

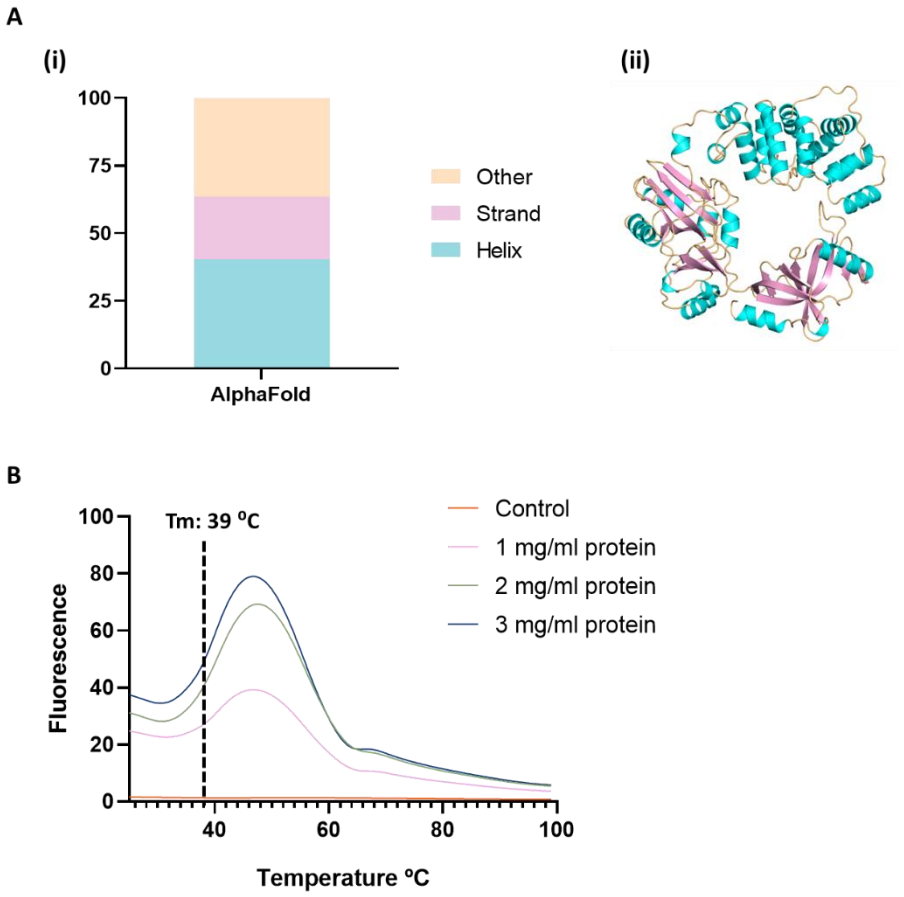


Figure 3.26. DSF and AlphaFold secondary structural composition of DV-Lig2 protein. **A**) (i) A graph representing secondary structural predictions from AlphaFold prediction model, using PDBsum analysis (Laskowski, 2022). (ii) AlphaFold 3D structural prediction of DV-Lig2, coloured based on secondary structure (Helix in blue, strand in pink and other orange. (John Jumper, 2021). **B**) DSF with four different concentrations (1, 2 & 3 mg/ml) of DV-Lig2 protein. T_m values were determined from the midpoint in the unfolding equilibrium and are indicated on the graph, by a dotted line. Each concentration was carried out in replicates of three. Graphs were designed using Prism version 8 (GraphPadSoftware).

3.2.7 Biochemical characterisation of DV-Lig2

Purification of DV-Lig2 resulted in tagged and un-tagged protein sample, which made it difficult to quantify protein concentration. Therefore, in the following activity assays protein concentration was based on nanodrop readings, as described in section **2.4.8**.

DNA binding assays using an EMSA were attempted for DV-Lig2 protein, with nicked DNA substrate, however no binding of protein to DNA was observed. Optimisation with different protein concentrations (5, 4, 3, 2, 1 & 0.5 mg/ml), temperatures (10, 15, 20 & 25 ºC) and time (0.2, 0.5, 1, 1.5 & 2 hours) all gave unsuccessful results.

The following section details the binding ability of DV-Lig2 to DNA substrates in ligation activity on nicked DNA substrates, under different conditions. Additional activity assays show ligation ability on different DNA substrates, using gel-based activity assays.

3.2.7.1 Protein concentration optimisation for DV-Lig2 assays

A protein concentration gradient was carried out to determine the optimal protein concentration to work with in future activity assays.

DV-Lig2 was capable of ligating nicked DNA substrate at all protein concentrations. Ligation reached close to 100 %, with undiluted DV-Lig2. A final concentration of 2 mg/ml was chosen for further activity assays, as higher concentrations often resulted in degradation of the nicked DNA substrate, due to the presence of contaminating nucleases (**Figure 3.27**).

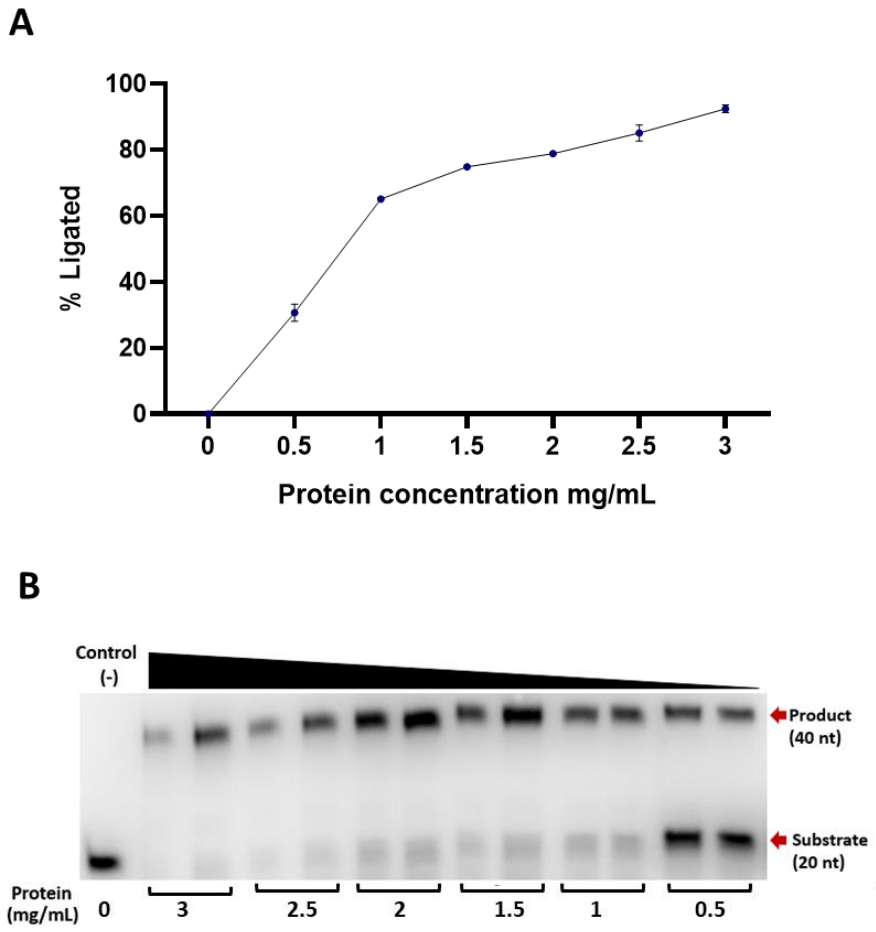


Figure 3.27. Shows ligation of nicked DNA substrate at different concentrations of DV-Lig2 protein. **A**) quantification of ligation by DV-Lig2 on nicked DNA, with different protein concentrations. Points on the graph represent averages of each concentration. Standard deviation error bars are included. **B**) TBE urea gel showing results of DV-Lig2 protein concentration gradient. Control reaction, with nicked DNA, doesn't contain any protein (-). Numbers under gel refer to protein concentration in mg/ml. Product (40 nt) and substrate (20 nt) are indicated by red arrows. Activity against each substrate was carried out in replicates of two. Reactions were carried out for 2 hours, at 25 ^{o}C, with varying final protein concentrations (0.5, 1, 1.5, 2, 2.5 & 3 mg/ml), 1 mM final ATP concentration and 10 mM final concentration of magnesium ions. Results of activity assays were visualized using iBright™ CL750 Imaging System, Invitrogen™. The graph was generated using GraphPad Prism version 8 (GraphPadSoftware).

3.2.7.2 Metal ion preference of DV-Lig2

To determine the optimal divalent metal ion cofactor, an activity assay was performed on nicked DNA substrate, with the addition of magnesium or manganese metal ions. DV-Lig2 can ligate nicked DNA substrate with the addition of both magnesium and manganese. Reactions with the addition of magnesium saw more ligated products formed compared to reactions with manganese, suggesting a preference for magnesium (**Figure 3.28**).

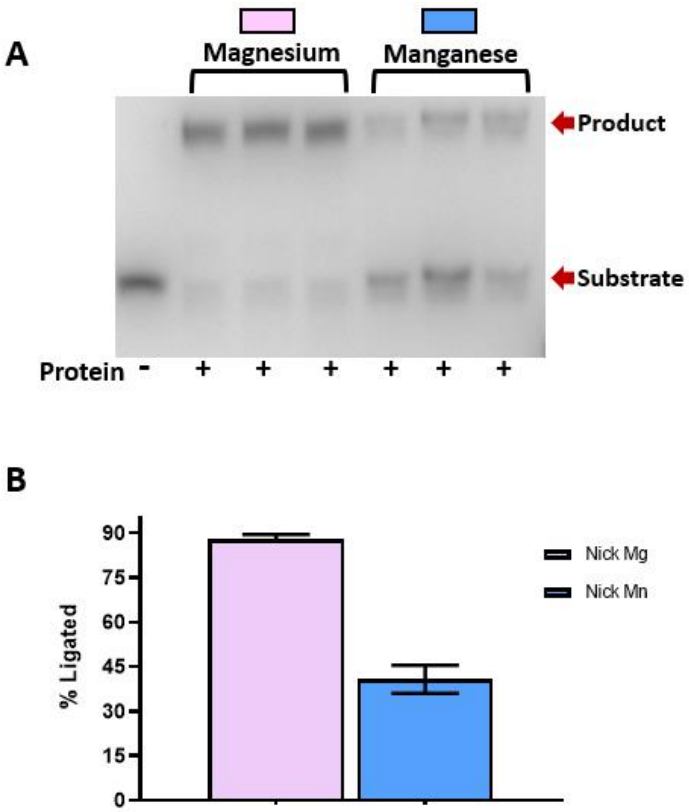


Figure 3.28. Ligation of nicked DNA substrate, by DV-Lig2 protein, with magnesium (Mg) or magnesium. **A**) TBE urea PAGE showing results of ligation with Mg or Mn, in replicates of 3. Addition of protein to the reaction is indicated by a plus symbol (+). Control reactions are indicated by (C), that don't contain protein (-). Product (40 nt) and substrate (20 nt) are indicated by red arrows. **B**) quantification of ligation by DV-Lig2, with either magnesium (Mg) or manganese (Mn) as a cofactor. Bar graphs represent average ligation percentage, for reactions with Mg or Mn. Standard deviation error bars are included. Reactions were carried

out for 3 hours, at 25 ^{o}C, with 2 mg/ml of protein, 1 mM final ATP concentration and 10 mM final metal ion concentrations. Results of activity assays were visualized using iBright™ CL750 Imaging System, Invitrogen™. The graph was generated using GraphPad Prism version 8 (GraphPadSoftware).

3.2.7.3 Nucleotide cofactor specificity of DV-Lig2

To determine if DV-Lig2 could utilize different nucleotide cofactors, for ligation on nicked DNA substrate, an activity assay was performed with the addition of ATP, ADP and NAD nucleotide cofactors.

Ligation on nicked DNA substrate, by DV-Lig2, cannot occur with the addition of either ATP or ADP nucleotide cofactors. NAD nucleotide cofactor does not support ligation, by DV-Lig2. Overall, ATP is the preferred nucleotide for ligation by DV-Lig2 on nicked DNA (**Figure 3.29**).

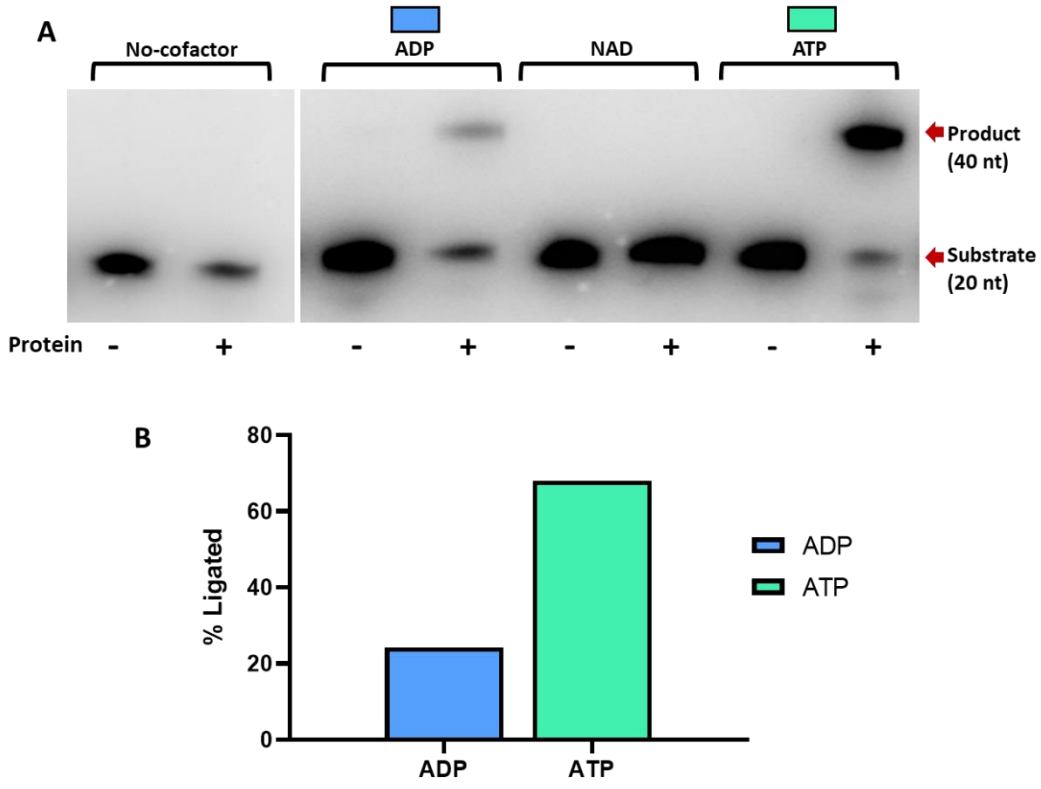


Figure 3.29. Ligation of nicked DNA substrate, by DV-Lig2 protein, with different cofactors. **A**) TBE urea PAGE showing results of ligation by DV-Lig2, with and without the addition of different cofactors. Addition of protein to the reaction is indicated by a plus symbol (+). Controls reactions were used that don't contain protein (-) or don't contain cofactor (No cofactor). Product (40 nt) and substrate (20 nt) are indicated by red arrows. **B**) quantification of ligation by DV-Lig2 on nicked DNA, with different cofactors (ATP & ADP). Reactions were incubated for 2 hours at 25 ^{o}C with nicked DNA substrate, 2 mg/ml of protein, 10 mM magnesium ion and different cofactors at 1 mM final concentration. Results of activity assays were visualized using iBright™ CL750 Imaging System, Invitrogen™. The graph was generated using GraphPad Prism version 8 (GraphPadSoftware).

3.2.7.4 Temperature dependence of DV-Lig2

A temperature gradient assay, from 10 ºC to 55 ºC, was used to determine the optimal temperature for ligation by DV-Lig2 protein,

DV-Lig2 can ligate nicked DNA substrate from 10 to 55 ºC, with best ligation occurring between 20 to 30 ºC. Ligation rates declined at reactions temperatures 40 ºC and above. The optimal ligation temperature for DV-Lig2 is between 20 to 40 ºC (**Figure 3.30**). Further activity assays were performed at 25 ºC, to ensure optimal ligation.

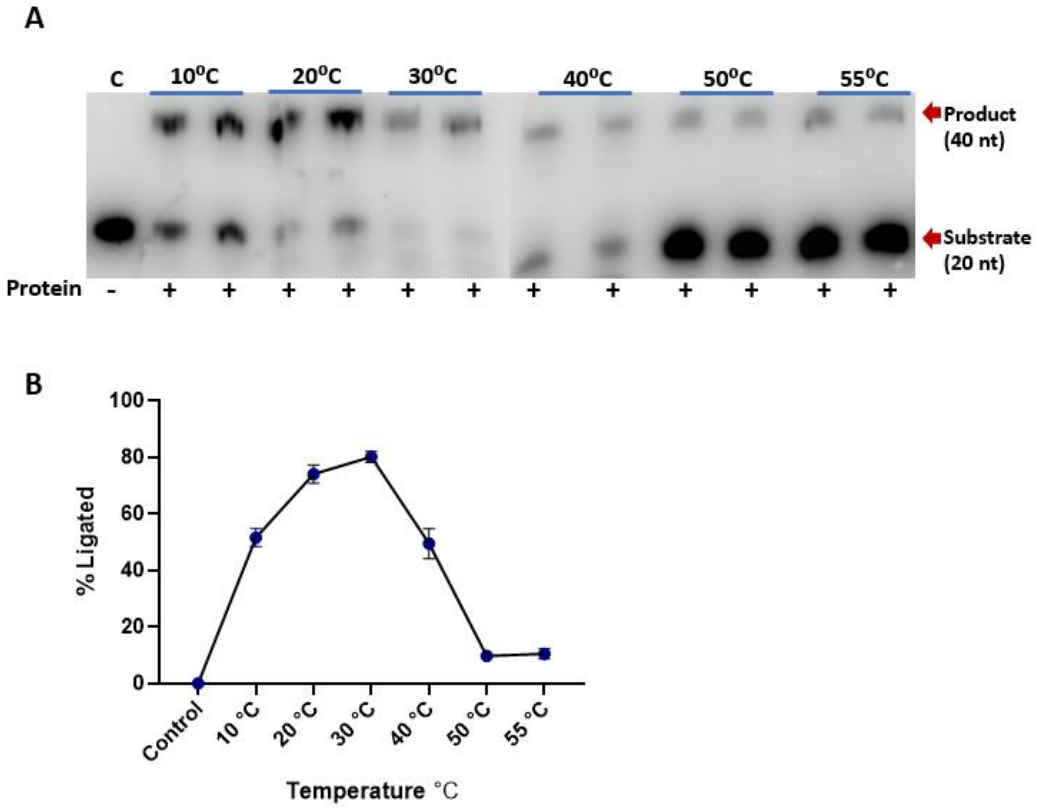


Figure 3.30. Ligation of nicked DNA substrate, by DV-Lig2 protein, at varying reaction temperatures. **A**) TBE urea PAGE showing results of ligation by DV-Lig2, at different temperatures (10, 20, 30, 40, 50 & 55 ºC). Addition of protein to the reaction is indicated by a plus symbol (+). Controls reactions (C) don't contain any protein (-). Product (40 nt) and substrate (20 nt) are indicated by red arrows. Activity against each substrate was carried out in replicates of two. **B**) quantification of ligation by DV-Lig2 on nicked DNA, with different reaction temperatures. Plots on the graph represent averages of each reaction temperature. Reactions were carried out for 5 hours, at varying temperatures, with 2 mg/ml final protein concentration, 1 mM final ATP concentration and 10 mM final magnesium ion concentration. Results of activity assays were visualized using iBright™ CL750 Imaging System, Invitrogen™. The graph was generated using GraphPad Prism version 8 (GraphPadSoftware).

3.2.7.5 DNA substrate specificity of DV-Lig2

Previous activity assays were conducted with nick DNA substrate and show that DV-lig2 is capable of ligating nick DNA substrates. The same DNA substrates (cohesive, blunt, mismatch and gapped) tested against DV-Lig5 (**Section 3.2.4.6**), were also used in activity assays with DV-Lig2, to determine if DV-Lig2 was capable of ligating DNA substrates with different types of DNA breaks.

DV-Lig2 can ligate A/C mismatch and cohesive DNA substrates, in addition to nicked DNA substrate. No ligation was observed on blunt or gapped DNA substrates. DV-Lig2 shows the greatest ligation ability on nicked DNA, followed by cohesive, then mismatch DNA (**Figure 3.31**).

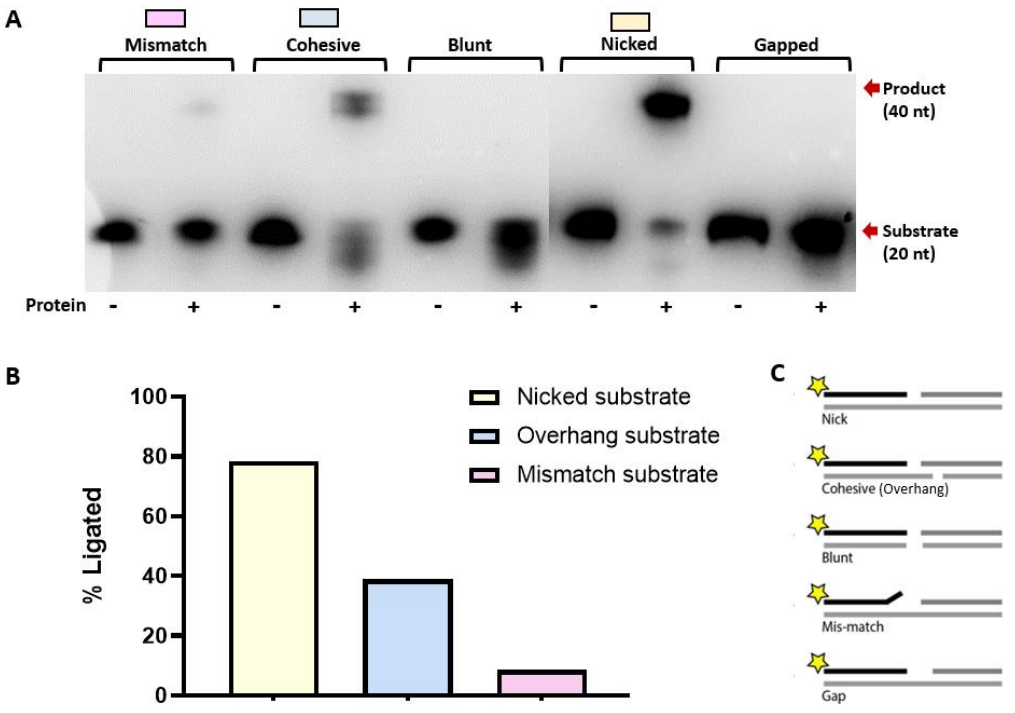


Figure 3.31. Results of ligation, by DV-Lig2, on different DNA substrates. **A**) TBE urea PAGE showing results of ligation, by DV-Lig2 on 5 different DNA substrates. Substrate (20 nt) and product (40 nt) are indicated by red arrows. Addition of protein to reaction is indicated by a plus symbol (+), controls (-) don't contain any protein. **B**) quantification of ligation by DV-Lig2 on nicked, cohesive (overhang) and mismatch DNA substrates. **C**) schematic of DNA substrates used in reactions. Star indicates fluorescent label. Reactions were carried out for 8 hours, at 25^{o}C, with 2 mg/ml final protein concentration, 1 mM final concentration of ATP and 10 mM final concentration of magnesium. Results of activity assays were visualized using iBright™ CL750 Imaging System, Invitrogen™. The graph was generated using GraphPad Prism version 8 (GraphPadSoftware).

3.2.7.6 Ligation of non-canonical DNA substrates by DV-Lig2

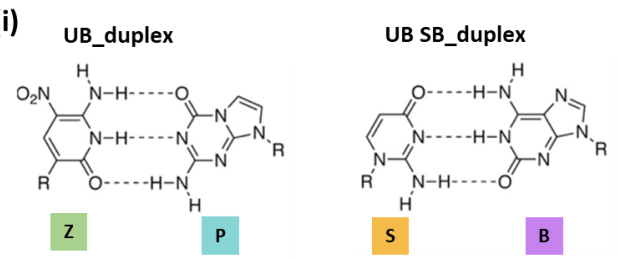
The ligation ability of DV-Lig2 protein was tested on nicked DNA substrates containing unnatural base pairs (UBPs), to determine how the type and placement of these UBPs, in the DNA substrate would affect the efficiency of ligation.

The UBPs used in this experiment have been previously described in **Section 1.10** and **Section 1.1.1.1.** Overall, seven different non-canonical DNA substrates (UB_duplex 13, UB_duplex 14, UB_duplex 15, UB SB_duplex 2, UB SB_duplex 13, UB SB_duplex 14 and UB SB_duplex 15) (**Figure 3.32, A**) were used in gel-based activity assays with DV-Lig2. Nick DNA substrate, with non-modified base pairs was used as a positive control.

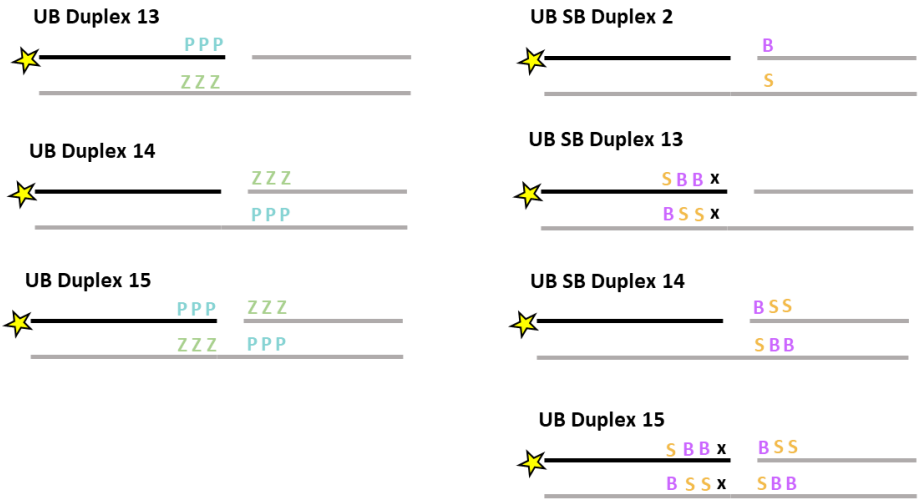
Analysis of results from the ligation assays showed that DV-Lig2 struggled to successfully ligate several of the nicked DNA substrates containing UBPs (**Figure 3.32, B, C).** Ligation was only observed on one P-Z UBP containing substrate (UB_duplex 13), and two S-B UBP containing substrates (UB SB_duplex 2 and 14). Activity was observed with both magnesium and manganese as metal ion cofactors. Activity on nicked DNA substrates, with non-modified bases and P-Z containing substrates, showed better ligation activity with the addition of magnesium over manganese. While activity on S-B containing substrates, was improved with the addition of manganese over magnesium. Overall, the ligation efficiency on these UBP containing substrates was very low when compared to ligation of the natural nicked DNA substrate. Further testing with these UBPs is required for DV-Lig2, with more replicate reactions included. However, DV-Lig2 is very difficult to purify, and optimisation of protein purifications is essential for future experiments.

A

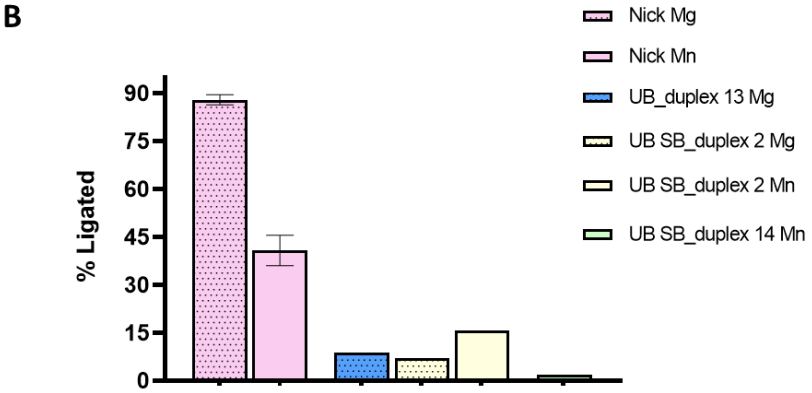
(i)


UB_duplex
UB SB_duplex
Z
P
S
B

(ii)


UB Duplex 13
P P P
Z Z Z
UB SB Duplex 2
B
S
UB Duplex 14
Z Z Z
P P P
UB SB Duplex 13
S B B x
B S S x
UB Duplex 15
P P P
Z Z Z
Z Z Z
P P P
UB SB Duplex 14
B S S
S B B
UB Duplex 15
S B B x
B S S x
B S S
S B B

B


% Ligated
90
75
60
45
30
15
0
Nick Mg
Nick Mn
UB_duplex 13 Mg
UB SB_duplex 2 Mg
UB SB_duplex 2 Mn
UB SB_duplex 14 Mn

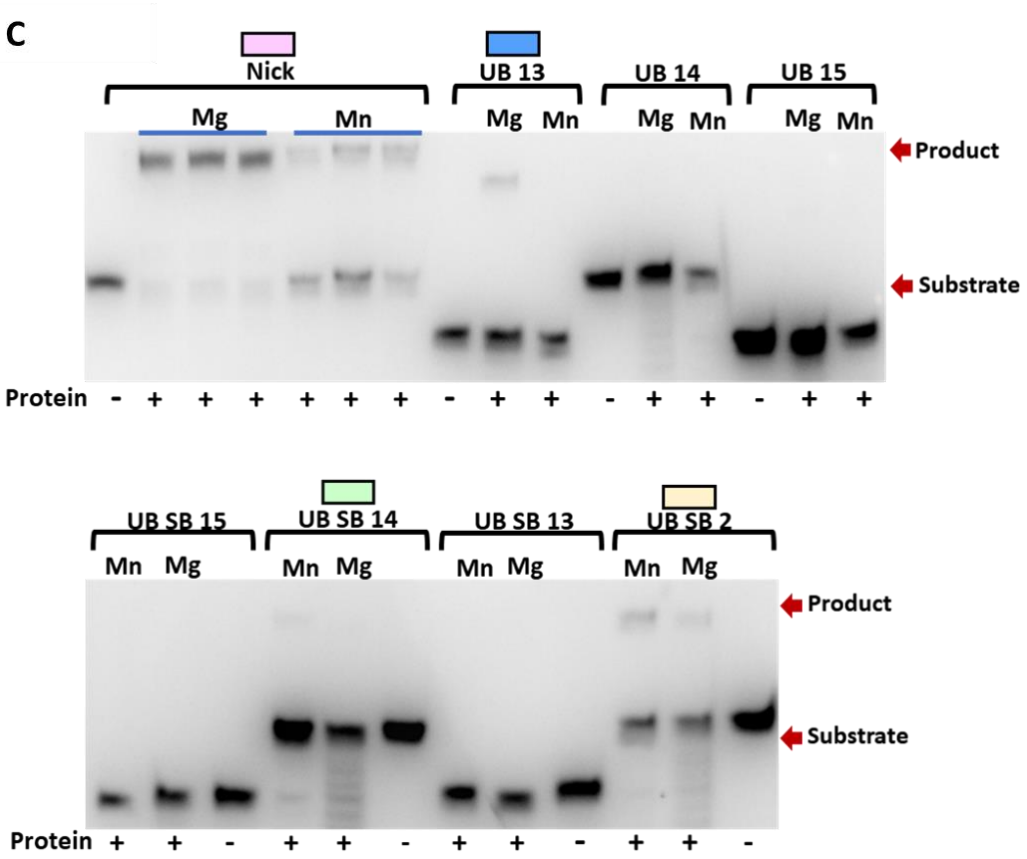


Figure 3.32. Represents the ligation ability of DV-Lig2 on a range of substrates with 3-6 non-canonical expanded base-pair substrates, with either magnesium (Mg) or manganese (Mn) as the divalent metal cofactor. **A**) **i** represents chemical modification of DNA to generate UB and SB DNA duplexes. **ii** represents the seven non-canonical DNA substrates, containing P and Z or S and B UBPs. X on the figures represents natural DNA bases. Stars represent labelling with the 6-carboxyfluorescein at the 5' terminus (5'FAM). **B**) represents the quantitative summary of ligation by DV-Lig2 on nicked DNA and seven different non-canonical substrates. Error bars represent standard deviation. **C**) represents the results of these ligation activity assays shown on urea PAGE gels. Nick DNA substrate, with non-modified bases, is indicated by a pink box. Controls contain no protein. No replicates were carried out. Product and substrate bands are indicated on the gel, by red arrows. Reactions were carried out for 2 hours, at 25^{o}C, with 2 mg/ml final protein concentration, 1mM final concentration of ATP and 10 mM final concentration of metals. Results of activity assays were visualized using iBright™ CL750 Imaging System, Invitrogen™. The graph was generated using GraphPad Prism version 8 (GraphPadSoftware).

3.3 Discussion

DV-Lig2 and DV-Lig5 were identified through bioinformatic analysis as belonging to the Lig B family of ATP-dependent DNA ligases. DV-Lig2 belongs to a four-cluster gene operon similar to that previously discovered in *P. putida* and other bacteria (Ejaz & Shuman, 2018). The DV-Lig2 operon arrangement belongs to a class I configuration of genes, in the order of an exo nuclease, followed by an ATP dependent DNA ligase, helicase Lhr-core, and an endonuclease MPE. The Lhr helicase and MPE, belonging to this gene cluster, have been previously characterized in *P. putida* (Ejaz et al., 2019; Ejaz & Shuman, 2018; Ghosh et al.,

2021). While the ligase and exo nuclease from this operon have not been biochemically characterized, *in silico* analysis showed that the ligase belongs to the ATP-dependent DNA ligase clade, with domains homologous to human DNA ligase I. The exo nuclease belongs to the MBL-β-CASP sub family of nucleases and shares homology to other members of this family such as SNM1B (Ejaz & Shuman, 2018). The clustering of genes encoding enzymes with related biochemical activities is suggestive of their participation in a common physiological pathway. Ejaz and Shuman have speculated that the enzymes, in this operon might play a role in a dedicated DNA repair pathway potentially involving inter-strand DNA cross-links or NHEJ, due to the apparent role of mycobacterial Lhr and eukaryal SNM1s in mitomycin C sensitivity (Ejaz & Shuman, 2018). DV-Lig2 and other enzymes in this operon share sequence similarity to those in the *P. putida* gene cluster and are likely to have a similar role in DNA repair. DV-Lig5 is positioned in a distinct gene arrangement different from DV-Lig2, where upstream genes encode proteins with DNA modifying functions. Several hypothetical proteins are also present in this gene region, however, *in silico* prediction of the two hypothetical proteins flanking DV-Lig5 did not reveal any potential function for either.

As well as being identified in bacteria, Lig B type DNA ligases are found as replicative enzymes in Archaea, and share structural homology to human DNA ligase I (hLigI) (Williamson & Leiros, 2020). In agreement with sequence-based predictions, AlphaFold models of DV-Lig2 and DV-Lig5 indicate both proteins are made up of three main domains typical of LigB type ATP-dependent DNA ligases: a DNA binding (DB) domain, an adenylation (AD) domain and an oligonucleotide binding (OB) domain. The models of DV-Lig2 and DV-Lig5 are highly similar to crystal structures from *A. fulgidus*, *P. furiosus*, *T. sibiricus*, *S. solfataricus* and hLigI. These ligases have been implicated as important enzymes in DNA repair pathways as well as replication, for example, hLigI is involved in BER, NER and SSBR (Sallmyr et al., 2020).

DV-Lig2 and DV-Lig5 adopt the typical arrangement of domains, seen for Lig B type ligases and contain positive contacts within the centre of these domains where DNA is likely to bind, as shown in the crystal structure of human

DNA ligase I (1X9N), bound to nicked DNA substrate (Pascal et al., 2004). The AD domain and OB domain are common to all ATP-dependent DNA ligases, while in bacteria the DB domain in conjunction with AD domain and OB domain is exclusive to Lig B type ligases (Williamson & Leiros, 2020). The presence of this DB domain allows the ligase to completely wrap around DNA substrates (Shi et al., 2018). DV-Lig2 has additional alpha helices in the DB domain and is predicted to form polar contacts with the OB domain when superimposed onto hLigI DB and OB domains in a DNA-bound configuration. One of these polar contacts forms a salt bridge between Arg-473 from the OB domain and Asp-74 from the DB domain. Salt bridges between the OB, and DB domains have been observed in DNA-bound crystal structures of other DNA ligases, such as T4 DNA ligase and *Chlorella* virus ligase, where they complete the encirclement of DNA (Shi et al., 2018). There are also additional salt bridge interactions predicted between the AD and OB domains in DV-Lig2 and the AD and DB domains in both proteins. In h-LigI a salt bridge is formed between the OB and AD domains, which orientates these domains as a continuous binding surface and stabilizes the AD and OB domain interface (Pascal et al., 2004). Analysis of the predicted catalytic sites of DV-Lig2 and DV-Lig5 within the AD and OB domains are highly similar to other previously characterised DNA ligases. As expected, both contain six motifs, as shown in **Appendix C.2**, that are conserved among the nucleotidyl transferase family that includes mRNA capping enzymes and RNA ligases in addition to DNA ligases (Shuman, 2009). Motif I, from the AD domain, contains the active site lysine residues (K-230 in DV-Lig2 and K-208 in DV-Lig5) that forms polar contacts with the AMP residue when superimposed onto the crystal structure of h-LigI bound to nicked DNA. This super imposition also reveals key residues from all domains in DV-Lig2 and DV-Lig5 that form polar contacts with the DNA. Residues from the DB domain only interact with the DNA backbone, similar to the interactions seen by h-LigI (Pascal et al., 2004).

In DNA binding experiments, DV-Lig5 demonstrated robust binding to a nicked DNA substrate. No binding was observed with DV-Lig2 in gel shift experiments, although this is likely due to sample preparation, specifically the presence of the residual MBP tag remaining on a large fraction of the enzyme and does not imply that DV-Lig2 does not bind to DNA. Similar to other DNA ligases

such as T4 DNA ligase, h-Lig1, Chlorella virus DNA ligase, and *E. coli* Lig A (Chauleau & Shuman, 2016; Cherepanov & De Vries, 2003; Odell & Shuman, 1999; Pascal et al., 2004), both DV-Lig2 and DV-Lig5 exhibit the highest ligase activity on nicked DNA, as shown from activity assays. DNA ligases have a general preference for double-stranded DNA (dsDNA) substrates that exhibit Watson-Crick base pairing, as opposed to substrates containing one or more mismatches. However, some ligases are capable of ligating certain mismatches to a significant extent. Active ligases, like T4 DNA Ligase, demonstrate high efficiency in ligating nicks near the ligation junction that contain one or more mismatches (Lohman et al., 2016; Wu & Wallace, 1989). The mechanism by which ligases ensure proper base pairing involves interrogating the dsDNA through interactions with the minor groove, rather than reading specific base sequences (Shuman, 2009). Ligases are sensitive to distortions in the helix shape, which guides their recognition of suitable base pairing configurations (Liu et al., 2004). h-LigI has the ability to ligate a G:T mismatch, as its active site can accommodate a G:T pairing in the wobble conformation (Tang et al., 2022). Both DV-Lig2 and DV-Lig5 were able to ligate an A:C mismatch DNA, although less ligation occurred compared to the nicked DNA substrate. It is likely that some ligases are more tolerant of mis-matches due to a more flexible active site (Bilotti et al., 2022).

While most ligases have strong activity on substrates containing a single strand break in one strand of a duplex (i.e. nick ligation), not all DNA ligases efficiently join two DNA fragments with short complementary overhangs and few join blunt ends in the absence of accessory proteins (Bilotti et al., 2022). In most cases, blunt and cohesive-end sealing activity requires the presence of specific DNA binding domains in the ligase. In hLigI and T4 DNA ligases, the addition of an N-terminal DB domain allows these ligases to completely encircle the DNA substrate. Removal of residues from the N-terminal domain of T4 DNA ligase, results in loss of activity on blunt-ended and cohesive DNA substrates (Bauer et al., 2017). While both DV-Lig2 and DV-Lig5 contain an equivalent DB domain, only DV-Lig2 was able to ligate the cohesive DNA substrate. T4 DNA ligase contains a salt bridge between Lys-384 from the OB domain and Asp-112 from the DB domain, which completes the encircling of DNA by T4 DNA ligase (Shi

et al., 2018). A salt bridge is also predicted in DV-Lig2 between Arg-474 from the OB domain, and Asp-74 from the DB domain, which likely gives this protein the ability to ligate cohesive DNA substrates, similar to that observed in T4 DNA ligase. DV-Lig5 is missing an equivalent interaction between the DB and OB domains, which might contribute to this enzyme not being able to ligate cohesive DNA substrates. No ligation was observed with blunt or gapped DNA substrates, from either ligase. In literature, it was often observed that ligation of blunt or gapped DNA substrates, required the addition of macromolecular crowding reagents, such as polyethylene glycol (PEG), to increase the efficiency of ligation (Wang et al., 2019). This could be examined in future work.

As explained in **Section 1.10**, unnatural base pairs (UBPs) represent an artificially expanded genetic system where the traditional nucleotides are extended by the inclusion of two unnatural pyrimidine analogs (S and Z) and their complementary partners (B and P) (Benner et al., 2016). Bentz and colleagues identified bacterial DNA polymerases, capable of replicating DNA containing UBPs. For example Klen Taq polymerase has the ability to replicate unnatural base pairs by inducing Watson-Crick geometry (Betz et al., 2012). Recent studies have shown that several commercially available ligases can catalyse ligation of modified nucleic acids known as xeno nucleic acids (XNAs) including 2'OMe, HNA, LNA, TNA, and FANA (Duffy et al., 2020).

The ligation ability of both DV-Lig2 and DV-Lig5 was tested on these non-canonical DNA substrates to determine if these enzymes had the ability to join UBP-DNA. DV-Lig5 could ligate all non-canonical DNA substrates, particularly the SB UBPs. DV-Lig5 also had no preference for UBP being at the 5' or 3' end of the nick, however less activity was observed on substrates with the UBP at each end of the nick (duplex_15). DV-lig2 was less effective at ligating non-canonical DNA substrates, compared to DV-Lig5 and was only capable of a small degree of ligation on UBPs, UB_duplex 13, UB SB_duplex 2 and UB SB_duplex 14. DV-Lig2 could ligate substrates with the UBP positioned at the 3' and 5' end of the nick, but not at each end of the nick. Best ligation on these non-canonical substrates was observed with UB SB_duplex 2, with the UBP positioned at the 5' end of the nick.

Phosphoryl transfer enzymes, such as ligases, can use magnesium as the physiological cofactor. In addition to magnesium many can also use manganese (Taylor, 2014). Both DV-Lig5 and DV-Lig2 showed the ability to ligate nicked DNA substrates with the addition of magnesium or manganese, however these enzymes had improved ligation efficiency with magnesium rather than manganese. This indicates that while manganese is accepted, magnesium is the preferred divalent ion. This preference for magnesium over manganese is also observed by *M. tuberculosis* Lig A (Srivastava et al., 2005). In differential scanning fluorimetry (DSF) thermal melts with DV-Lig5, the addition of metal ions at a concentration of 5 mM saw the same T_m for DV-Lig5 as melts without metal ion additives. Increasing the concentration of metal ions in these thermal melts saw a decrease in T_m for DV-Lig5, with manganese having a destabilizing effect at higher concentrations than magnesium. In literature, ligation by DNA ligases, on nicked DNA has also been achieved with the addition of DNA. Ligases have also been shown to have ligation with the addition of other metal ions, such as calcium (Zhang & Tripathi, 2017) or cobalt (Zhu & Shuman, 2007). It would be interesting to see if either enzyme would also be able to ligate DNA substrate with the addition of these other metal ions.

Both DV-Lig2 and DV-Lig5 were able to utilize ADP and ATP as a nucleotide cofactor, which is required for ligation of a nicked DNA substrate. Interestingly, the models of DV-Lig2 and DV-Lig5 are most similar to crystal structures of Archaeal ATP dependent DNA ligases. Some Archaeal ATP-dependent DNA ligases have been reported to utilize additional nucleotide cofactors, like ADP, such as *Desulfurococcales* from the phylum crenarchaeota (Seo et al., 2007). Meanwhile *Sulfophobococcus zilligii*, which also belongs to crenarchaeota, was found to utilize GTP in addition to both ATP and ADP as a nucleotide cofactor (Sun et al., 2008). The use of GTP as a nucleotide cofactor was also tested for DV-Lig5, it was found to slightly increase ligation of product over the basal levels observed. However, due to the background remaining from pre-adenylation from the *E. coli* expression, further experiments are needed to confirm this which could involve observing ligation activity over a range of GTP concentrations. If the protein is using GTP, then there should be some increase in ligation with higher GTP concentration up until a point where the concentration of

GTP becomes inhibitory for ligation. Ideally, protein needs to be to be pre-incubated with unlabeled DNA substrate, to a point where the protein is no longer adenylated and does not show any ligation ability without the addition of a cofactor. This was attempted several times for DV-Lig5, however despite this pre-treatment, there were always basal levels of ligation observed in reactions without cofactor present.

Currently, only Archaea have been reported to utilize additional nucleotide cofactors like ADP and GTP, with many classified as hyperthermophilic Archaea, which inhabit extreme environments. It has been suggested that enzymes from hyperthermophiles might possess biochemical features of an ancestral prototype that existed under extreme, limiting environments, where it is likely that the reaction energy source was not only ATP, but also additional NTPs (Adul Rahman et al., 1997; Fujiwara et al., 1996). While DV-Lig2 and DV-Lig5 are derived from bacterial lineages and not Archaea, these bacteria also occupy an extreme environment under sub optimal conditions. It is possible that DV-Lig2 and DV-Lig5 can utilize additional nucleotide cofactors, as ATP may not always be available as an energy source.

As these results are unusual for bacterial DNA ligases, one must be skeptical about the validity of these ligases utilizing ADP and GTP as alternative nucleotide cofactors. Chen and his colleagues raised a concern that contaminating *E. coli* kinase in protein preps, can result in conversion of ADP into ATP and AMP and thereby confound interpretation of the ADP/ATP specificity experiments (Chen et al., 2009). Capturing crystal structures of these enzymes with each nucleotide bound is one option to validate these results. Other options would be to probe the potential generation of ATP by *E. coli* contaminants by mass-spectrometry of the sample in the presence of ADP, or to use mass spectrometry to detect the kinase contaminant in the sample directly.

Due to various factors, including nuclease contamination, it was not possible to conduct complete temperature-dependent ligation activity assays for DV-Lig2 and DV-Lig5. Preliminary findings indicate that DV-Lig2 is capable of ligating nicked DNA substrate within a temperature range of 10 ºC to 55 ºC.

However, the efficiency of ligation began to decline beyond 40 ºC. The most favorable temperature for ligation was observed to be 30 ºC. These results align with the DSF thermal melts analysis, which demonstrated a melting temperature (T_m) of approximately 39 ºC for the protein. These findings also demonstrated that DV-Lig5 was able to ligate nicked DNA substrate from -40 ºC to 20 ºC. Conclusive results regarding activity at higher temperatures were impeded by the presence of contaminating nuclease in the sample preparation. Further investigation is required to determine a lower temperature threshold that inhibits product ligation. Thermal melts generated for DV-Lig5 using DSF SYPRO and CD techniques indicated a higher T_m compared to DV-Lig2, suggesting a potential greater tolerance of ligation at elevated temperatures. The optimal temperature for most DNA ligases is around 37 ºC (Suzuki et al., 2016). DNA ligases from extremophiles can exhibit ligation on DNA substrates at temperature extremes, such as the DNA ligase from *Thermococcus fumicolans*, with an optimum temperature of 65 ºC (Rolland et al., 2004). A DNA ligase from psychrophilic *Aliivibrio salmonicida* has an optimum of 15 ºC. Of note, not all DNA ligase from psychrophiles are cold-adapted, for example DNA ligases from *Psychromonas* sp. strain SP041 (Psy-Lig) and *Pseudoalteromonas artica* (Par-Lig) both have temperature optima in the range of 35-40 ºC and unfolding temperatures greater than 45 ºC (Suzuki et al., 2016). DV-Lig2 and DV-Lig5, while still being able to ligate at lower temperatures, have similar temperature optimums, as Psy-Lig and Par-Lig, which suggests these enzymes are also not cold adapted.

Based on findings from structural and functional characterisation of DV-Lig2 and DV-Lig5, it is evident that these enzymes are ATP-dependent DNA ligases of the Lig B subfamily. Ligases from this family participate in several DNA repair pathways such as the ligase from *P. putida*, which has been suggested to participate in NHEJ (Ejaz & Shuman, 2018).

4 Chapter 4

DV-Nuclease-ligase fusion protein from a unique gene cluster

4.1 Introduction

Operons or clusters of coregulated genes are used by Bacteria, Archaea, and certain Fungi to organize their genomic information. A gene cluster comprises a syntenic group of genes, their intervening non-coding sequences, and nearby regulatory elements. Typically, the genes within a cluster are involved in the same pathway or process (Kountz & Balskus, 2021). Grouping related genes under a common control mechanism allows these organisms to rapidly adapt to changes in their environment. Within many bacterial and archaeal genomes various types of gene clusters encoding DNA repair proteins have been discovered. The grouping of these DNA repair genes within the same operon or together in a gene cluster, suggests their involvement in the same DNA repair pathway.

Present in many bacterial species is a gene cluster made up of an Lhr superfamily 2 helicase, a binuclear metallo-phosphoesterase (MPE), an ATP dependent DNA ligase and a metallo-β-lactamase exonuclease (Ejaz & Shuman, 2018; Ordonez & Shuman, 2013). Genetic linkage of a helicase and DNA nuclease with a ligase and a putative exonuclease, suggests that these enzymes participate in a bacterial DNA repair pathway (Ejaz & Shuman, 2018). This gene cluster was previously discussed in section 3.1 and has been included here, due to the discovery of a nuclease-ligase fusion protein from the DV-metagenomes. This fusion protein has structural and sequence homology to the exo nuclease-ligase fusion protein, identified in the four gene cluster from *O. terrae*, discussed below in **Section 4.2.1** (Ejaz & Shuman, 2018).

There is limited information in the literature on these nuclease-ligase fusion proteins, aside from their identification in *O. terrae.* However, the function of these domains as separate proteins, has been characterized *in silico* from *P. putida* (Ejaz & Shuman, 2018). Here the ligase protein is classified as an ATP-

dependent DNA ligase, made up of three domains: a DB domain, an AD domain and an OB-fold domain, common to Lig B type DNA ligases. (More detail on Lig B type ligases in **Section 3.1**). The adjacent exo nuclease is described as an MBL-β-CASP nuclease, with homology to SNM1A, SNM1C and SNM1B type nucleases. Metallo-β-lactamase fold proteins belong to a vast superfamily of proteins that exhibit the capability to interact with a wide range of substrates. Within this superfamily, there are many smaller family groups, including class B β-lactamases that hydrolyze lactams, glyoxalase II, aryl sulfatases, cytidine monophosphate-N-acetyl neuraminic acid (CMP-NeuAc) hydrolases, cAMP phosphodiesterases, and the phnP protein. (Yosaatmadja et al., 2021). Another family within the Metallo-β-lactamase fold superfamily acts specifically on DNA substrates (Fernandez et al., 2011). These proteins are recognized by a distant globular domain, known as the β-CASP motif, after metallo-β- lactamase associated C PSF A rtemis S NM1/ P SO2 (Yosaatmadja et al., 2021) The proteins from which this motif is named after, include the 73 kDa subunit of cleavage and polyadenylation specificity factor (CPSF) and its yeast orthologue Ysh1p, involved in RNA processing (Huang et al., 2023), SNM1 and PSO2, implicated in DNA crosslink repair and finally Artemis a novel member of this group, identified to be involved in V(D)J recombination and DNA repair (Fernandez et al., 2011; Yosaatmadja et al., 2021).

Examples in the literature of other ligase nuclease fusion proteins where the nuclease is not of the MBL-type have been identified in the genomes of *M. tuberculosis* and *P. aeruginosa.* (Della et al., 2004; Malyarchuk et al., 2007). These fusion proteins have been classified as Lig D and consist of a third domain fusion to a DNA polymerase (**Figure 4.1**). Mt-Lig D has three distinct domains, and each domain possesses individual activities that participate in DNA end processing or DNA ligation. The N-terminal domain, also known as the polymerase domain, has been shown to have terminal transferase activity, DNA-dependent RNA primase and DNA dependent DNA/RNA gap-filling polymerase functions. The central domain, known as the phosphoesterase-nuclease (PE) domain, has 3'-5' single stranded DNA exonuclease activity, requiring magnesium or manganese. The C-terminal domain of Mt-Lig D is the ligase

domain and is an active adenyltransferase, catalyzing nick sealing in double-stranded DNA (Wright et al., 2010).

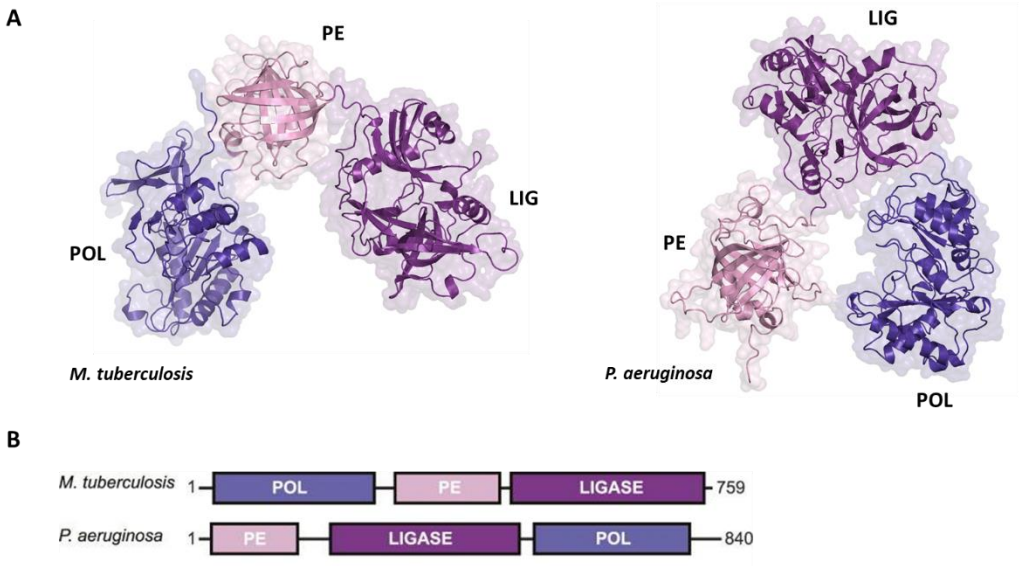


Figure 4.1. Domains of Lig D from *M. tuberculosis* and *P. aeruginosa*. A) *In silico* predictions of Lig D atomic structure from *M. tuberculosis* and *P. aeruginosa*, predicted by AlphaFold (Jumper et al., 2021), displayed as both cartoon and surface representations. POL–blue; PE–pink; LIG–purple. B) Domain arrangement of Lig D in *M. tuberculosis* and *P. aeruginosa*. POL, polymerase domain; PE, phosphoesterase domain; LIGASE, ligase domain. Figure adapted from (Amare et al., 2021).

4.2 Results

4.2.1 Discovery of a novel gene cluster in Antarctic Dry Valley metagenome

Using sequence similarity network (SSN) analysis, a novel cluster of DNA modifying enzymes were identified from Antarctic Dry Valley metagenomes. These genes were identified in a contig (Ga0136611_10000860) predicted to belong to the *Chthoniobacter* genus and are hypothesized to contribute to the extreme survival capabilities of bacteria living in this extreme, DNA damaging environment. Genes within this contig were predicted to encode for protein homologs of an ATP dependent DNA ligase, an unknown protein with a RecA-like domain, an ImuB DNA polymerase, and an error prone DNA polymerase. All five proteins have low sequence identity to homologs in the NCBI databases (ranging from 69-43 %), and in all cases the top hits were to uncharacterised gene products, predominantly from uncultivated organisms.

Of particular interest is the ATP-dependent DNA ligase, belonging to this cluster. This predicted Lig B DNA ligase contains an N-terminal fusion with a metallo-hydrolase nuclease domain. The Lig B class DNA_ligase_A_N (PF04675) N-terminal DNA binding domain was present in 1791 DV-metagenome proteins. At the 50% threshold, most of these proteins formed groups with numerous UniRef sequences. Cluster #1 (**Figure 4.2**, **A**), however, contained fewer UniRef representatives and included DV-metagenome Lig B sequences that were significantly longer than others in the network. Further searches using hmmscan revealed 13 sequences in this cluster that contained a fusion of the Zn-dependent metallo-hydrolase RNA specificity domain RMMBL (PF07521) at the N-terminus of the DNA ligase DNA binding domain (**Figure 4.2**, **B**). As this gene cluster has not been studied in other Bacteria, it is of interest to determine what role these proteins may play in a potential DNA repair pathway.

Within the Dry Valley metagenomes, other nuclease N-terminally fused ligase proteins were identified that shared 60% sequence homology, with a majority of these proteins been identified by IMG to originate from *Candidatus Udaeobacter copiosus* species. Searching publicly available datasets outside of the Dry Valleys has revealed other proteins with this nuclease ligase fusion, with top hits found in species of *Chthoniobacter flavus*, *C. Udaeobacter copiosus*, *Verrucomicrobia* bacterium verI-A, *Nibricoccus aquaticus* and *O. terrae.* A search was conducted to compare the synteny and gene conservation of genes surrounding the Dry Valley ligase-nuclease with homologous domain-fused enzymes to determine whether they are part of conserved gene clusters. Comparison of contigs that contained flanking genes indicated that none of the Dry Valley ligase-nuclease genes were present in the LigB/Lhr helicase/ phosphodiesterase/ metallo-beta-lactamase configuration observed in the *O. terrae* genome, and there was little overall synteny between clusters. However, there was synteny observed between DV contigs for several genes involved in protein expression, such as sigma factors, transcription factors, ribosomal subunit proteins as well as enzymes involved in tRNA or rRNA modification. None of these regions were predicted to be complete or partial phage.

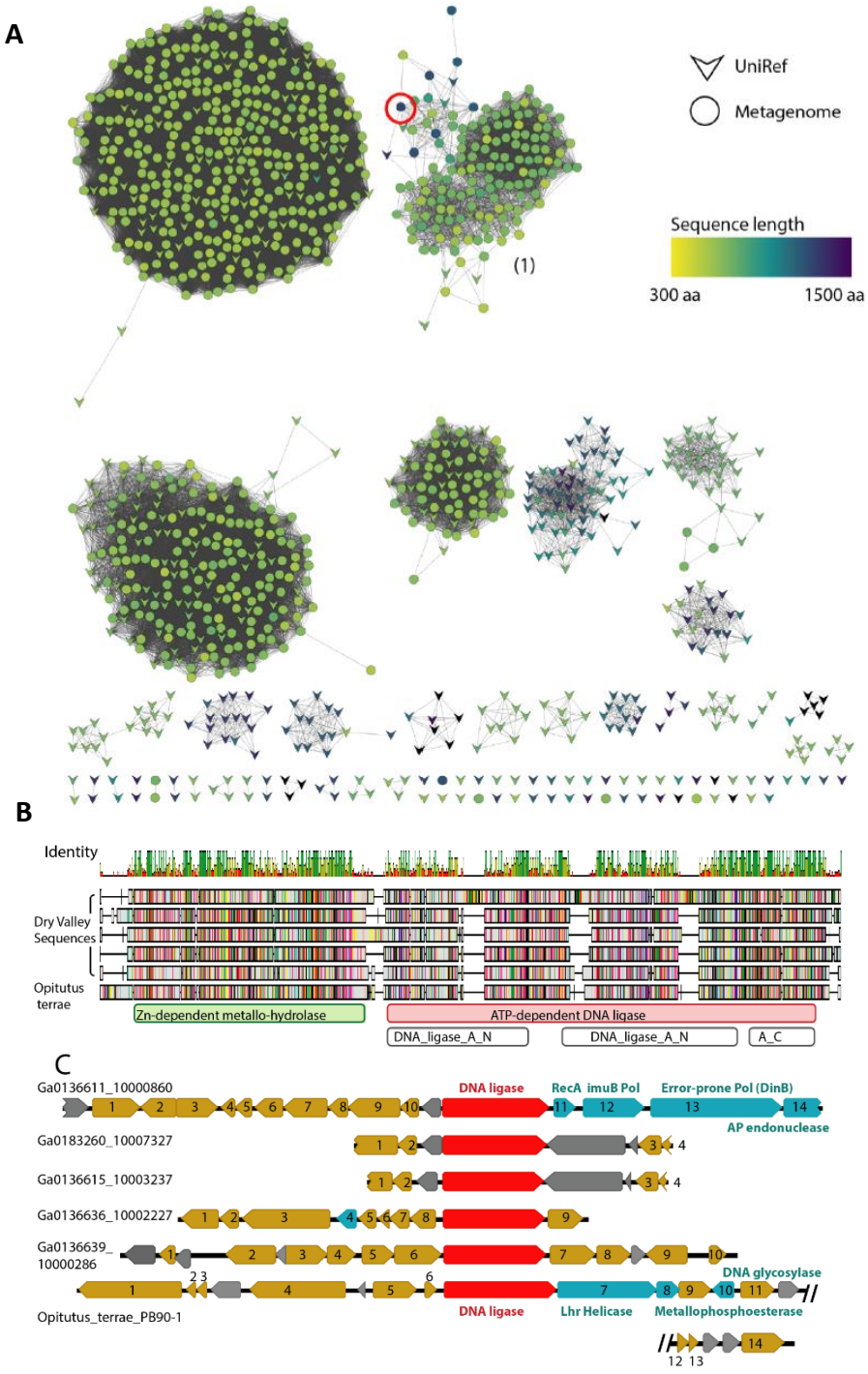


Figure 4.2. Identification of ligase nuclease fusion proteins using sequence similarity networks (SSNs) Sequence similarity networks were constructed for each set of sequences identified by hmmsearch using the EFI-EST server. **A**) SSN of metagenome hits to LigB-type DNA ligases at 50 % identity edge threshold; other network parameters are detailed in **Appendix B.1.** Domain compositions include the catalytic DNA_ligase_A_M domain together with the N-terminal DNA binding domain DNA_ligase_A_N. Nodes are coloured by sequence length. **B**) Sequence alignment of full-length DV-metagenome sequences from Cluster #1 where an N-terminal metallonuclease (RMMBL) domain was detected with hmmscan Sequences are aligned to the *Opititus terrae* DNA ligase gene OTER_RS15935 from the genome sequence NC_010571_Opitutus_terrae_PB90-1. Domain boundaries from Pfam/Interpro sequences searches are shown below the alignment. **C**) Genomic context of metallonuclease-ligase fusion proteins from DV-metagenomes and *O. terrae*. The nuclease-ligase gene is shown in red, putative DNA-repair genes in cyan, other annotated genes in gold and hypothetical proteins in gray. Figure is sourced from (Rzoska-Smith et al., 2023).

4.2.2 Structural characterisation of DV-1-1-Lig-Nuc protein

As there are no current crystal structures of this nuclease ligase fusion protein structural modelling was used to gain insight into the interactions of this protein. The polypeptide sequence of the Dry Valley nuclease ligase fusion protein (Ga0136611_1000086013), hence forth known as DV-1-1-Lig-Nuc, was used in the Google Colab AlphaFold2_Advanced structural prediction software (John Jumper, 2021; Varadi et al., 2022) to generate a 3D structural prediction, as described in **Section 2.1**. This version of AlphaFold was used, as the polypeptide sequence was too big to put through Google Colab AlphaFold2. The 3D structures for the separate DV-1-1-Nuc domain and DV-1-1-Lig domain were also predicted, this time through Google Colab AlphaFold2.

The predicted local distance difference test (pLDDT) for the DV-1-1-Nuc domain, gives a high confidence score for each amino acid, with only a small region of low confidence in the C-terminal region of the protein (**Appendix C.1**). For both DV-1-1-Lig-Nuc (fusion protein) and DV-1-1-Lig proteins, most amino acids were modelled with high confidence, however the interdomain linker had a particularly low pLDDT score which is consistent with the predicted flexibility of this region.

To get an overview of the placement of secondary structural elements and compare these with DNA ligases of known structure, topology maps were generated for the AlphaFold models. Consistent with the structural and sequence alignments, these indicate that secondary structural elements are positioned equivalently with known DNA ligase structures (**Figure 4.3**). The topology map generated for DV-1-1-Lig shows an arrangement of secondary structural elements as expected for Lig B type DNA ligases. The DBD and OBD are in close proximity to each other as seen with DV-Lig2 in **Section 3.2.1**, **Figure 3.26**. The topology map generated for DV-1-1-Nuc shows an arrangement of secondary structural elements that is also observed in homologous proteins such as hSNM1A, where the MBL and β-CASP domains, are connected through a series of linker regions (Baddock et al., 2020).

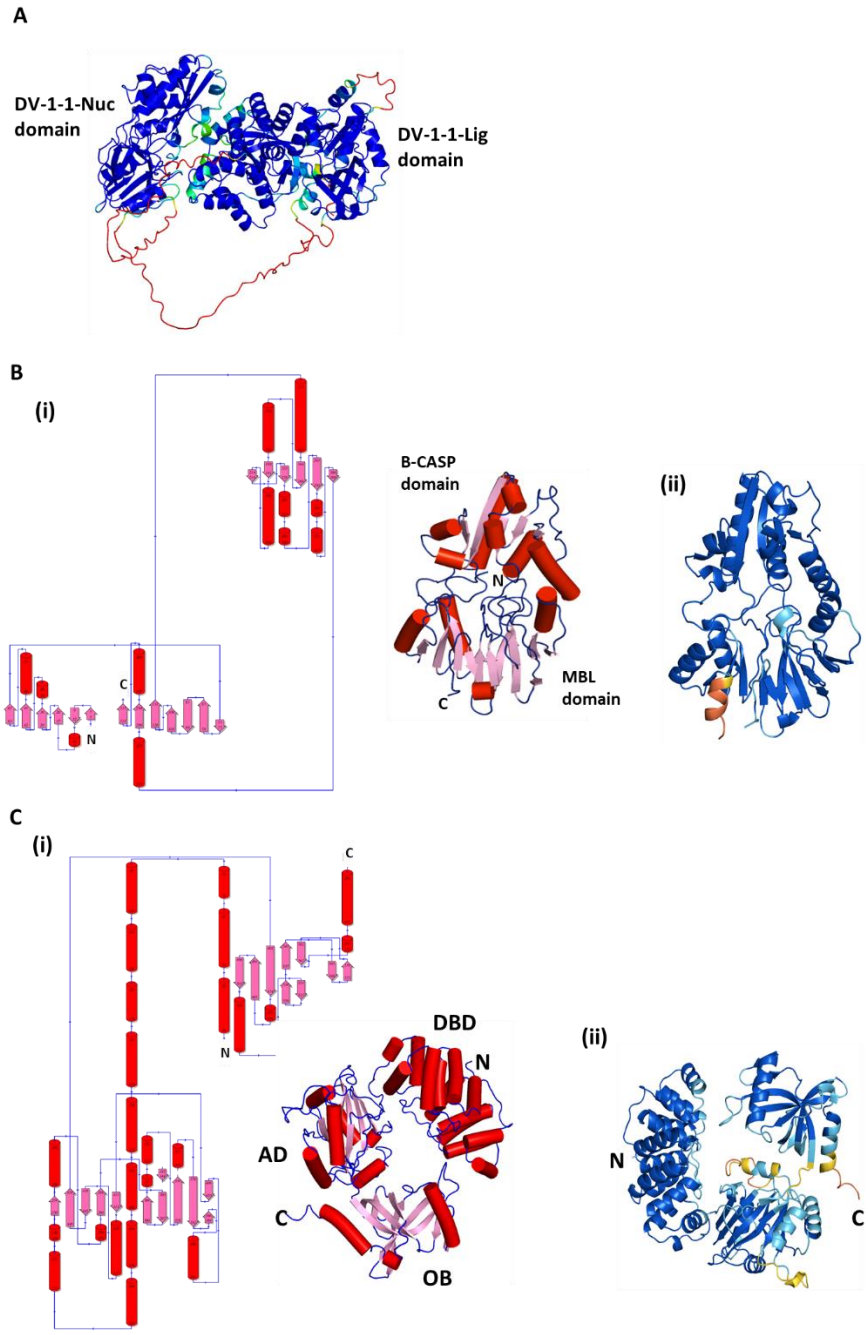


Figure 4.3. Structural arrangements of DV-1-1-Lig-Nuc, DV-1-1-Nuc and DV-1-1-Lig. **A**) AlphaFold structural prediction of DV-1-1-Lig-Nuc. **B**) Topology map and AlphaFold predicted structure of DV-1-1-Nuc. **C**) Topology map and AlphaFold predicted structure of DV-1-1-Lig. Secondary structural elements are colour coded in reference to secondary structure, helices in red, strands in pink and loops in blue. AlphaFold models in (ii) are coloured based on pLDDT confidence, with high confidence residues coloured in blue and lower confidence in yellow, orange and red. Protein topology maps were generated in PDBsum (Laskowski, 2022) using AlphaFold predicted structural models for DV-1-1-Lig and DV-1-1-Nuc (John Jumper, 2021; Varadi et al., 2022).

As mentioned above in **Section 4.2.1**, DV-1-1-Lig-Nuc, annotated as an ATP-dependent DNA ligase, is found within a gene cluster with other DNA modifying proteins. DV-1-1-Lig-Nuc is made up of a N-terminal nuclease domain, fused to a C-terminal ligase domain. These domains are connected by a

45-polypeptide linker. The predicted structure of the nuclease domain contains a β-CASP domain, sandwiched between a MBL domain. The predicted structure of the ligase domain shows the canonical Lig B type arrangement of three domains; an N-terminus DB domain, followed by an AD domain and an OB domain at the C-terminus (**Figure 4.4**).

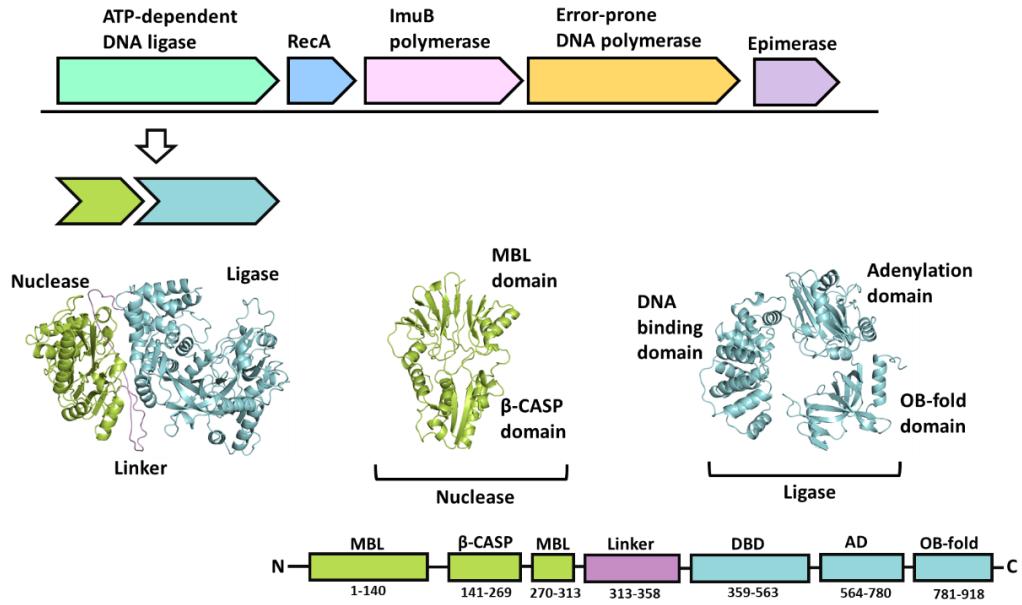


Figure 4.4. DV-1-1-Lig-Nuc (ATP-dependent DNA ligase) is in a gene cluster with genes encoding for DNA modifying proteins (RecA, ImuB polymerase, error-prone polymerase, and an epimerase). DV-1-1-Lig-Nuc is made up of an N-terminus nuclease domain (green), connected to C-terminus ligase domain (blue), by a polypeptide linker (purple). Predicted structural models were generated by AlphaFold2_Advanced and AlphaFold2 from Google Colab, version v2.3.1 (Jumper, Evans et al. 2021). Models were presented using PyMOL (Schrödinger, 2020).

Predicted structural analysis of this ligase-nuclease fusion protein (DV-1-1-Lig-Nuc), shows the nuclease domain in close proximity to the ligase domain. α-helices from the β-CASP domain, are contacting the DNA binding (DB) domain and OB (oligonucleotide/oligosaccharide-binding) fold domain, of the ligase component. An analysis of potential interactions between the ligase and nuclease domain was conducted in Pymol, through a search of polar contact between domains (**Figure 4.5**). Several polar contacts are made between the ligase and nuclease domains. From the MBL domain residue Arg-47 is forming a polar contact with residue Gln-890 from the OB domain. In the β-CASP domain residue Pro-295 is forming a polar contact residue Tyr-485 (DB domain) and residues Arg-273 and Lys-275 are forming polar contacts to residues His-844, Lys-841 and Asp-874 (OB domain). There is also contact between the side chain of residue

Ala-353 (MBL domain) and residue Asn-358 from the linker. The protein contact potential was generated for the predicted structure of DV-1-1-Lig-Nuc, which predicts regions of positive charge on the surface of the nuclease and ligase domains.

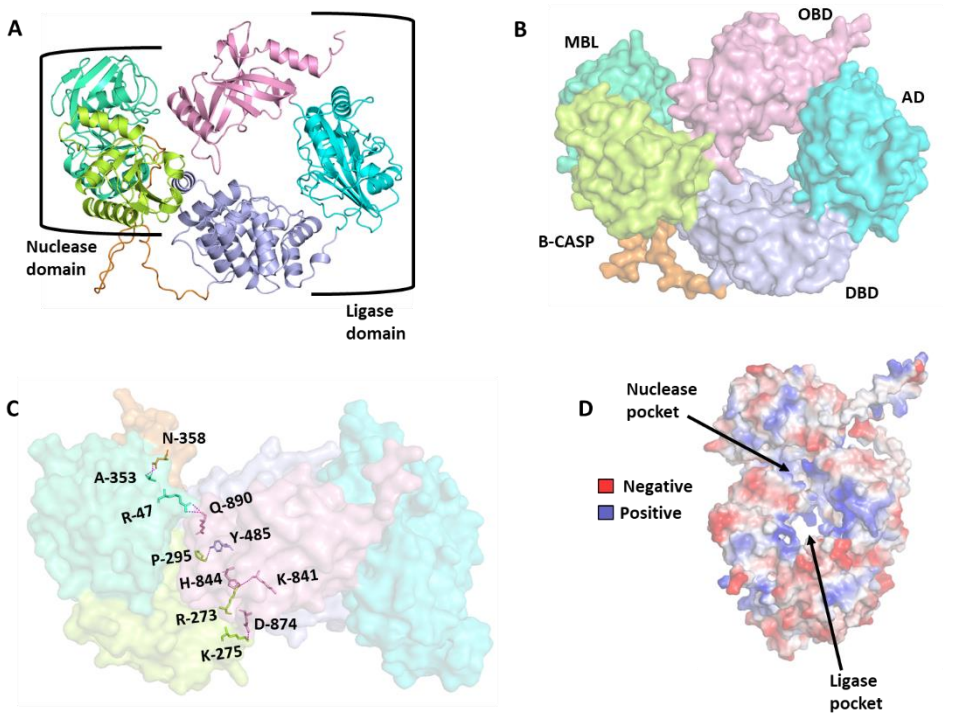


Figure 4.5. Characteristics of DV-1-1-Lig-Nuc structural model prediction. **A**, **B**) Arrangement of domains making up the ligase and nuclease domain. The nuclease domain consists of a MBL domain (dark green) and a β-CASP domain (light green). The ligase domain is made up of a DBD (purple), an AD (blue) and an OBD (pink). **C**) Polar interactions between the ligase, and nuclease domains, and the linker domain (orange). **D**) the electrostatic surface potential of DV-1-1-Lig-Nuc, where red is more electronegative, and blue more electropositive. The predicted model for DV-1-1-Lig-Nuc was generated by AlphaFold2_Advanced, from Google Colab, version v2.3.1 (John Jumper, 2021; Varadi et al., 2022) and all structures were presented in PyMOL (Schrödinger, 2020).

Currently, no crystal structures of this ligase-nuclease fusion protein have been solved, however there are AlphaFold models which show similar protein fusions from *C. flavus Ellin428*, *T. sacchariphilum*, *N. aquaticus* and *O. terrae*. The structure of DV-1-1-Lig-Nuc was superimposed onto the predicted structures of these fusion proteins. All proteins overlayed with a reasonable degree of similarity, with RMSD values below 3. The individual domains of DV-1-1-Lig-Nuc and *O. terrae*-Lig-Nuc were superimposed separately to account for changes in orientation around the linker since this is expected to be flexible in solution. The nuclease domains are structurally very similar, with an RMSD value of 0.677. The ligase domains also showed high structural similarity, with small differences

in the positioning and length of linkers between the DB, AD (adenylation) and OB domains. The DB, AD and OB domains of DV-1-1-Lig were superimposed onto the matching domains from *O. terrae*-Lig. RMSD values were generated for each separate overlay; OB-0.366, AD-2.787 and DB-0.531 (**Figure 4.6**). Looking at sequence alignments between these protein homologs and DV-1-1-Lig-Nuc, there is high sequence similarity, with many conserved domains in red (**Appendix C.2**) The Lig-Nuc protein from *O. terrae*, is less conserved with DV-1-1-Lig-Nuc, compared to the other proteins.

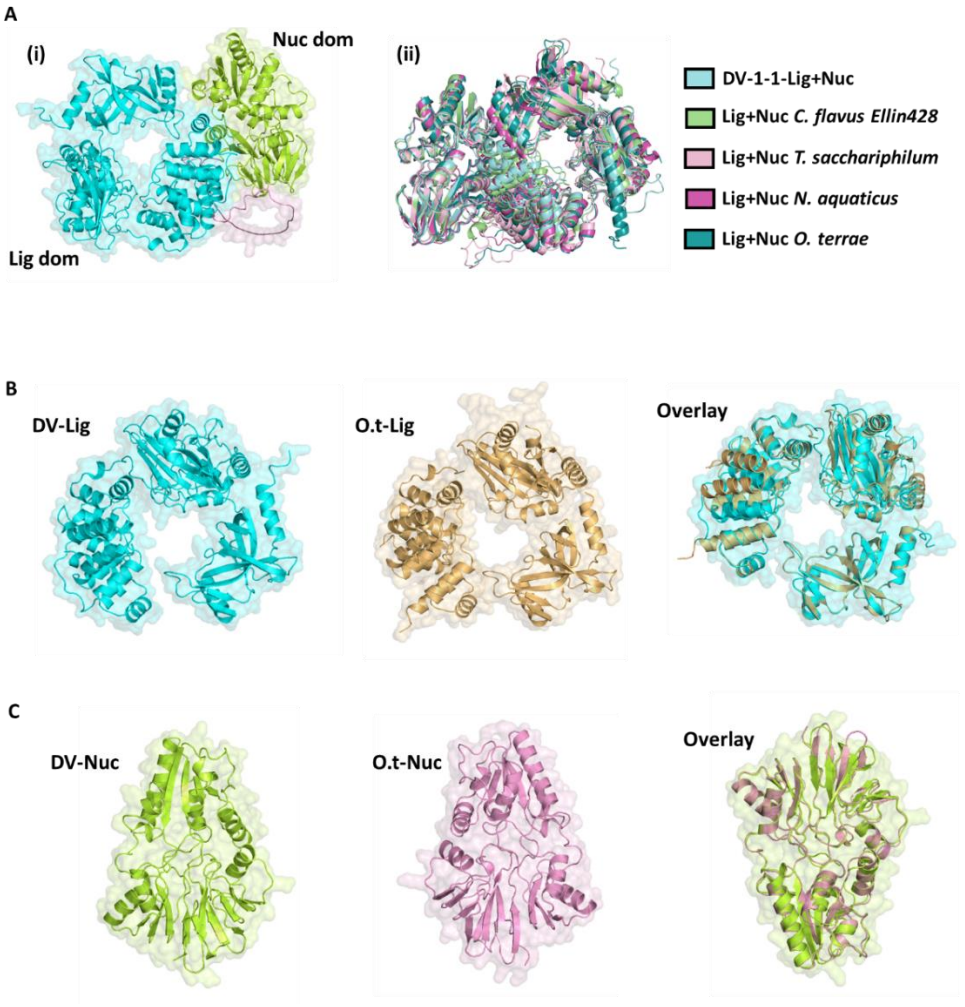


Figure 4.6. Structural alignments of ligase-nuclease fusion proteins with DV-1-1-Lig-Nuc. **A**) (i) DV-1-1-Lig-Nuc fusion protein, with ligase domain shown in blue, linker in pink and nuclease domain in green. (ii) Predicted structures of ligase-nuclease fusion proteins from, C. flavus Ellin428, T. saccharophilum, N. aquaticus and O. terrae, overlayed onto the structure of DV-1-1-Lig-Nuc. **B**) The ligase domains from DV-1-1-Lig-Nuc (blue) and *O. terrae*-Lig-Nuc (orange), overlayed onto each other. (RMSD: OB: 0.366, AD: 2.787, BDB: 0.531). **C**) The nuclease domains from DV-1-1-Lig-Nuc (green) and *O. terrae*-Lig-Nuc (pink), overlayed onto each other. (RMSD: 0.677). DV-1-1-Lig-Nuc predicted model was generated by AlphaFold2_Advanced, from Google Colab, version v2.3.1 Other predicted models were sourced from AlphaFold database (John Jumper, 2021; Varadi et al., 2022) and all structures were presented in PyMOL (Schrödinger, 2020).

As described above, the nuclease domain (DV-1-1-Nuc) is made up of a MBL domain and a β-CASP domain. The MBL fold is identified by a four-layered αβ/βα structure, which features a broad and shallow active site located consistently on one side of the fold (**Figure 4.7**, **A**, **B**). The generation of electrostatic surface potential for DV-1-1-Nuc shows a cluster of electronegative residues on the surface of the active site pocket. The β-CASP domain consists of a five-stranded parallel β-sheet, flanked by α-helices either side and is inserted between strands of the MBL domain (**Figure 4.7**, **C**). There are several polar interactions between the MBL domain and β-CASP domain, with interactions from residue Glu177 in the β-CASP domain forming a salt bridge with residues Lys110 and Arg-112 from the MBL domain (**Figure 4.7**, **D**).

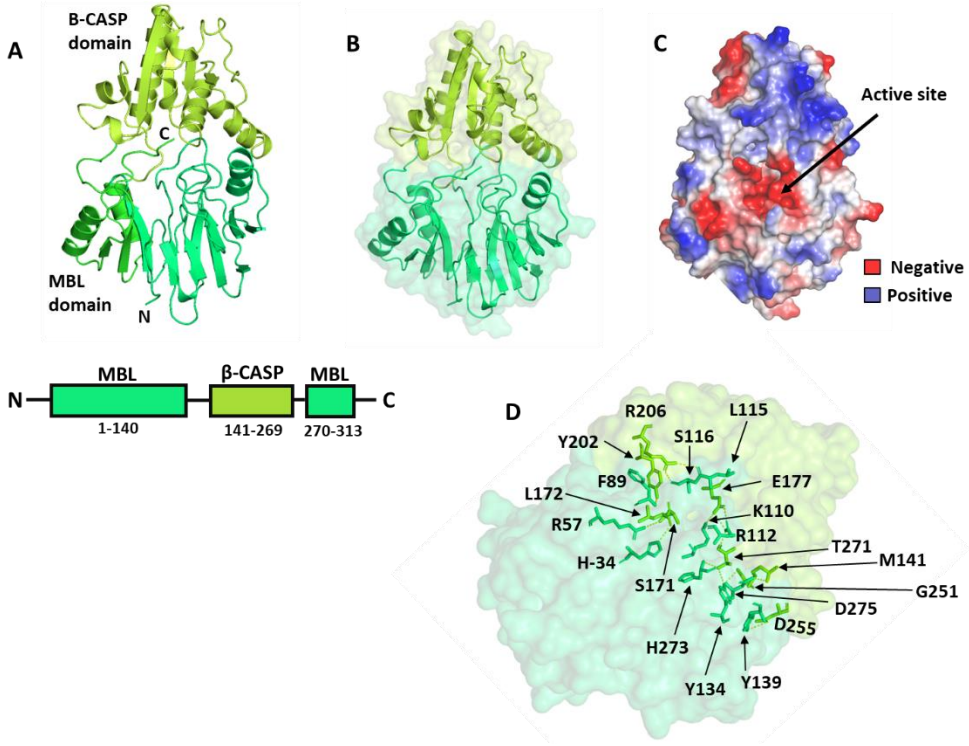


Figure 4.7. Characteristics of DV-1-1-Nuc structural model prediction. **A**, **B**) Arrangement of domains making up the nuclease domain. The nuclease domain consists of a MBL domain (dark green) and a β-CASP domain (light green). **C**) the electrostatic surface potential of DV-1-1-Nuc, where red is more electronegative, and blue more electropositive. **D**) Polar interactions between residues from the MBL and β-CASP domains. The predicted model for DV-1-1-Nuc was generated by AlphaFold2, from Google Colab, version v2.3.1 (John Jumper, 2021; Varadi et al., 2022) and all structures were presented in PyMOL (Schrödinger, 2020).

The overall fold of the catalytic core of DV-1-1-Nuc is similar to that of other MBL-β-CASP nucleases, such as SNM1B, SNM1C and SNM1A (Yosaatmadja et al., 2021). Structural overlays between DV-1-1-Nuc and SNM1B, SNM1C and SNM1A, show a comparable arrangement of secondary

structural elements, all giving RMSD values below 3 (**Figure 4.8**, **A**). DV-1-1-Nuc has all the key structural characteristics of MBL fold nucleases, with a metal binding active site at the interface between the MBL and β-CASP domains. DV-1-1-Nuc was super imposed onto the crystal structure of SNM1C (7AF1), which was solved with two zinc ions bound to the active site. The same residues involved in zinc coordination, from SNM1C are also observed in DV-1-1-Nuc, which form polar contacts to both zinc ions. In DV-1-1-Nuc four residues His32, His34, His37 and Asp108 show polar interactions with the first zinc ion M1 and three residues Asp36, His37 and Asp108 show polar interactions with the second zinc ion M2 (**Figure 4.8**, **B**).

A polypeptide sequence alignment was generated for DV-1-1-Nuc and other typical MBL-β-CASP nucleases (**Appendix C.2**). Here there is low sequence conservation between proteins. However, despite the low sequence identity between members, all proteins contain four sequence motifs conserved in the MBL family. Motif II comprises the well characterized HxHxDH sequence motif, that is nearly absolutely conserved among all the MBLs, and where the first His and Asp residues are completely invariable. The β-CASP domain is characterised by three motifs; motif A, consisting of an acidic residue (D or E), motif B (His) and motif C which is a Val residue in SNM1A, SNM1B and SNM1C and a His in CSP-73 and DV-1-1-Nuc (Dominski, 2007; Fernandez et al., 2011).

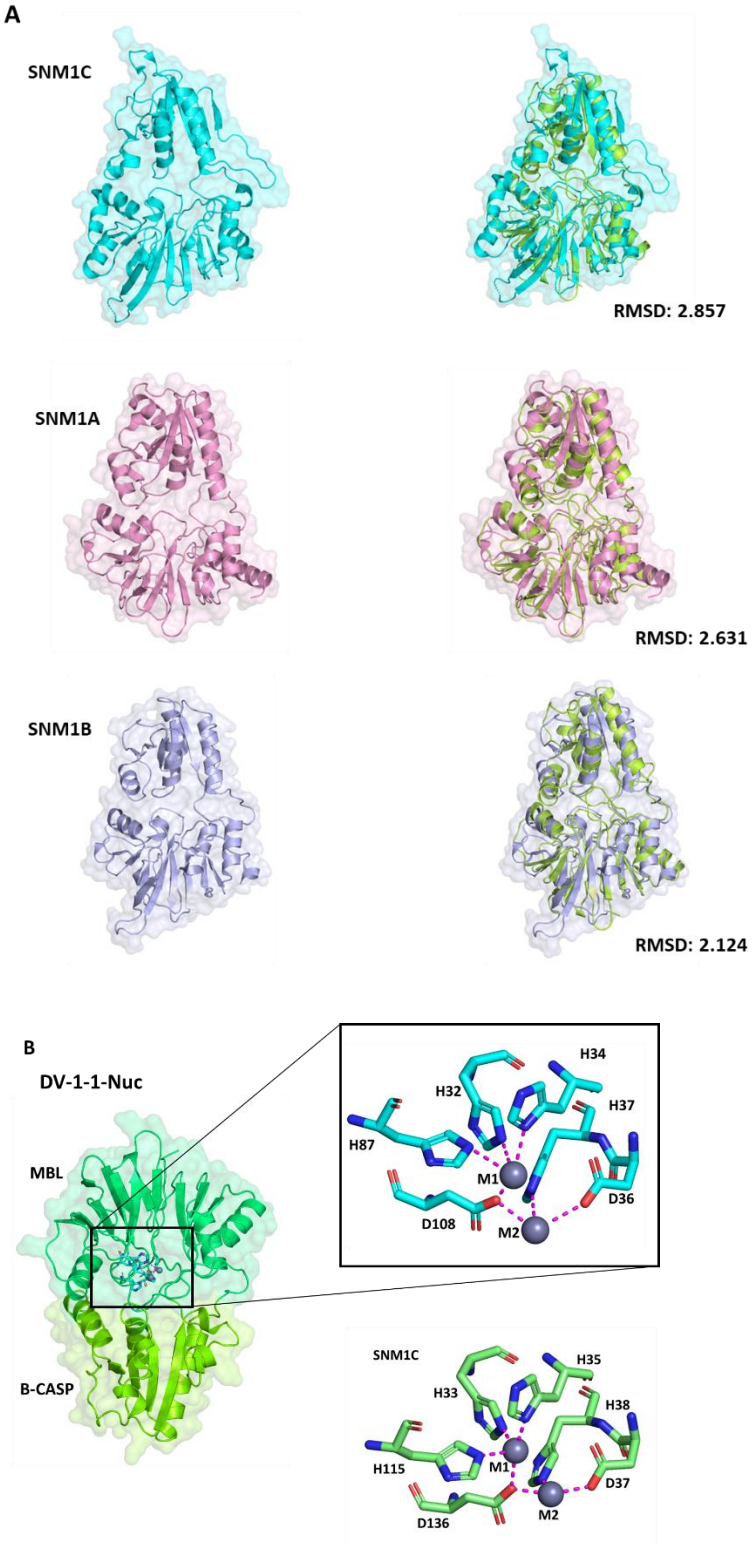


Figure 4.8. Structural comparison of DV-1-1-Nuc to other MBL-β-CASP nucleases. **A**) Crystal structures of SNM1C (7AF1), SNM1A (5AHR) and SNM1B (7A1F) superimposed onto DV-1-1-Nuc structural prediction. RMSD values included in figure for each structural overlay. **B**) DV-1-1-Nuc superimposed onto the structure of SNM1C, bound to two zinc ions in the active site. DV-1-1-Nuc make 6 polar interactions with the zinc ions and these are the same residue interactions seen with SNM1C. DV-1-1-Nuc predicted model was generated by AlphaFold2, from Google Colab, version v2.3.1 (John Jumper, 2021; Varadi et al., 2022) Other structural models were sourced from PDB. PDB identifiers are included above, for each structure. All structures were presented in PyMOL (Schrödinger, 2020).

The ligase domain is connected to the C-terminus of the nuclease ligase fusion protein. Analysis of its structure and sequence has revealed it to be a LigB type ATP-dependent DNA ligase. Details on LigB type DNA ligases have been previously discussed in **Section 3.2.1**. As of typical LigB type ligases, the protein is made up of three main domains (Williamson et al., 2016). Domain I (residues 359-563), at the N-terminus of the ligase domain, forms a triangular configuration, representing the DNA binding domain (DBD). Domain II (residues 564-780) is the central nucleotidyltransferase (NTase) adenylation (AD) domain, which contains the KxDG motif (motif I) (**Appendix C.2**), including a Lys residue (Lys 610) for adenylation. Domain III (residues 781-918) is at the C-terminus of the protein, consists of the oligonucleotide-binding fold (OB-fold), with the addition of two α-helices and is classified as the oligonucleotide/oligosaccharide binding (OB)-fold domain (**Figure 4.9, A**) (Nishida et al., 2006).

Interactions between domains of DV-1-1-Lig were investigated in PyMol. Here there are several interactions, with salt bridges forming between the DB and AD domains and the AD and OB domains. With residue Asp-211 from the DB domain forming a polar contact with Arg-449 from the AD domain, and residue Glu-422 from the AD domain forming a polar contact with Lys-495 from the OB domain. There are also mutliple interactions between the DB and OB domains, none of these interactions form salt bridges. Here residues Ala-122, Tyr-121 and Arg-119 from the DB domain are forming polar contacts with residues Gly-476, Ser-479 and Gln-473 from the OB domain. The generation of electrostatic surface potential for DV-1-1-Lig shows a region of electropositive surface residues in the centre of the three domains (**Figure 4.9**, **A**).

DV-1-1-Lig was super imposed onto other homologous DNA ligases, that contain a DB, an AD and an OB domain from; *A. fulgidus*, *P. furiosus*, *T. sibiricus* and *S. solfataricus*. The individual domains from each DNA ligase were also super imposed separately onto the equialevent domains of DV-1-1-Lig. These overlays show the structural similarity of domains between these different proteins (**Figure 4.9, B**). The structure of Human DNA ligase I (1X9N) was also super imposed onto DV-1-1-Lig, but did not overlay as well as the archael DNA

ligases, with structural differences observed between the DB domains (data not shown).

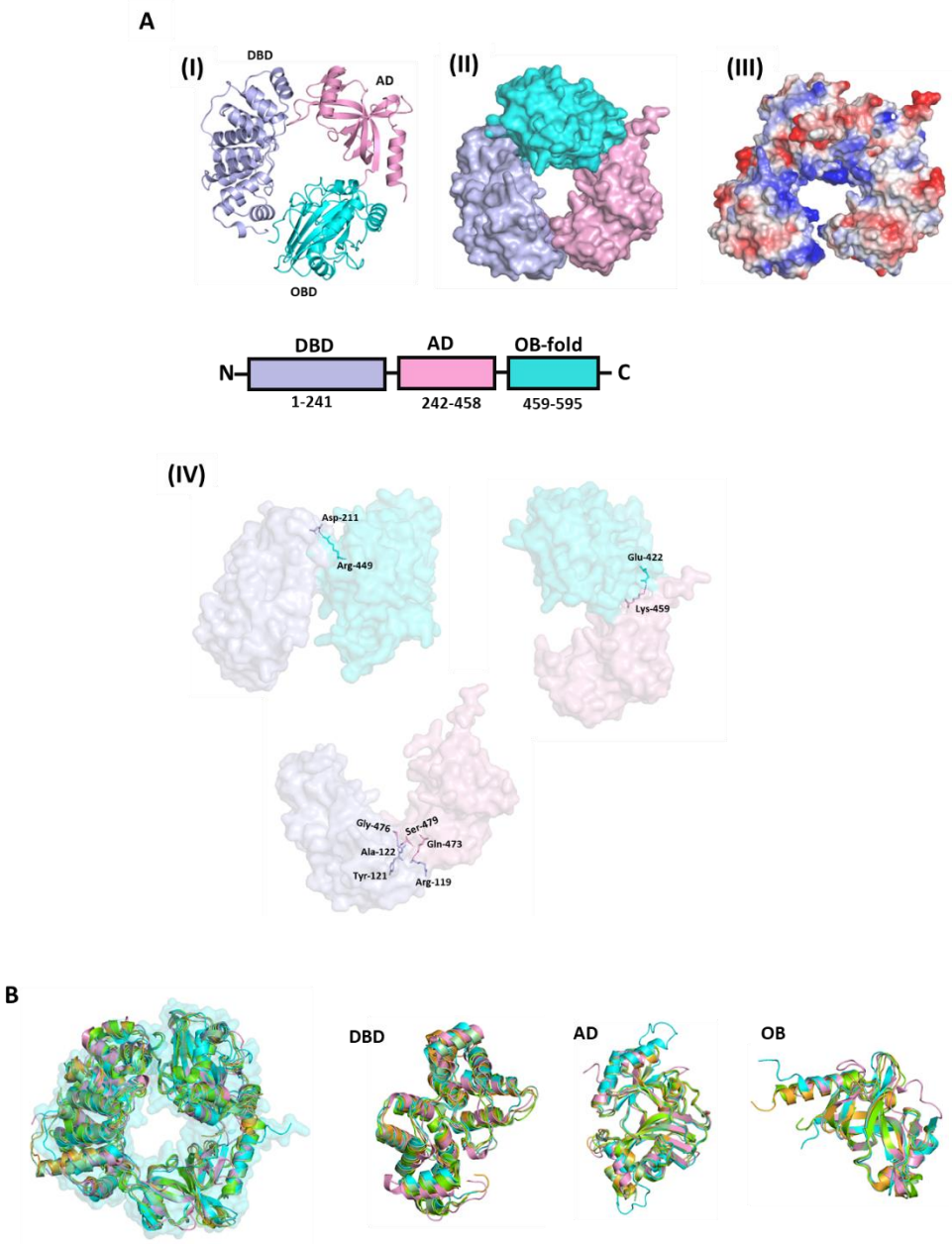


Figure 4.9. Structural arrangement of DV-1-1-Lig domain and comparison with homologous DNA ligases. **A**) (I) cartoon representation of domain making up the ligase domain. DBD in purple, AD in blue and the OBD in pink. (II) the ligase domain represented as a surface. (III) the electrostatic surface potential of DV-1-1-Lig, where red is more electronegative, and blue more electropositive. (IV) Polar interactions between residues from the DBD, AD and OBD. **B**) Structural comparison of DV-1-1-Lig to other homologous DNA ligases (*A. fulgidus* (3GDE), *P. furiosus* (2CFM), *T. sibiricus* (4EQ5), and *S. solfataricus* (2HIV). Here ligases were super imposed onto DV-1-1-Lig as a whole protein or as separate domains. RMSD values from overlays were below 2.5. The predicted model for DV-1-1-Lig was generated by AlphaFold2, from Google Colab, version v2.3.1 (John Jumper, 2021; Varadi et al., 2022). Other structural models were sourced from PDB. PDB identifiers are included above, for each structure. All structures were presented in PyMOL (Schrödinger, 2020).

To model the possible DNA-bound conformation, DV-1-1-Lig was super imposed onto the crystal structure of h-LigI, which is bound to a nicked DNA

duplex (1X9N). Interactions between domains of DV-1-1-Lig and this DNA duplex were investigated in PyMOL (**Figure 4.10**). The extensive DNA contacts made by DV-1-1-Lig domain involve a total of 31 amino acids residues from the three structural domains. The N-terminal DB domain of DV-1-1-Lig is a bundle of 13 α-helices and it is predicted to engage two regions of the DNA minor groove separated by one turn of the helix. A total of nine residues from the DNA binding domain could form polar contacts to the DNA duplex. Lys-194 forms polar contacts with a cytosine base (DC-12), while all other residues are forming polar contacts to the phosphate backbone. The AD and OB domains constitute the catalytic core of DNA ligases. The AD domain comprises two lobes, consisting of a layer of β-sheet flanked by α-helices and the cleft formed between the two lobes, serves as the AMP binding site. The AMP phosphate group in the active site, is within hydrogen binding distance from four lysine residues; Lys-286 (motif I), Lys-439, Lys-458 and Lys-456. Lys-286 is the catalytic residue that gets adenylated in the ligase-AMP intermediate. The adjacent DNA 5'-phospahte interacts with Arg-306. Other interactions with the AMP groups are formed by polar contacts from residues; Glu-284, Asp-285, Phe-287 and Glu-334. Residues Arg-263, Arg-306, Thr-305 and Lys-309 form polar contacts with the DNA phosphate backbone, while residue Lys-356 forms polar contacts with the ribose sugar on an adenine nucleobase. The OB domain is a barrel of 7 β-strands, capped on either end by two short α-helices. The β-barrel with an oblong shape, fits in the minor groove of DNA, making extensive contacts into the groove. The OB domain shows a 2-fold symmetrical overall fold, mirroring the dyad of the bound DNA. Nine residues from the OB domain are interacting with the DNA duplex. Residue Ser-505 is forming polar contacts with the ribose sugar of a cytosine nucleobase (DC-3). Tyr-504 is forming three polar contacts to two guanine (DG1, DG-16) and one cytosine (DC-15) nucleobases. Arg-551 is forming polar contacts to a guanine nucleobase (DG-19). All other residues are forming polar contacts with the phosphate backbone of the DNA.

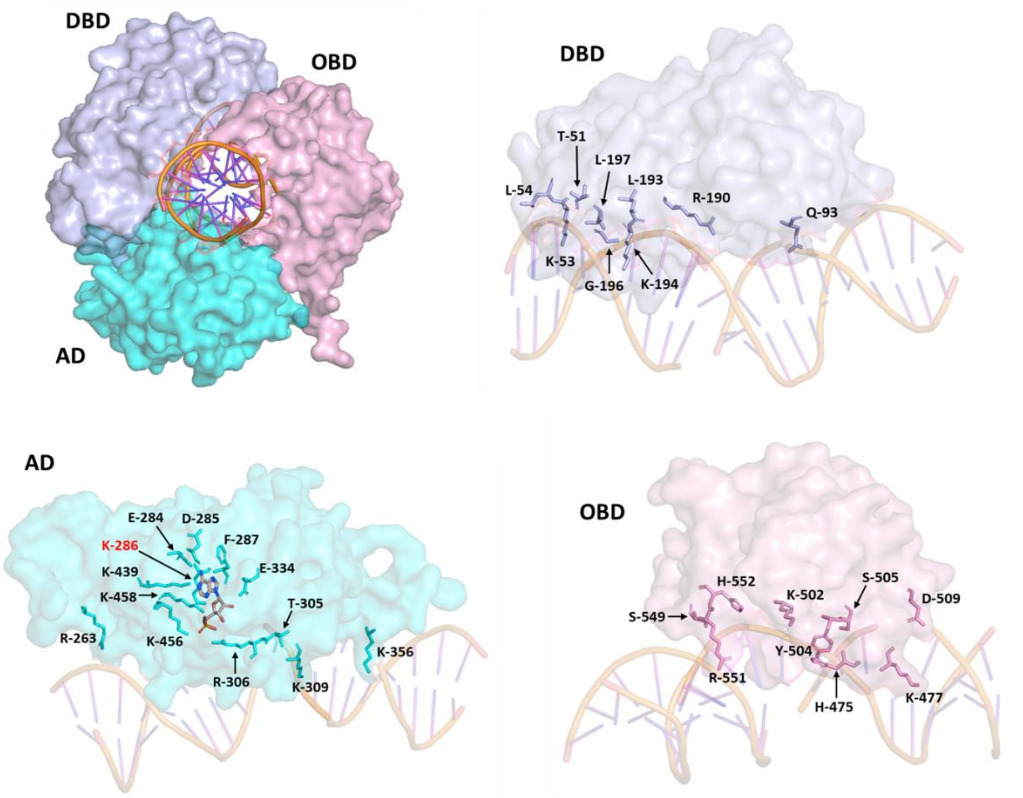


Figure 4.10. Structural arrangement of DV-1-1-Lig around a DNA duplex from h-LigI (1X9N). DV-1-1-Lig was super imposed onto 1X9N, which was solved bound to a nicked DNA duplex. Polar contacts from residues in the DBD (purple), AD (blue) and OBD (pink) were investigated in PyMOL. The catalytic lysine residue from the AD is indicated by red text (K-285). The predicted structure for DV-1-1-Lig domain protein was generated by AlphaFold2, from Google Colab, version v2.3.1 (Jumper, Evans et al. 2021). Models were presented using PyMOL (Schrödinger, 2020).

4.2.3 Preliminary small-scale expression of DV gene cluster proteins

Before I joined this project, preliminary protein expression trials of proteins belonging to the gene cluster in **Section 4.2.1**, were conducted to test expression and solubility of these proteins on a small scale (50 ml). Proteins were cloned into pHMGWA and pDEST17 plasmids and expressed in pLysS BL21 (DE3) *E. coli* cells, at 25 ºC. DV-1-1-Ligase (Hence forth called DV-1-1-Lig-Nuc) protein showed a small amount of soluble protein expression in the pHMGWA vector. DV-1-2-RecA protein did not have any soluble protein expression, at the correct size. While DV-1-3-Polymerase showed a small amount of soluble protein expression in pHMGWA vector (**Figure 4.11**).

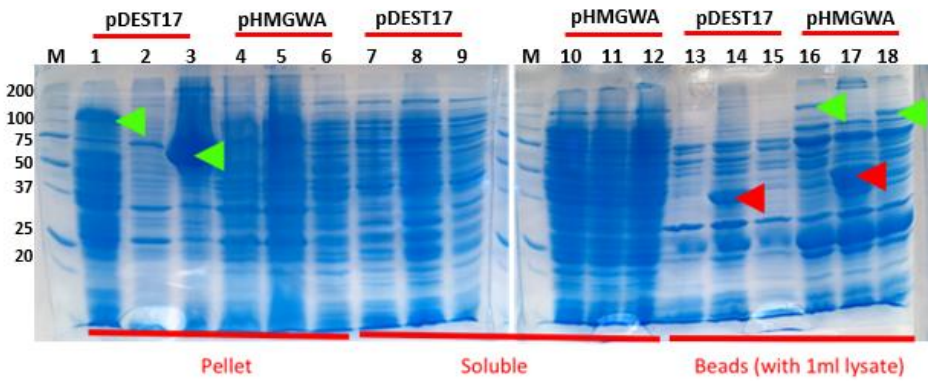


Figure 4.11. SDS PAGE showing results of a small-scale protein expression trials of proteins from the DV-1-4 gene cluster of the Dry Valley metagenome (metaG UQ223). DV-1-1-Lig-Nuc protein (lanes: 1, 4, 7,10, 13 and 16) is 109 kDa in pDEST17 plasmid and 150 kDa in pHMGWA plasmid. DV-1-2-RecA protein (lanes: 2, 5, 8, 11, 14 and 17) is 25.6 kDa in pDEST17 plasmid and 66.1 kDa in pHMGWA plasmid. DV-1-3-Polymerase protein (lanes: 3, 6, 9, 12, 15, and 18) is 64.9 kDa in pDEST17 plasmid and 105 kDa in pHMGWA plasmid. A precision plus 250 kDa protein marker, was used. Green markers indicate proteins of the correct size, while the red markers indicate protein expression, but at the wrong protein size. Gels were stained with Blue Safe protein stain (GelCode™).

When I joined this project, I examined the small-scale expression results of these proteins from the DV-gene cluster and decided to proceed further with DV-1-1-Lig-Nuc and DV-1-3-Polymerase proteins. DV-1-3-Polymerase protein went through a large-scale (1 L) purification and resulted in some soluble protein expression, but this protein precipitated when up concentration was attempted (data not shown). Further attempts at purification gave similar results, and no further experiments were conducted on this protein.

4.2.4 Recombinant production of DV-1-1-Lig-Nuc

Following on from expression of soluble protein in small scale expression trials, as described above, large scale growth and purifications were attempted with DV-1-1-Lig-Nuc protein, as described in **Sections 2.3.3** and **2.4**. IMAC purification of DV-1-1-Lig-Nuc expressed from BL21 pLysS cells showed some enrichment of the target protein in the eluted fractions, but overall expression yields were low (**Figure 4.1.2**).

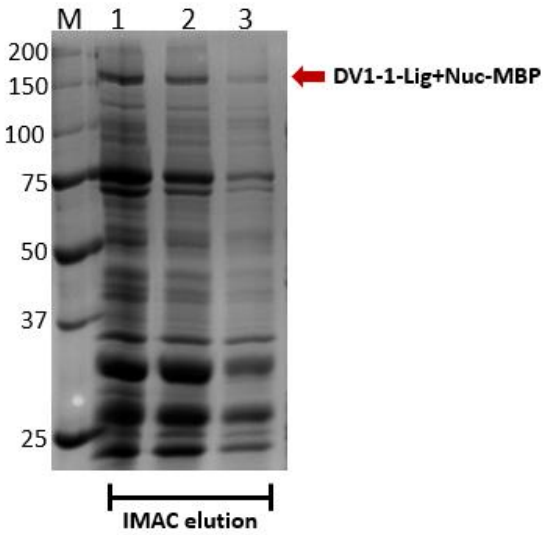


Figure 4.12. Elution peak fractions from an IMAC purification of DV1-1-Lig-Nuc protein (original start sequence), shows low protein expression and contaminating *E. coli* proteins. Lane M contains a precision plus protein standard. Lanes 1, 2 & 3 are fractions from the IMAC elution peak. DV1-1-Lig-Nuc protein is 150 kDa in size (indicated by red arrow).

Further growth expression trials were attempted with DV-1-1-Lig-Nuc protein in BL21 pLysS cells, to try and improve overall protein yield. Here different concentrations of IPTG, growth temperature (15, 20, 30 ºC), additions of 2 % glucose, 1 mM ATP nucleotide and 10 mM magnesium metal ions were introduced separately during the expression of DV-1-1-Lig-Nuc protein, in BL21 pLysS cells. Results from these trials saw no improvement in the yield of DV-1-1-Lig-Nuc protein (data not shown) and no further purifications of this full length DV-1-1-Lig-Nuc construct were attempted.

To address the challenges associated with producing the full length DV-1-1-Lig-Nuc protein, constructs were designed to isolate and express the ligase and nuclease domains as individual smaller proteins. By separating these domains and expressing them independently, it was anticipated that protein expression would improve, because of the reduced protein size.

4.2.5 Construct design for splitting DV-1-1-Lig-Nuc

Constructs for each domain were based on domain arrangements seen from AlphaFold predicted models. Each new construct was designed to keep the original N-terminus for the nuclease domain and the C-terminus for the ligase domain, with modifications to remove sections of the linker region from each

domain (**Figure 4.13)** The original DV1-1-Lig-Nuc plasmid in pDONR221 was used as a template for each new domain construct, with primers designed to amplify each domain separately. Several PCR steps, followed by Gateway BP and LR reactions, as described in **Section 2.2.2**, resulted in each domain construct being successfully cloned into pHMGWA and pDEST17 expression plasmids. PCR using T7 and MBP forward primers, as described in **Section 2.2.3.2** indicated the correct size for each construct (**Figure 4.13**, **B**) which was then confirmed by sequencing. Following confirmation of correct sequences, DV-1-1-Lig and DV-1-1-Nuc proteins were expressed in *E. coli* expression strains and small-scale expression trials were used to check for soluble protein expression.

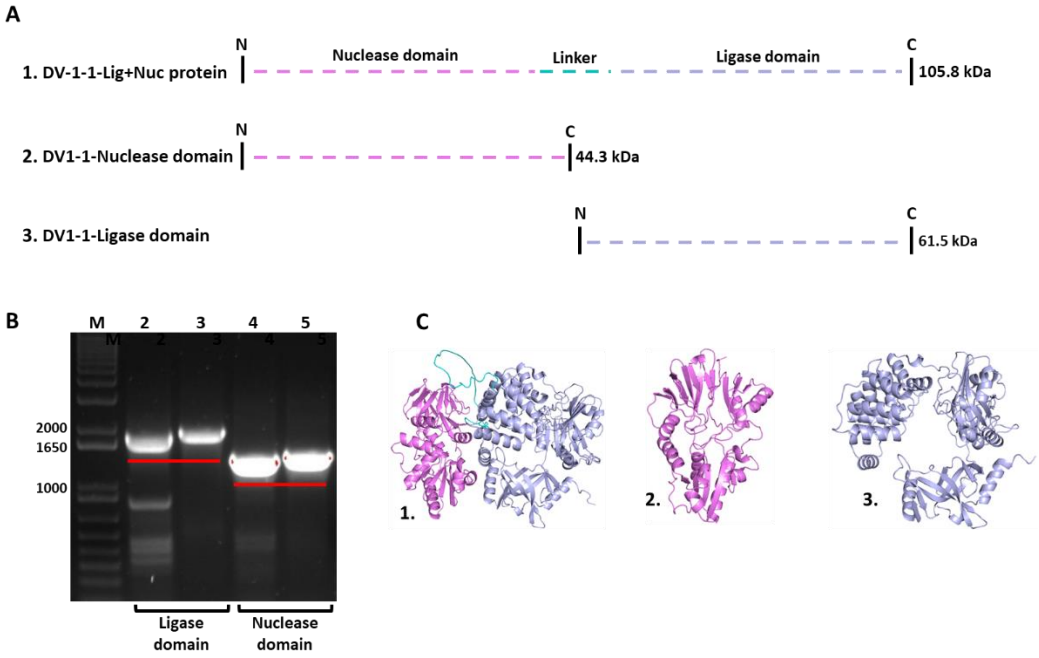


Figure 4.13. Design of new constructs to separate DV-1-1-Lig-Nuc protein into separate domains. **A**) Schematic of new construct designs for DV-1-1-Nuc domain and DV-1-1-Lig domain. **B**) Results of PCR, on an agarose gel, confirming correct sizes expected for DV-1-1-Lig and DV-1-1-Nuc genes in pDEST17 and pHMGWA plasmids. Lanes 2, 4 are constructs cloned into pHMGWA (1.904 bp & 1.418 bp), lanes 3, 5 are constructs cloned into pDEST7 (1.947 bp & 1.46 bp). 1 kb+ molecular weight ladder was used (M). **C**) Represent AlphaFold predicted models of 1. DV-1-1-Lig-Nuc, 2. DV-1-1-Nuc and 3. DV-1-1-Lig proteins. The predicted structures for protein models were generated by AlphaFold2_advanced, from Google Colab, version v2.3.1 (Jumper, Evans et al. 2021). Structural models were presented using PyMOL (Schrödinger, 2020).

The following sections detail expression, purification, and characterisation of DV-1-1-Lig and DV-1-1-Nuc proteins, starting with characterisation of the ligase domain.

4.2.6 DV-1-1-Lig protein expression, purification & crystallisation

4.2.6.1 Small scale protein expression testing

The gene construct for DV-1-1-Lig domain was cloned into pDEST17 and pHMGWA expression plasmids and transformed into pLysS, Arctic express and Origami *E. coli* expression strains, as described in **Section 2.2.2**. Small scale expression trials were performed using the three *E. coli* strains, at 15 and 20 ºC and results of these trials were run on SDS PAGE, as described in **Section 2.4.7**. Soluble protein expression was observed in all expression strains, in both pDEST17 and pHMGWA expression plasmids (data not shown). The best protein expression was seen in the Origami strain at 15 ºC, with results shown below (**Figure 4.14**).

Both pHMGWA (MBP-tagged) and pDEST17 (His-tagged) expression plasmids gave soluble expression of DV-1-1-Lig protein, best expression was seen with His-tagged protein.

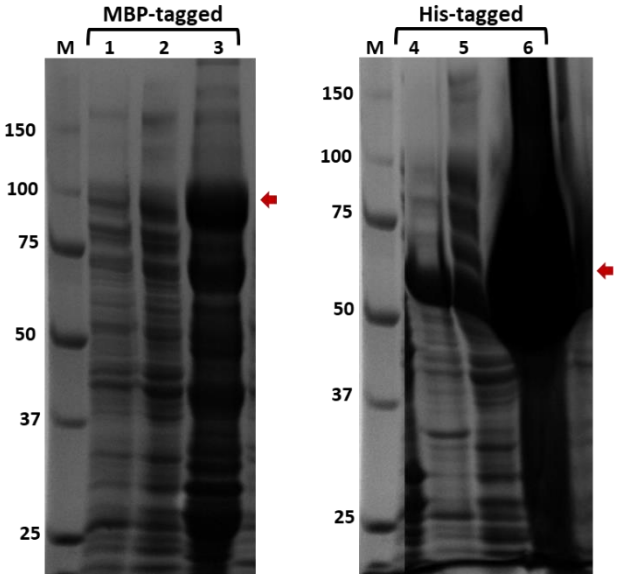


Figure 4.14. SDS PAGE of small-scale protein expression results for DV-1-1-Lig, in *E. coli* (DE3) Origami. Lanes 1 & 4 represent insoluble protein, lanes 2 & 5 represent soluble protein and lanes 3 & 6 represent soluble protein bound to Ni beads. Red arrows indicate expression of of DV-1-1-Lig protein, at the expected size for MBP tagged protein (105.4 kDa) and His-tagged protein (64.5 kDa). A precision plus protein ladder was used as a molecular weight marker (M).

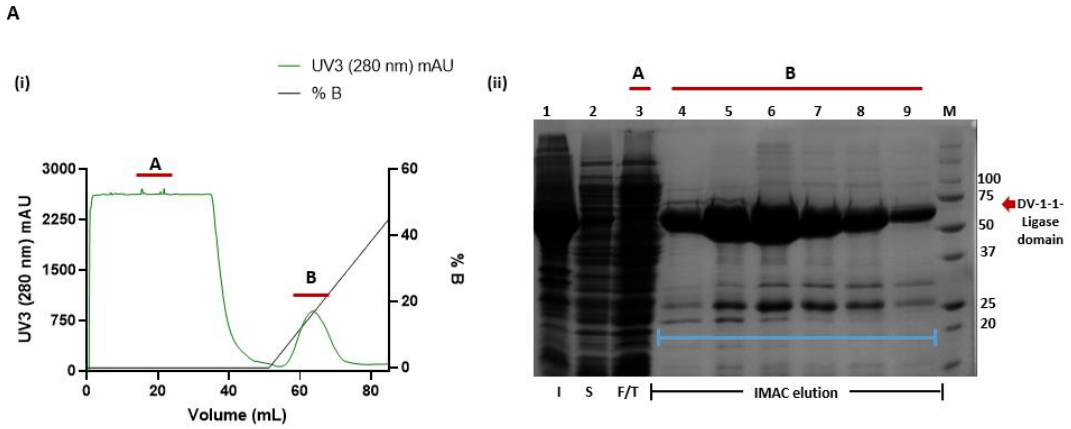
4.2.6.2 Large scale purifications

Results of small-scale expression trials, saw His-tagged DV-1-1-Lig protein, give the best soluble expression in *E. coli* (DE3) Origami, and therefore was used when scaling up protein expression cultures, as described in **Section 2.3.3**.

In the first purification attempts, TEV protease was added to the protein sample after the IMAC purification step, to remove the His-tag from the protein. Unfortunately, this resulted in DV-1-1-Lig protein precipitating overnight and prevented further purification of this protein (data not shown).

In the following purifications the His-tag was left on and a two-step purification *via* IMAC and gel filtration chromatography (as described in **Section 2.4**), produced soluble, active protein, suitable for characterisation experiments. The chromatograms and corresponding SDS-PAGE gels in **Figure 4.15** depict the purification, column load and flow through fractions.

A typical IMAC purification resulted in fractions of soluble, highly expressed DV-1-1-Lig protein, that eluted off the column with the addition of 15 % buffer B, along with some *E. coli* contaminating proteins (**Figure 4.15**, **A**). A gel filtration purification resulted in a single peak eluting off the column and with fractions containing relatively pure DV-1-1-Lig protein (**Figure 4.15**, **B**).



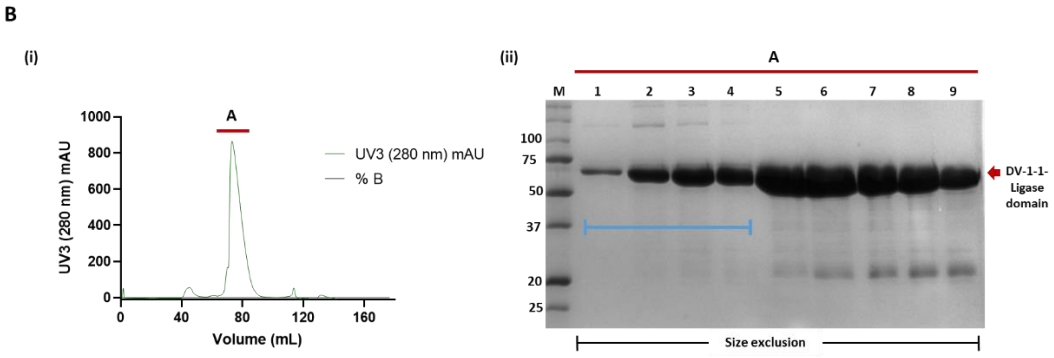


Figure 4.15. IMAC and gel filtration chromatography of DV-1-1-Lig. **A**) IMAC chromatogram (i) and SDS PAGE gel for production of DV-1-1-Lig protein from *E. coli* (DE3) Origami (ii). (i) Peak A represents flow through during IMAC purification, peak B represents fractions of proteins that eluted during the elution step of the IMAC purification, including DV-1-1-Lig protein (64.5 kDa). (ii) Lanes 1-3 represent insoluble (P), soluble (S) and flowthrough (F/T) samples. Lanes 4-9 represent fractions containing proteins that eluted off the column during the elution step, with the addition of buffer B. The blue bar indicates fractions that were pooled, up concentrated, and further purified *via* gel filtration chromatography. **B**) Gel filtration chromatogram (i) and SDS PAGE gel for production of DV-1-1-Lig protein from *E. coli* (DE3) Origami (ii). (i) Peak A represents where DV-1-1-Lig protein eluted off the gel filtration column. (ii) Lanes 1-9 represent the following proteins present in fractions from peak A (i). The blue bar indicates fractions that were pooled and up concentrated and stored for further use. A precision plus protein ladder was used as a molecular weight marker (M). Chromatogram graph was designed in GraphPad Prism, version 9.0.0.

To determine the molecular weight (MW) of DV-1-1-Lig and predict protein structural conformation in solution, purified DV-1-1-Lig protein was put through a Superdex™ S200 10/300 gel filtration column, as described in **Section 2.4.6**. Results from this purification (**Figure 4.16**) show DV-1-1-Lig protein eluted as a single peak, with an elution volume of 14.81 mls. Using the equation, from **Section 2.4.6**, the calculated MW of DV-1-1-Lig protein was 75.61 kDa. The predicted molecular weight of DV-1-1-Lig domain protein is 65.019 kDa, which suggests the protein elutes as a monomer (SEC MW/ sequence MW = 1.1629).

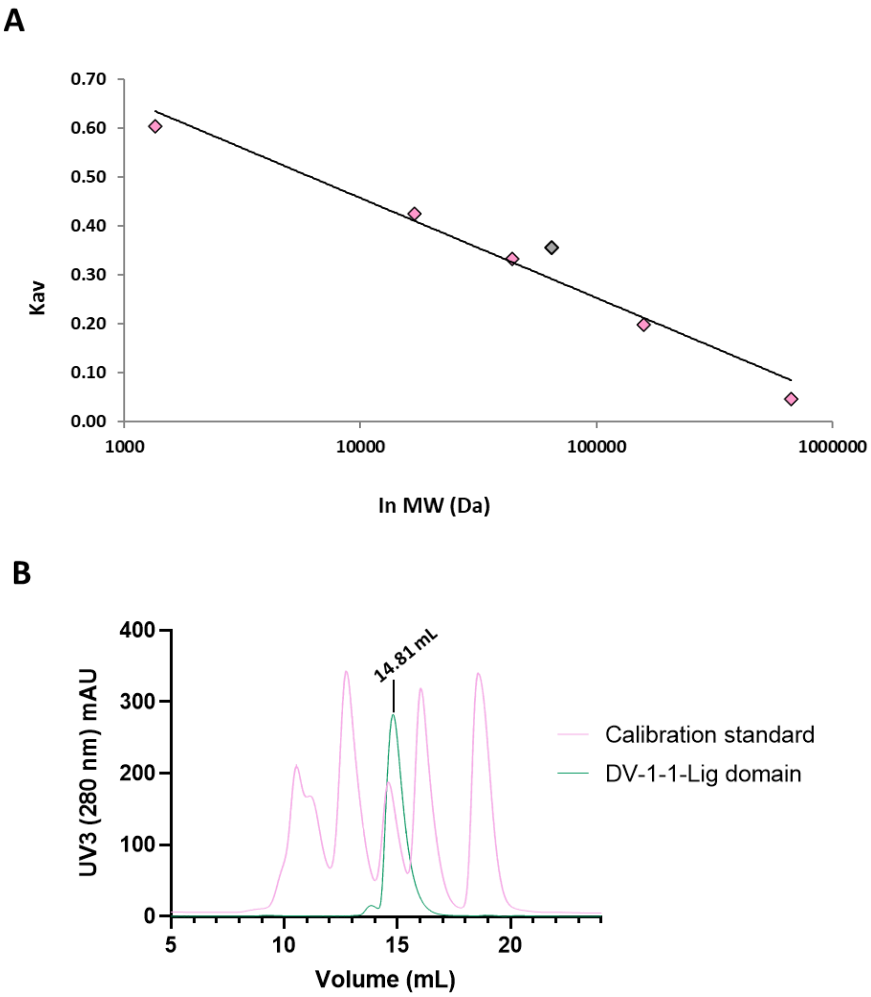


Figure 4.16. Analytical gel filtration chromatography of DV-1-1-Lig protein and protein standards to determine MW. **A**) Calibration Curve for S200 10/300 gel filtration column. K_{av} values of kit proteins plotted against the log molecular weight. DV-1-1Lig protein was also included and is indicated in green. **B**) Resulting chromatogram from gel filtration chromatography. Protein calibration standards are represented in pink and DV-1-1-Lig is represented in green. The elution volume for DV-1-1-Lig is 14.81 mls. Graphs were generated in Microsoft Excel version 2304 and GraphPad Prism version 8 (GraphPadSoftware).

4.2.6.3 Protein crystallization

Purified DV-1-1-Lig protein (8-15 μM) was used in crystallization trials to try and obtain well-diffracting crystals for structural characterization. Below (**Figure 4.17**) shows results of crystallization trials from robot screens, fine screens and seeding. In robot screens, protein crystals only formed from conditions in the Natrix screen in the presence of nicked DNA. Different conditions resulted in varying crystal morphology, ranging from rods to hexagons. Crystals with the hexagon morphology were often formed in unsymmetrical layers and were therefore unsuitable for X-ray diffraction experiments. Crystals formed in fine screens, were surrounded by a heavy precipitated protein film that made

them difficult to loop. Seeding, using a cat whisker, as described in **Section 2.6.2**, was successful in decreasing the extent of precipitation around the crystals, however the crystals that developed were too small for x-ray diffraction. Other techniques including additive screens, varying protein concentration, changing pH of crystallization condition and keeping screens at 4 ºC, did not improve crystal formation (data not shown). Some crystals were sent for data collection at Australian Synchrotron (beamline MX1), however none diffracted to a suitable resolution for further processing. It was observed that many of the crystals lost their morphology when removed from the crystallisation solution during looping, indicating they became disordered.

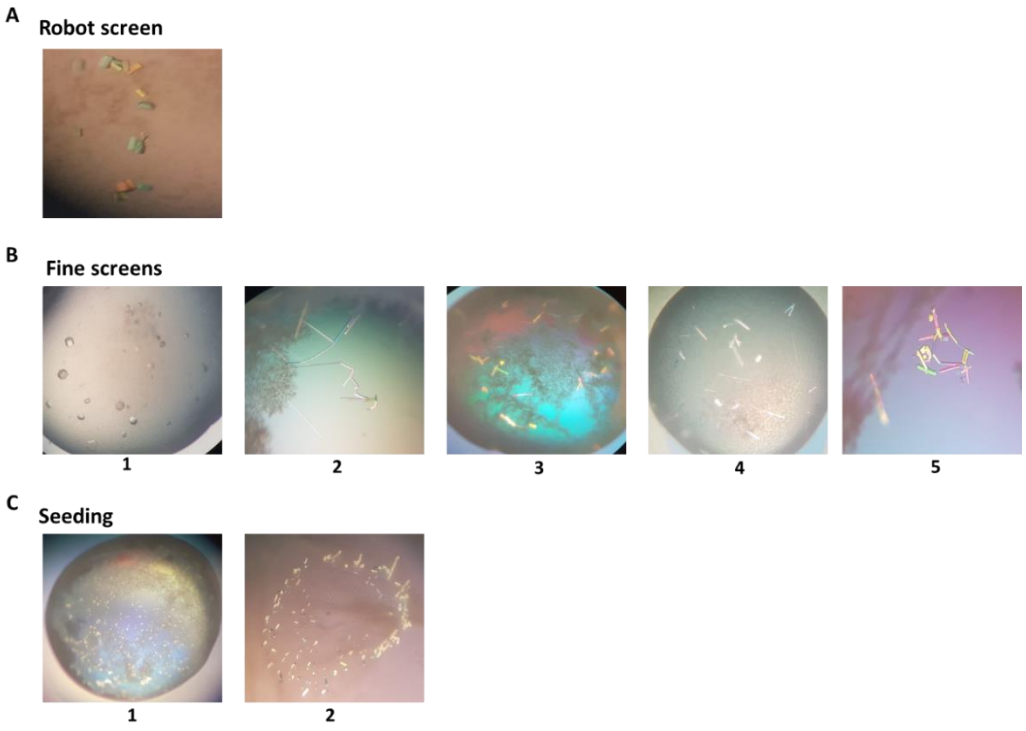


Figure 4.17. Crystal formation of DV-1-1-Lig protein in robot and fine screens. **A**) formation of protein crystals in a robot screen with natrix crystallization conditions (1. H1, 2. H2, 3. H3, 4. H5, 5. F3) crystallization solution makeup is described in **Section 2.6**. **B**) different protein crystal morphology resulting from different crystallization conditions, in fine screens. **C**) formation of protein crystals after seeding in fine screens using natrix conditions from above (1. H1, 2. H2).

4.2.7 Protein folding and stability of DV-1-1-Lig

The secondary structure of DV-1-1-Lig protein was analysed using circular dichroism (CD), which indicates the protein is well-folded in solution. Secondary structure content derived from the CD spectra (using BeStSel) was similar to the AlphaFold models described in **Section 4.2.2** (analyzed using

PDBsum) with only small discrepancies between helix and strand percentage contributions (**Figure 4.18**).

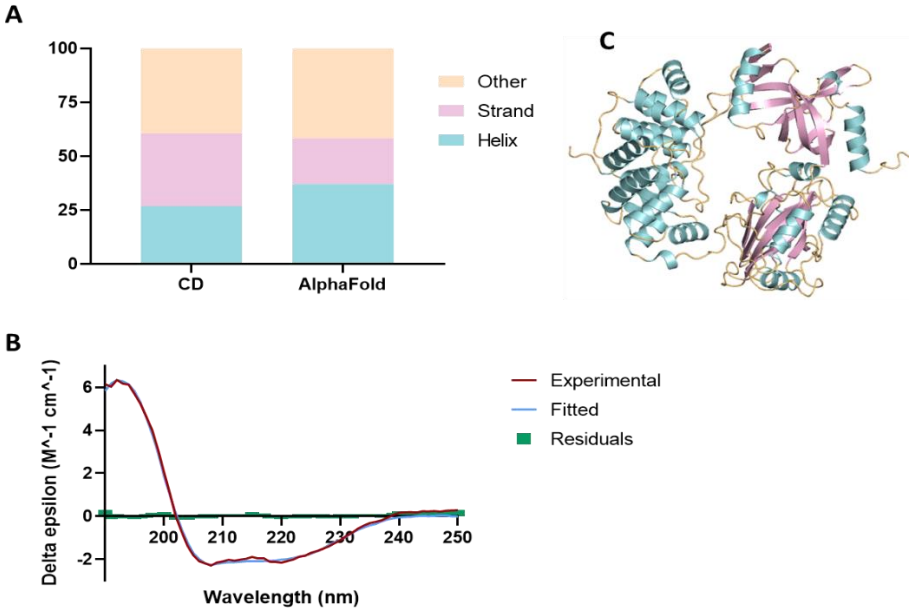


Figure 4.18. Circular dichroism (CD) and AlphaFold secondary structural composition of DV-1-1-Lig protein. **A**) A graph showing comparison of secondary structural predictions from CD and AlphaFold prediction model. **B**) Single spectrum analysis of CD spectra, using BeStSel database (Micsonai et al., 2018). **C**) AlphaFold 3D structural prediction of DV-1-1-Lig, coloured based on secondary structure (Helix in blue, strand in pink and other orange. (John Jumper, 2021). Graphs were designed using Prism version 8 (GraphPadSoftware). Wavelength range (190-250 nm) and scale factor (1). RMSD value (0.1361). NRMSD value (0.01578).

Determination of thermal stability using CD, was attempted for DV-1-1-Lig, however protein aggregation prevented data collection by this method. Instead, Differential Scanning Fluorimetry (DSF) (**Section 2.8**) was used to estimate the melting temperature (T_m) of DV-1-1-Lig protein and to further confirm the protein was folded. DSF was also used to compare the stability of DV-1-1-Lig protein in different pH conditions and determine the optimal pH range suitable for its storage and activity. Measurement over a range of protein concentrations indicates a T_m of 45 ºC. Thermal melts with different pH conditions, indicate an optimum pH range between 7 and 8 (**Figure 4.19**).

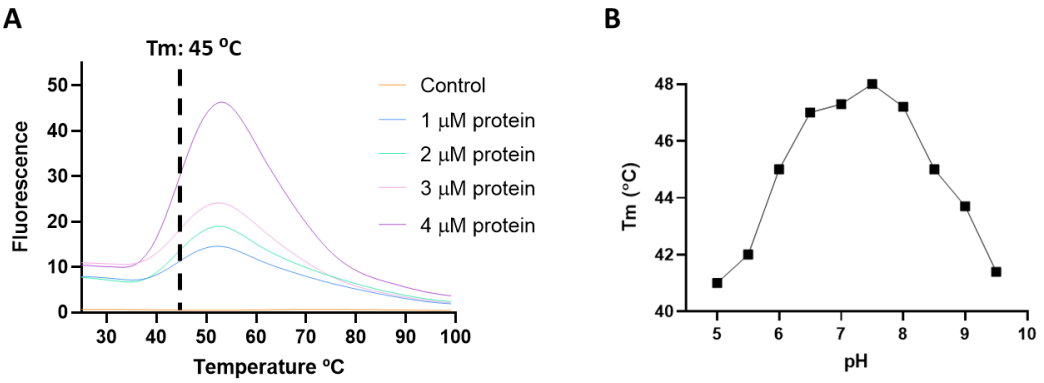


Figure 4.19. Results of differential scanning fluorimetry (DSF), with SYPRO orange, with DV-1-1-Lig. **A**) DSF with four different concentrations (1, 2, 3 & 4 µM) of DV-1-1-Lig protein. T_m values were determined from the midpoint in the unfolding equilibrium and are indicated on the graph, by a dotted line. Each concentration was carried out in replicates of three. **B**) DSF with DV-1-1-Lig protein, where reactions are of a different pH value. Reactions were carried out in replicates of three, for each pH value. T_m values were determined as above and are indicated in a plot graph. Protein was at a final concentration of 2.5 µM. Graphs were generated using GraphPad Prism version 8 (GraphPadSoftware).

4.2.8 Biochemical characterisation

4.2.8.1 DNA binding by DV-1-1-Lig domain

To evaluate the extent of DNA-binding, by the isolated DV-1-1-Lig domain, an EMSA (gel-shift) assay, as described in **Section 2.9**, was performed with a range of varying protein concentrations and temperatures. EMSAs were visualized on native TBE gels. Here DV-1-1-Lig domain showed weak binding to nicked DNA substrate, indicated by bound substrate in **Figure 4.20**.

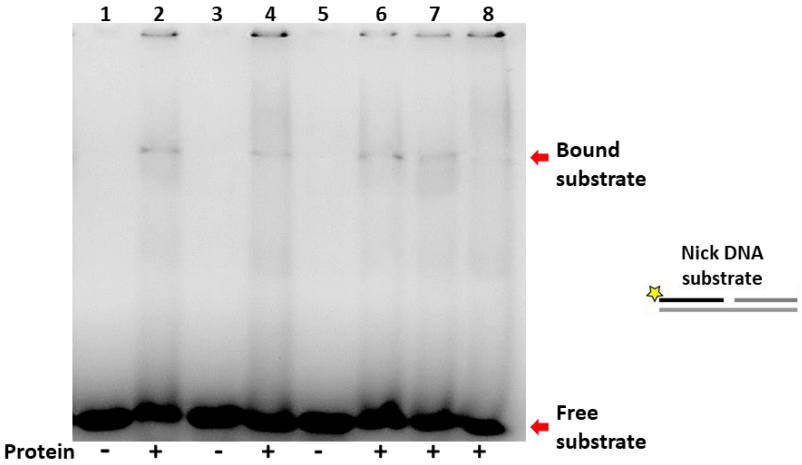


Figure 4.20. EMSA showing the binding ability of DV-1-1-Lig protein, to nicked DNA substrate. Lanes 1, 3 & 5 are control lanes, without protein (-). Lanes 2, 4, represent different incubation times, (0.5, 1 hour), with 1 µM protein. Lanes 6, 7 & 8, represent different protein concentrations (1, 2 & 4 µM), incubated for 1 hour. DV-1-1-Lig was incubated with nicked DNA substrate, at 25 °C, with 1 mM final ATP concentration and 40 mM EDTA. Reactions were run out on a 10 % native TBE gel, with native loading dye in the reactions. The

control lanes contain nicked DNA substrate with 1 mM, final ATP, EDTA, but no protein. Nicked DNA substrates are fluorescently labeled and were visualized using the Invitrogen™ iBright™ CL1500 Imaging System.

4.2.8.2 Protein concentration optimization for DV-1-1-Lig assays

To determine the best protein concentration to use in future assays, the activity of DV-1-1-Lig was measured over a range of concentrations. The results below (**Figure 4.21**) show that DV-1-1-Lig protein can ligate nicked DNA substrate down to a final protein concentration of 0.5 µM, while the maximum activity was observed with a protein final concentration of 6.5 µM, giving close to 80 % ligation of substrate. A final protein concentration of 4 µM, was chosen for future activity assays, as it would allow either increases or decreases in activity under different conditions to be detected.

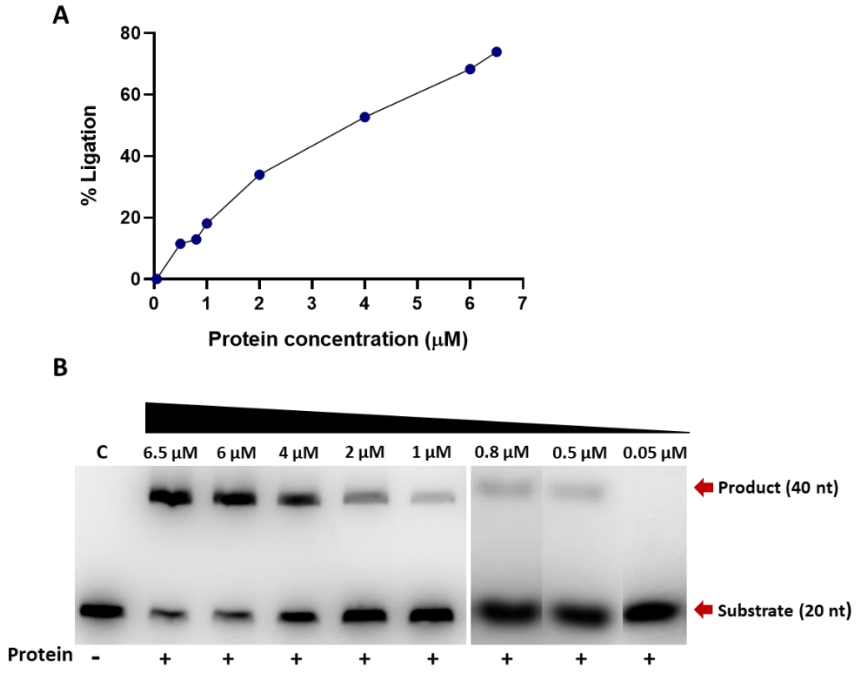


Figure 4.21. Shows ligation of nicked DNA substrate at different concentrations of DV-1-1-Lig protein. **A**) quantification of ligation by DV-1-1-Lig on nicked DNA, with different protein concentrations. **B**) TBE urea gel showing results of DV-Lig-1-1-Lig protein concentration gradient. Addition of protein to the reaction is indicated by a plus symbol (+). Control reaction is indicated by (C) and no protein (-). Product (40 nt) and substrate (20 nt) are indicated by red arrows. Reactions were carried out for 2 hours, at 25 ^{o}C, with varying final protein concentrations (6.5, 6, 4, 2, 1, 0.8, 0.5 & 0.05 µM), 1 mM final ATP concentration and 10 mM final concentration of magnesium ions. Results of activity assays were visualized using iBright™ CL750 Imaging System, Invitrogen™.

4.2.8.3 Time dependence of DV-1-1-Lig ligation

DV-1-1-Lig was incubated with nicked DNA substrate for different time periods (0.5, 1, 2, 3, 5 and 5 hour(s)) to determine what time point would give sufficient ligation of nicked DNA. The results below (**Figure 4.22**) show that ligation by DV-1-1-Lig is detectable after a one-hour incubation period and is almost complete by five hours. For future activity assays, DV-1-1-Lig was incubated with nicked DNA substrate for four hours, to ensure robust detection of ligation activity.

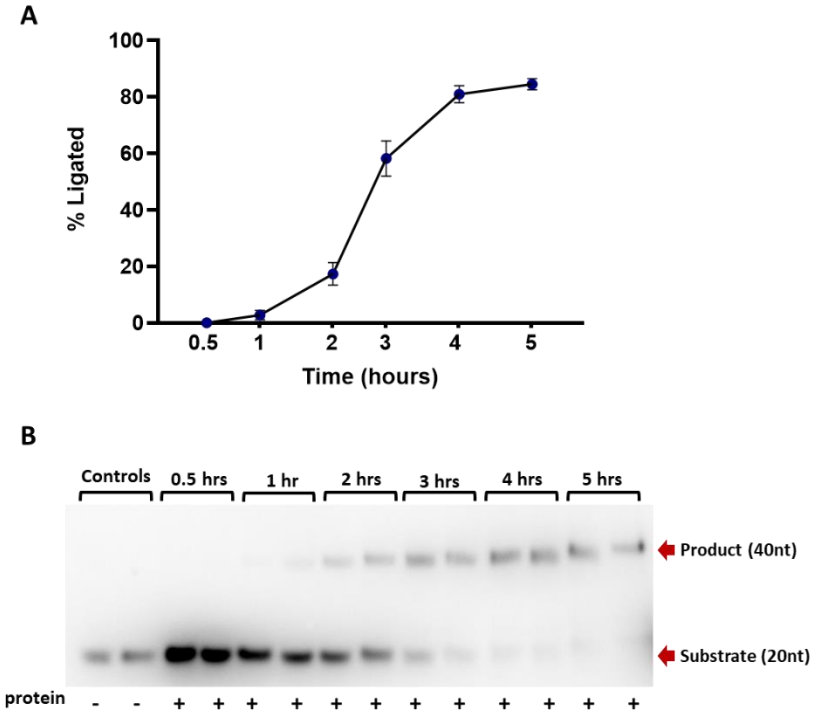


Figure 4.22. Ligation of nicked DNA substrate at different incubation time points, with of DV-1-1-Lig protein. **A**) Quantification of ligation by DV-1-1-Lig on nicked DNA, at different time points. Points on the graph represent averages of each time point. Standard deviation error bars are included. **B**) TBE urea gel showing results of DV-1-1-Lig ligation activity assay, on nicked DNA, at different time points. Addition of protein to the reaction is indicated by a plus symbol (+). Controls reactions are indicated by (Controls) and don't contain protein (-). Product (40 nt) and substrate (20 nt) are indicated by red arrows. Activity against each substrate was carried out in replicates of two. Reactions were incubated at different time points (0.5, 1, 2, 3, 4 and hour(s), at 25 ^{o}C, with 4 µM final protein concentration, 1 mM final ATP concentration and 10 mM final concentration of magnesium ion. Results of activity assays were visualized using iBright™ CL750 Imaging System, Invitrogen™.

4.2.8.4 Metal ion preference of DV-1-1-Lig

To determine the optimal divalent metal ion cofactor and the best concentration of either magnesium or manganese, activity assays were performed

with DV-1-1-Lig on nicked DNA substrate. Overall, a higher percentage of ligation is observed with the addition of manganese, compared to magnesium. Ligation of nicked DNA substrate only requires a low concentration of manganese, (1 mM), higher concentrations start to inhibit the reaction. Ligation reactions with magnesium require a higher concentration (5 mM) to reach optimal ligation, up until a point where an increase in magnesium also inhibits the reaction. No ligation of nicked DNA substrate is observed in reactions with no metal ion, or in reactions where EDTA is added in in addition to a metal ion (**Figure 4.23**).

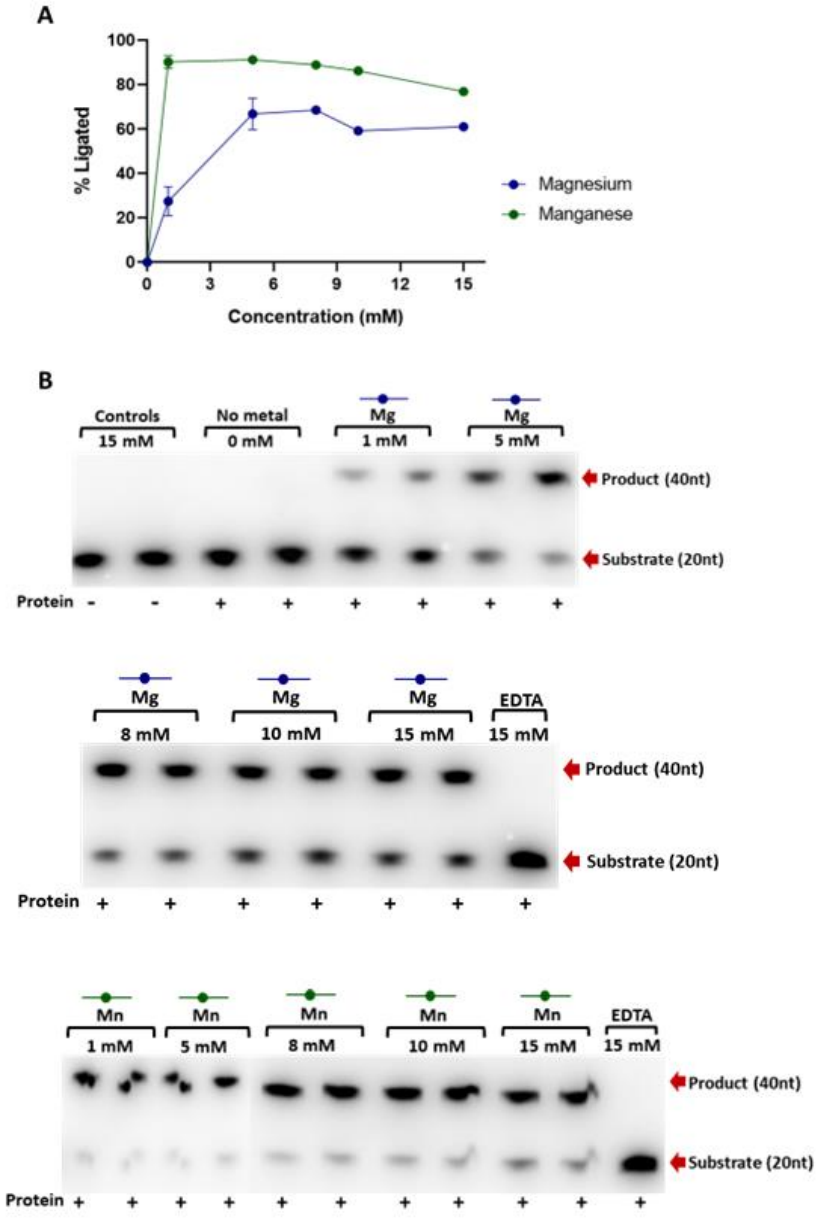


Figure 4.23. Ligation of nicked DNA substrate, by DV-1-1-Lig protein, with magnesium (Mg) or magnesium (Mn). **A**) Quantification of ligation by DV-1-1-Lig, with varying metal ion concentrations. Points on the graph represent the average ligation percentage for each metal concentration. Standard deviation error bars are included. **B**) TBE urea PAGE showing results of ligation with varying Mg and Mn ion concentrations, in

replicates of two. Addition of protein to the reaction is indicated by a plus symbol (+). Control reactions all contain nicked DNA and vary by; no protein (controls), protein with no metal ion (No metal) and protein, with 15 mM Mg or Mn and 40 mM EDTA (E). Product (40 nt) and substrate (20 nt) are indicated by red arrows. Reactions were carried out for 4 hours, at 25 ^{o}C, with 4 µM final protein concentration, 1 mM final ATP concentration and varying Mg or Mn ion final concentrations (1, 5, 8, 10 & 15 mM). Results of activity assays were visualized using iBright™ CL750 Imaging System, Invitrogen™.

4.2.8.5 Nucleotide cofactor specificity of DV-1-1-Lig

DV-1-1-Lig protein, like DV-Lig5 as described in **Section 3.2.4.4**, is pre-adenylated after purification from *E. coli* cells. This made identifying its required cofactors difficult as there was always a basal level of ligation seen across no cofactor controls and addition of more cofactor to the reaction often inhibited ligation. To solve this issue, the protein was pre-incubated with excess un-labelled nick DNA substrate to use up the cofactor already present, followed by the usual ligase activity assay procedure (**Section 2.9**).

While there is still a low basal level of ligation seen in the no cofactor control, it is evident that ligation is drastically increased with the addition of ATP and ADP. There is a moderate increase in ligation with GTP, in comparison to the no cofactor condition. Ligation with the addition of NAD does not show an increase in ligation of product, above background level and it is likely that the protein can't utilize NAD for ligation (**Figure 4.24**).

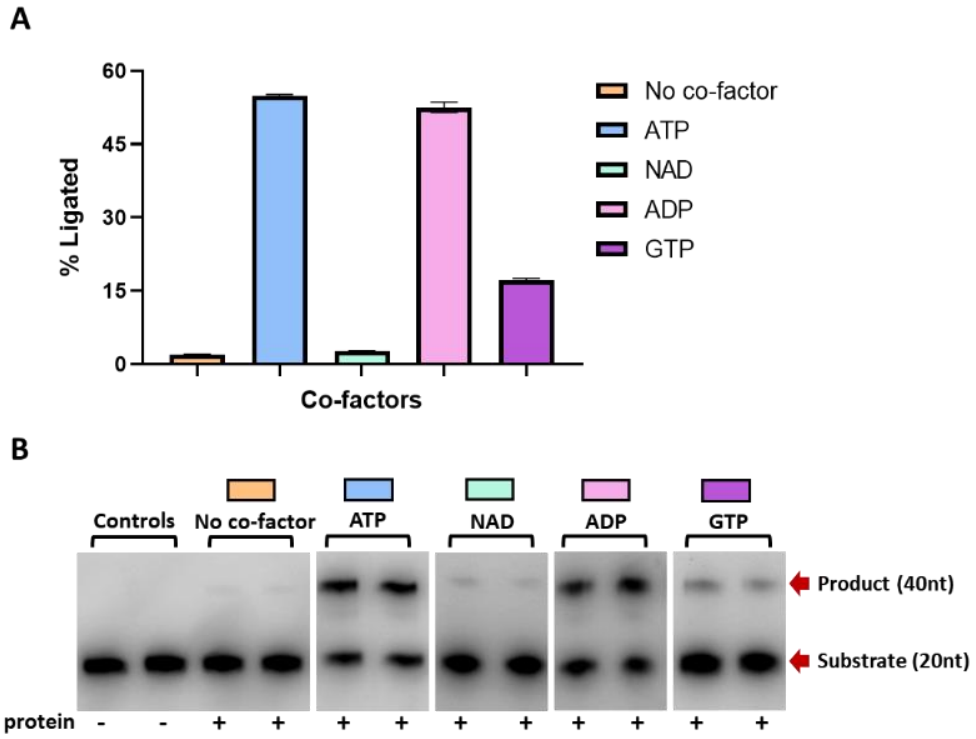


Figure 4.24. Ligation of nicked DNA substrate, by DV-1-1-Lig protein, with different cofactors. **A**) Quantification of ligation by DV-1-1-Lig on nicked DNA, with different cofactors (ATP, NAD, ADP & GTP). Points on the graph represent averages of each concentration. Standard deviation error bars are included. **B**) TBE urea PAGE showing results of ligation by DV-1-1-Lig, with and without the addition of different cofactors. Addition of protein to the reaction is indicated by a plus symbol (+). Controls reactions were used that don't contain protein (Controls) or don't contain cofactor (No cofactor). Product (40 nt) and substrate (20 nt) are indicated by red arrows. Activity against each substrate was carried out in replicates of two. Reactions were pre-incubated for 2 hours at 25 ^{o}C with unlabeled nicked DNA substrate, 4 µM protein and 5 mM magnesium ion, followed by a 4-hour incubation with the addition of labelled nicked DNA substrate, 5 mM magnesium, and different cofactors at 1 mM final concentration. Results of activity assays were visualized using iBright™ CL750 Imaging System, Invitrogen™.

A series of cofactor concentration gradient activity assays were carried out on nicked DNA substrate, using the cofactors (ADP, GTP and ATP), that supported ligation activity.

Here reactions with ADP showed slightly better ligation activity compared to those with ATP, which is surprising, as ATP is thought to be the main cofactor for Lig B type DNA ligases. Reactions with both ATP and ADP showed better ligation with a lower concentration of cofactor, between 0.15-0.5 mM. Reactions containing GTP showed low levels of ligation across all concentrations, with only a slight improvement seen up to 1 mM of GTP. No basal level of ligation was observed in reactions with no cofactor, signifying that pre-incubation with non-

labelled nicked DNA removed all pre-adenylated cofactor from DV-1-1-Lig protein (**Figure 4.25**).

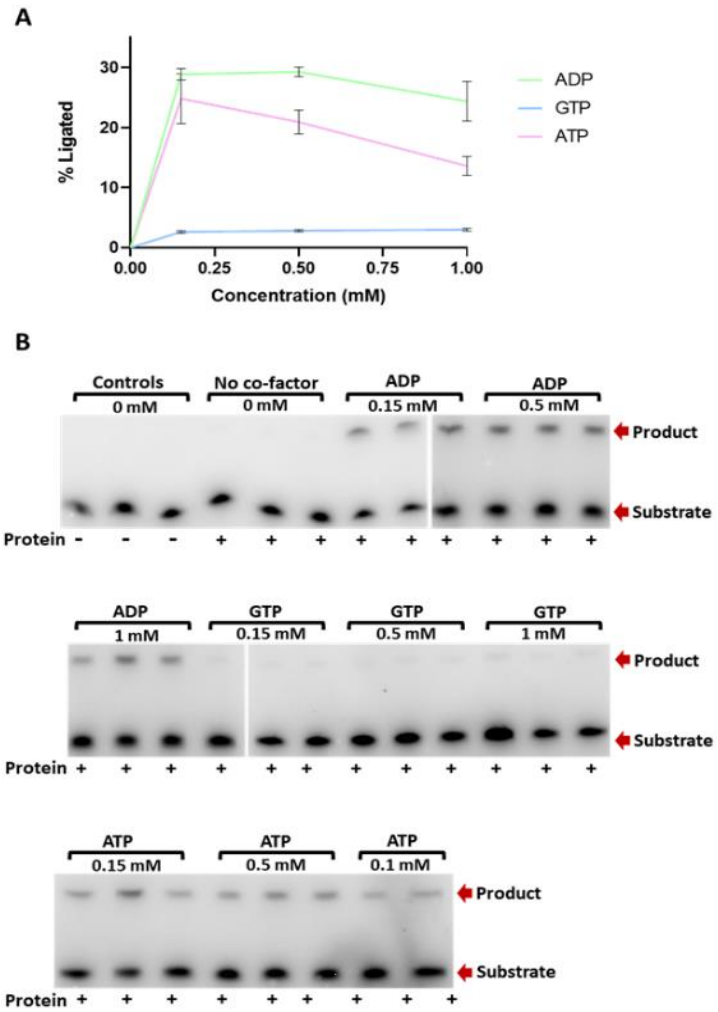


Figure 4.25. Ligation of nicked DNA substrate, by DV-1-1-Lig protein, with ADP, GTP and ATP. **A**) Quantification of ligation by DV-1-1-Lig on nicked DNA, with different cofactors, at varying concentrations. Points on the graph represent averages of each concentration. Standard deviation error bars are included. **B**) TBE urea PAGE showing results of ligation by DV-1-1-Lig, the addition of different cofactors, at increasing concentrations. Addition of protein to the reaction is indicated by a plus symbol (+). Controls reactions were used that don't contain protein (Controls) or don't contain cofactor (No cofactor). Product (40 nt) and substrate (20 nt) are indicated by red arrows. Activity against each substrate was carried out in replicates of three. Reactions were pre-incubated for 2.5 hours at 25 ^{o}C with unlabeled nicked DNA substrate, 3 µM protein and 5 mM magnesium ion, followed by a 2-hour incubation with the addition of labelled nicked DNA substrate, 5 mM magnesium, and different cofactors at 1 mM final concentration. Results of activity assays were visualized using iBright™ CL750 Imaging System, Invitrogen™.

4.2.8.6 Temperature dependence of DV-1-1-Lig

To determine the optimal temperature for DV-1-1-Lig protein activity, a temperature gradient assay, from -5 ºC to 80 ºC was performed. Ligation ability of DV-1-1-Lig at temperatures was compared with the addition of magnesium and manganese metal ions.

Overall, these results show that DV-1-1-Lig protein can ligate nicked DNA substrate at a wide temperature range (-1 to 80 ºC). At low and high temperatures more ligation of substrate is observed with the addition of manganese, compared to those with magnesium. The optimal temperature for ligation is between 25 and 30 ºC (**Figure 4.26**).

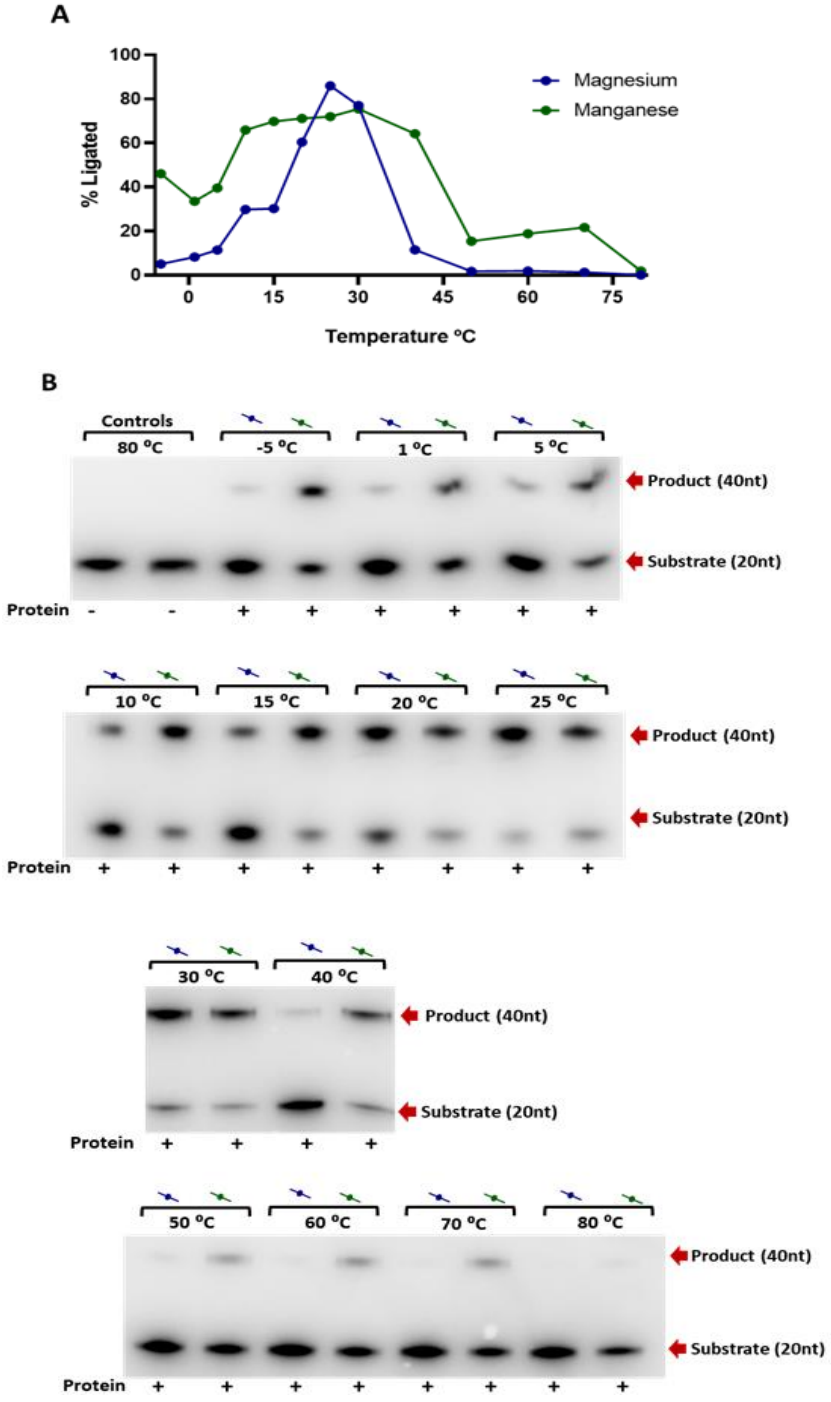


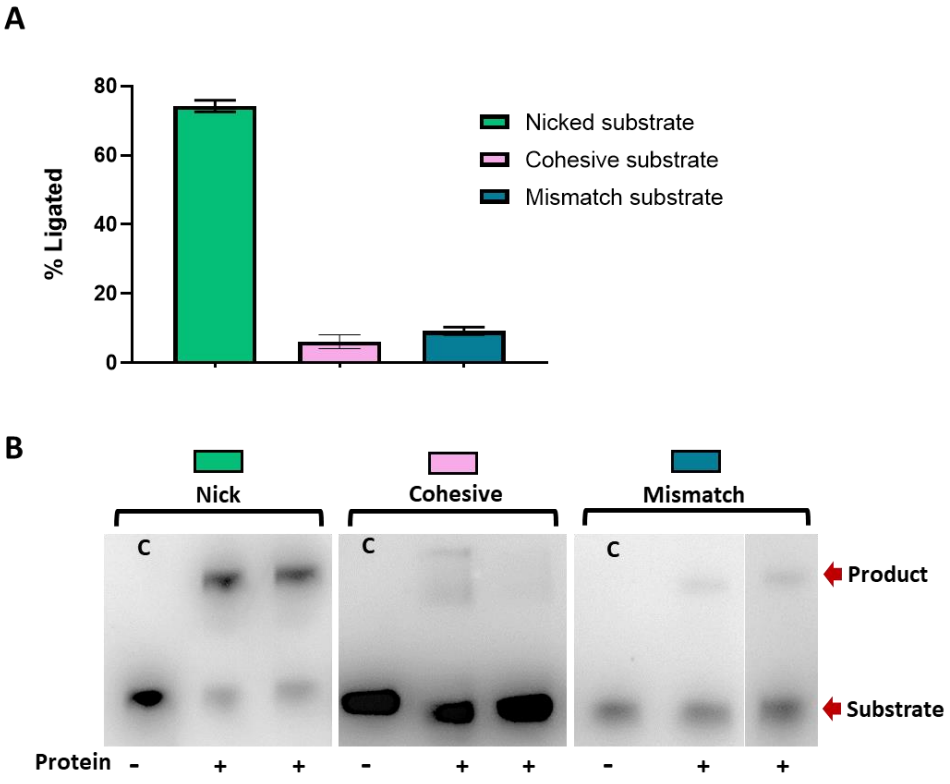
Figure 4.26. Ligation of nicked DNA substrate, by DV-1-1-Lig protein, at varying temperatures. **A**) Quantification of ligation by DV-1-1-Lig on nicked DNA, with different reaction temperatures (-5, 1, 5, 10, 15,

20, 25, 30, 40, 50, 60, 70, 80 °C). Points on the graph represent averages of each reaction temperature. **B**) TBE urea PAGE showing results of ligation by DV-1-1-Lig, at different temperatures. Addition of protein to the reaction is indicated by a plus symbol (+). Controls reactions (Controls) don't contain any protein (-) and were incubated at 80 °C, to ensure no degradation of substrate at higher temperatures. Product (40 nt) and substrate (20 nt) are indicated by red arrows. Reactions were carried out for 4 hours, at varying temperatures, with 4 µM final protein concentration, 1 mM final ATP concentration and 10 mM final metal ion concentrations. Results of activity assays were visualized using iBright™ CL750 Imaging System, Invitrogen™.

4.2.8.7 DNA substrate specificity of DV-1-1-Lig

Activity assays with a range of different DNA substrates (nick, cohesive, mismatch, gapped and blunt) (**Figure 4.27, C**), were designed to determine if DV-1-1-Lig protein was able to ligate substrates containing additional types of DNA breaks, alongside nick DNA substrates. Results of these activity assays were visualised using denaturing PAGE.

Analysis of results from the ligation assays showed that DV-1-1-Lig could ligate substrates with cohesive or mismatched DNA breaks, however the efficiency of ligation on these substrates was relatively low compared to that on nick DNA substrates. No ligation of gapped or blunt-ended substrates was detected (**Figure 4.27**).



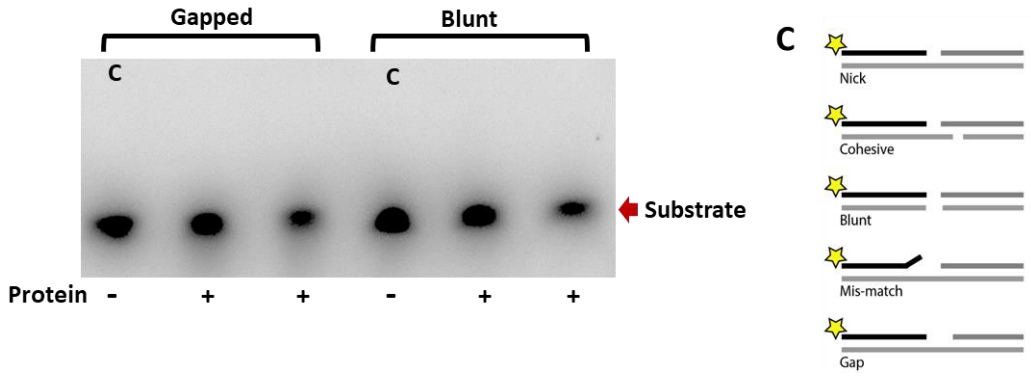


Figure 4.27. Results of ligation on different DNA substrates, by DV-1-1-Lig protein. **A**) Quantification of ligation by DV-1-1-Lig on nicked, cohesive and mismatch DNA substrates, from replicates of two. Standard deviation error bars are included. **B**) TBE urea PAGE showing results of ligation, by DV-1-1-Lig on 5 different DNA substrates. Substrate (20 nt) and product (40 nt) are indicated by red arrows. Addition of protein to reaction is indicated by a plus symbol (+), controls (C) don't contain any protein (-). Reactions were run in replicates of two. C) schematic of DNA substrates used in reactions. Star indicates fluorescent label. Reactions were carried out for 8 hours, at 20^{o}C, with 4 µM final protein concentration, 1 mM final concentration of ATP and 10 mM final concentration of magnesium. Results of activity assays were visualized using iBright™ CL750 Imaging System, Invitrogen™.

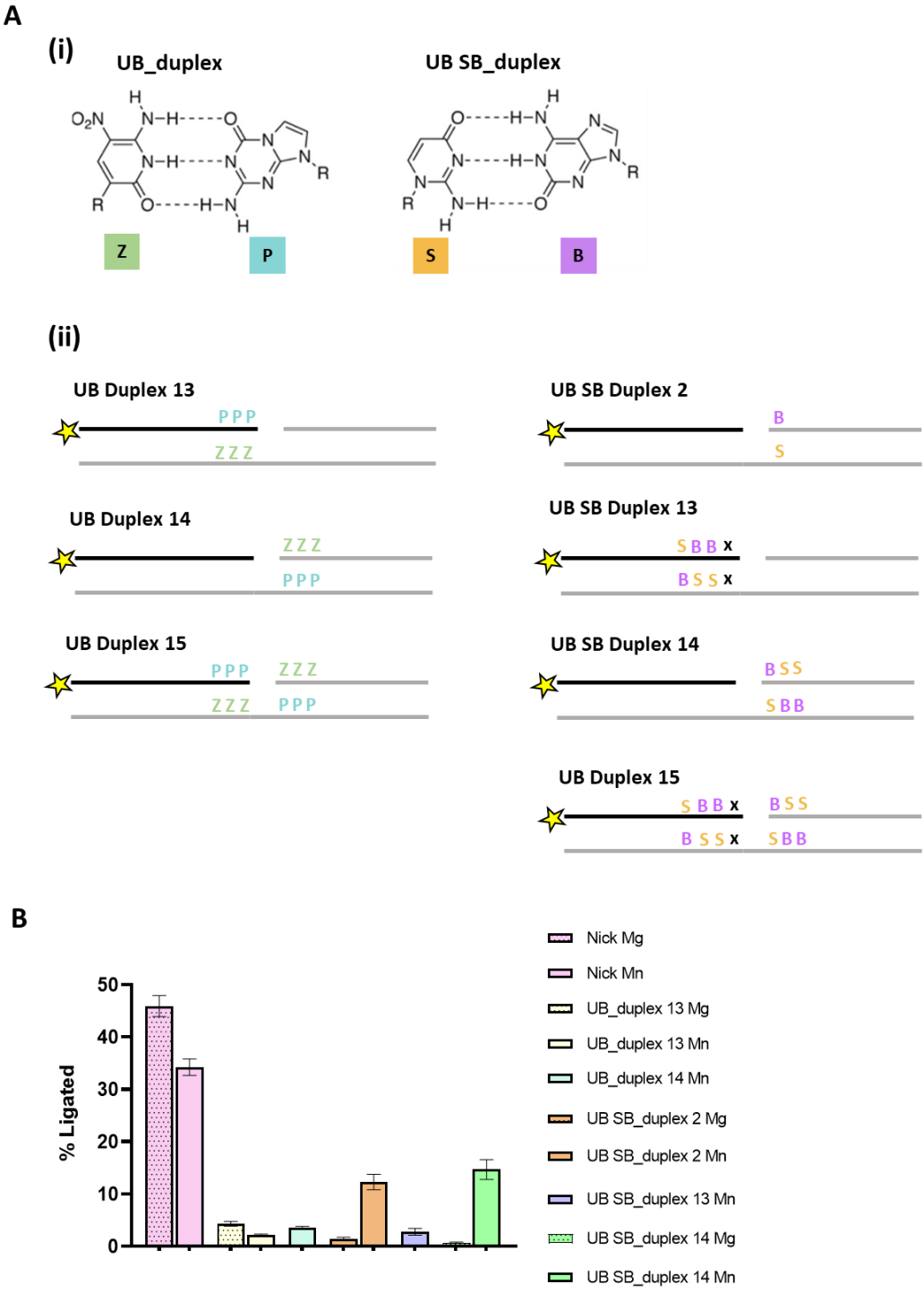
4.2.8.8 Ligation of non-canonical DNA substrates by DV-1-1-Lig

The ligation ability of DV-1-1-Lig protein was tested on nicked DNA substrates containing unnatural base pairs (UBPs), to determine how the type and placement of these UBPs, in the DNA substrate would affect the efficiency of ligation.

The UBPs used in this experiment have been previously described in **Section 1.10** and **Section 1.1.1.1.** Overall, seven different non-canonical DNA substrates (UB_duplex 13, UB_duplex 14, UB_duplex 15, UB SB_duplex 2, UB SB_duplex 13, UB SB_duplex 14 and UB SB_duplex 15) (**Figure 4.28, A**) were used in gel-based activity assays with DV-1-1-Lig. Nick DNA substrate, with non-modified base pairs was used as a positive control.

Analysis of results from the ligation assays showed that DV-1-1-Lig was able to ligate several of these UBP containing substrates, but the ligation efficiency was relatively low compared to that of nick DNA substrate with natural bases (**Figure 4.28, B, C**). The best ligation with the UBP substrates, was observed on the S-B bases and on substrates where the UBP was situated on the 5' end of the nick. Activity was observed with both magnesium and manganese as

metal ion cofactors. Interestingly activity on the UBP substrates was usually increased with the addition of manganese over magnesium.



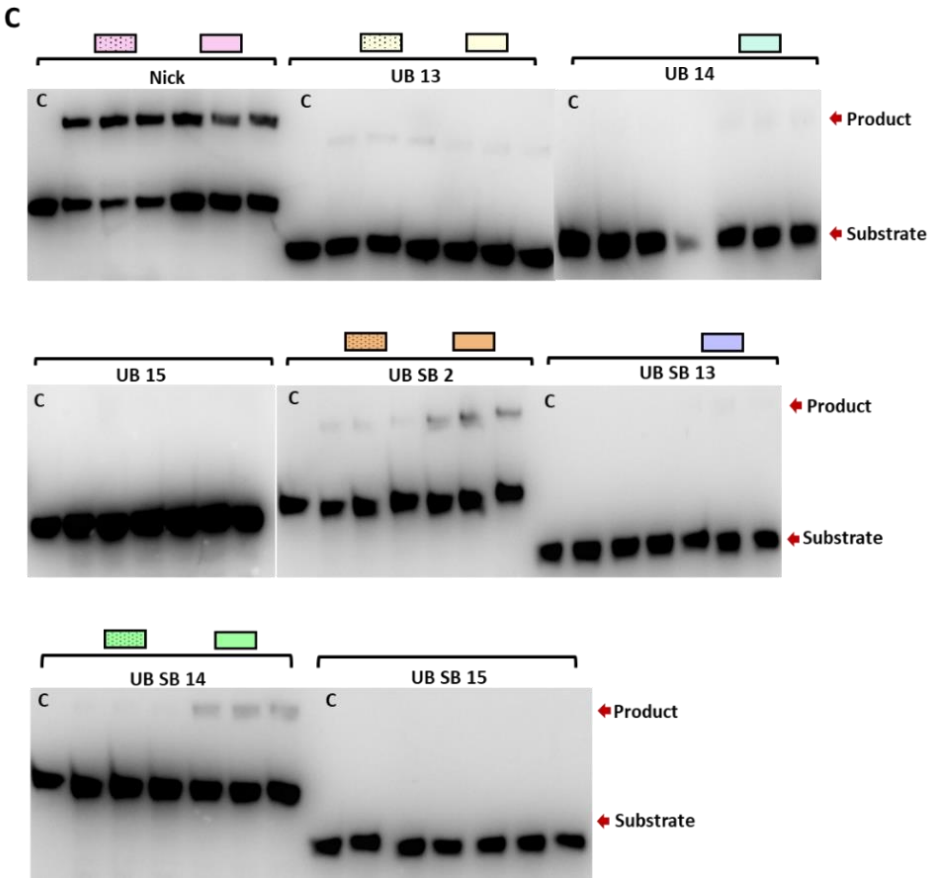


Figure 4.28. Represents the ligation ability of DV-1-1-Lig on a range of substrates with 3-6 non-canonical expanded base-pair substrates, with either magnesium (Mg) or manganese (Mn) as the divalent metal cofactor. **A**) **i** represents chemical modification of DNA to generate UB and SB DNA duplexes. **ii** represents the seven non-canonical DNA substrates, containing P and Z or S and B UBPs. X on the figures represents natural DNA bases. Stars represent labelling with the 6-carboxyfluorescein at the 5' terminus (5'FAM). **B**) represents the quantitative summary of ligation by DV-1-1-Lig on nicked DNA and seven different non-canonical substrates. Error bars represent standard deviation. **C**) represents the results of these ligation activity assays shown on urea PAGE gels. Nick DNA substrate, with non-modified bases, is indicated by a pink box. Controls are represented by C and contain no protein. Activity against each substrate was carried out in replicates of three. Product and substrate bands are indicated on the gel, by red arrows. Reactions were carried out for 2 hours, at 25^{o}C, with 4 µM final protein concentration, 1 mM final concentration of ATP and 10 mM final concentration of metals. Results of activity assays were visualized using iBright™ CL750 Imaging System, Invitrogen™. The graph was generated using GraphPad Prism version 8 (GraphPadSoftware).

4.2.9 DV-1-1-Nuc protein cloning, expression & purification

4.2.9.1 Small scale protein expression testing of DV-1-1-Nuc

The gene construct for DV-1-1-Nuc domain was cloned into pDEST17 (His-tag) and pHMGWA (MBP-tagged) expression plasmids and transformed into the BL21 (DE3) pLysS, Arctic express (DE3) and Origami (DE3) *E. coli* expression strains, as described in **Section 2.2.1.** Small scale expression trials showed no expression of DV-1-1-Nuc domain in BL21 plysS or Arctic express

strains (data not shown). Protein expression was observed with His-tagged DV-1-1-Nuc, in Origami, however this was in the insoluble fraction indicating it was produced as inclusion bodies (**Figure 4.29**).

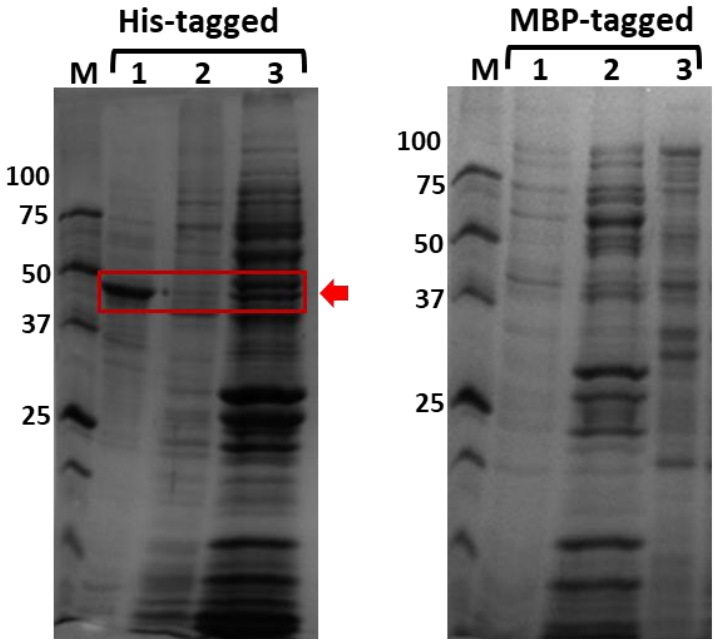


Figure 4.29. SDS PAGE of small-scale protein expression results for His-tagged and MBP tagged DV-1-1-Nuc. Lanes 1 represent insoluble protein, lanes 2 represent soluble protein and lanes 3 represent soluble protein bound to Ni beads. Red arrow indicates expression of of DV-1-1-Nuc protein, at the expected size for His tagged protein (47.7 kDa). No protein expression is seen in MBP tagged protein (88.14 kDa). A precision plus protein ladder was used as a molecular weight marker (M).

Following on from these results, three new constructs were designed for DV-1-1-Nuc domain, in an attempt to produce soluble protein, as described below.

4.2.9.2 DV-1-1-Nuc construct designs

New constructs for DV-1-1-Nuc domain were designed to remove regions from the beginning of the protein, which were likely incorrectly annotated and the end of the protein, which was predicted to have high disorder.

The protein sequence for DV-1-1-Lig-Nuc was annotated by in the IMG/JGI database (Chen et al., 2023), with a leucine (L) as the starting N-terminal amino acid. Alignments of the DV-1-1-Lig-Nuc protein sequence to other homologous proteins, through a NCBI Blast search (Altschul et al., 1990), revealed a potential error in the prediction of the start codon. AlphaFold2

structural model predictions of DV-1-1-Lig-Nuc protein, show 39 amino acids at the N-terminus as an unstructured region and these amino acids are not present in homologous proteins shown in **Figure 4.30**. A new start site start was proposed, based on this alignment, indicated by arrow 1 in **Figure 4.30**, **A**, that better aligned the protein sequence of DV-1-1-Lig-Nuc with other homologous proteins. The new start codon is also a leucine. In the design of new constructs for DV-1-1-Nuc domain, this new start site was used for construct 1 and construct 2, with construct 3 having the original start site (**Figure 4.30**, **B**).

Between the two protein domains of DV-1-1-Lig-Nuc, is a long flexible linker, which was predicted to have high disorder by EBI InterPro Scan (Jones et al., 2014). DV-1-1-Nuc domain was originally designed with some of this linker region included in the sequence, which may have contributed to the production of insoluble protein. Constructs 1 and 3 were designed with all of this linker region removed from the protein, as indicated in **Figure 4.30**, **B & C**, while construct 2 retained the original N-terminal from DV-1-1-Nuc domain.

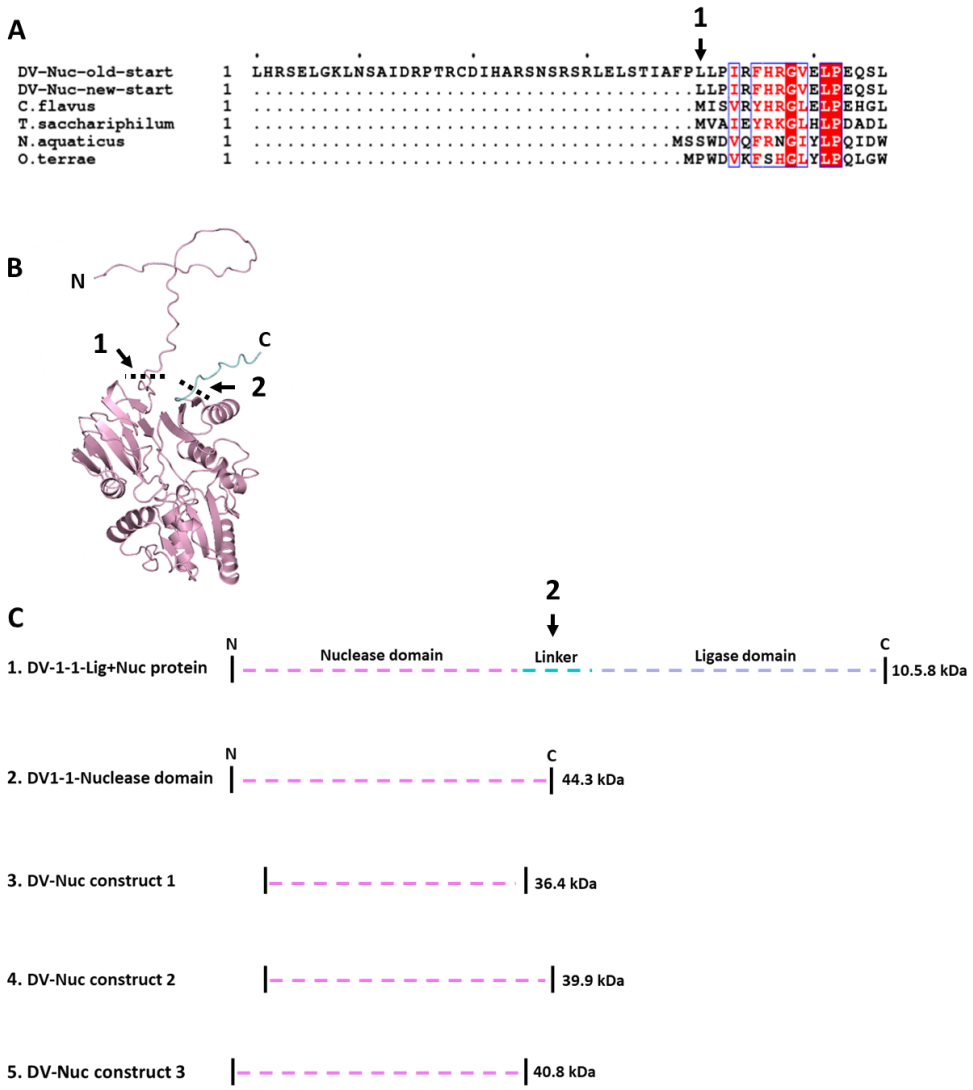


Figure 4.30. Schematic of new construct designs for DV-1-1-Nuc domain protein. **A**) Multiple sequence alignment of DV-1-1-Lig-Nuc (original and new start site) against other homologous proteins encoded by *A. fulgidus*, *P. furiosus*, *T. sibiricus, S. solfataricus* and human. New start site is indicated by arrow 1. **B**) Predicted structure of DV-1-1Nuc domain (pink) and linker region (blue). Indication of new N-terminal (1.) and C-terminal (2.) sites, are indicated by black arrows. **C**) Diagram showing design of 3 new constructs for DV-1-1-Nuc domain. Sequence alignment was created using Clustal Omega version 1.2.4. **Figure 4.30. A**, was created using ESPript 3 (Robert & Gouet, 2014). The predicted structure for DV-1-1-Ligase domain protein, **Figure 4.30. B**, was generated by AlphaFold2_advanced, from Google Colab, version v2.3.1 (Jumper, Evans et al. 2021). The model was presented using PyMOL (Schrödinger, 2020).

4.2.9.3 Small scale expression testing of new constructs

DV-1-1-Nuc domain constructs in pDEST17 and pHMGWA expression plasmids were transformed into *E. coli* expression strains, pLysS, Arctic express and Origami as described in **Section 2.2.1**. Small scale expression trials were performed using the three *E. coli* strains and results of these trials were analysed by SDS PAGE, as described in **Section 2.4.7**. The Origami strain gave the best

soluble protein expression for all three constructs, with results shown below in **Figure 4.31**. Soluble protein expression was observed in pHMGWA plasmids (MBP-tagged) for all three constructs, with construct 1 showing the most soluble protein expression. Protein was expressed in pDEST17 plasmids (His-tagged), for construct 2 and construct 3, however the expressed protein was insoluble.

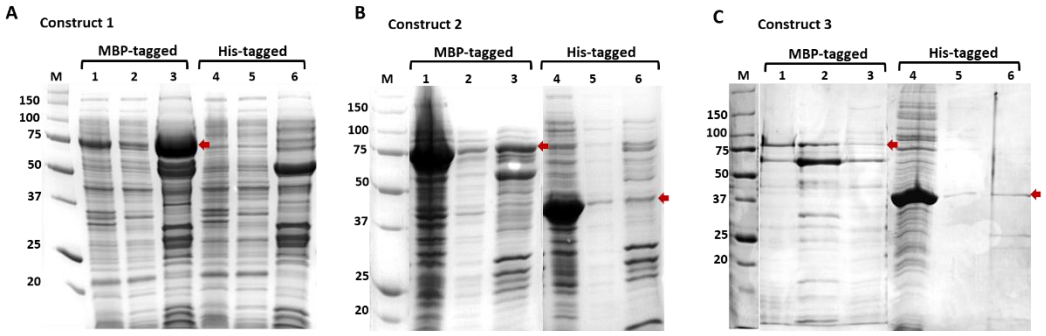


Figure 4.31. SDS PAGE gels of small-scale expression trials for DV-1-1-Nuc domain construct 1, construct 2, construct 3. **A**) represents protein expression for construct 1 in MBP-tagged protein (80.2 kDa) (pHMGWA plasmid) and His-tagged protein (39.8 kDa) (pDEST17 plasmid). **B**) represents protein expression for construct 2 in MBP-tagged protein (83.8 kDa) (pHMGWA plasmid) and His-tagged protein (43 kDa) (pDEST17 plasmid). **C**) represents protein expression for construct 3 in MBP-tagged protein (84.6 kDa) (pHMGWA plasmid) and His-tagged protein (44.2 kDa) (pDEST17 plasmid). Lanes 1,4, represent insoluble pellet samples, lanes 2,5, represent soluble samples and lanes 3,6, represent samples bound to Ni beads. Red arrows indicate expressed protein at the correct size. M in each gel stands for a precision plus 250 kDa protein marker. Proteins were recombinantly expressed in *E. coli* Origami cells and grown at 15 ^{o}C overnight.

4.2.9.4 Large scale purification of DV-1-1-Nuc

Following on from results of soluble protein expression of MBP-tagged DV-1-1-Nuc construct 1 (henceforth named DV-1-1-Nuc), in *E. coli* Origami, in small scale screens, protein expression cultures were scaled up following methods from **Section 2.3.3**. A three/four step purification *via* IMAC and an overnight TEV digest and reverse IMAC (if MBP tag was to be removed), followed by MBP purification and/or gel filtration chromatography; produced soluble, active protein, suitable for characterisation experiments. The chromatograms and corresponding SDS-PAGE gels in **Figure 4.32**, depict the purification, column load and flow through fractions.

An IMAC purification resulted in the elution of DV-1-1-Nuc-$_{MBP}$, with the at an imidazole concentration of 40 mM. A 60 kDa *E. coli* contaminant eluted off the IMAC column with the addition of wash buffer. DV-1-1-Nuc$_{MBP}$ was used in a

de-salting column, followed by incubation with TEV protease, which cleaved the MBP tag from DV-1-1-Nuc, with 80 % cleavage efficiency. A reverse IMAC purification and gel filtration resulted in highly pure DV1-1-Nuc as judged by SDS-PAGE.

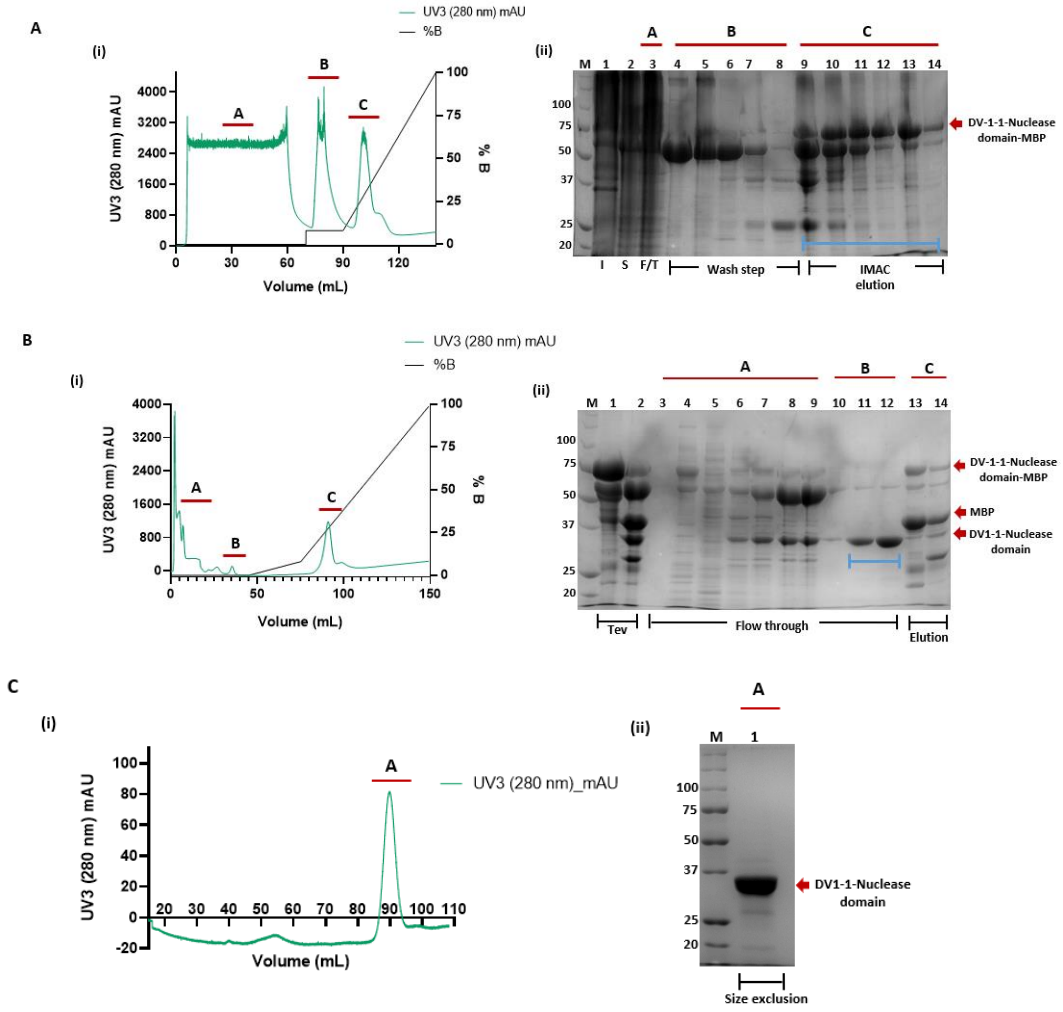


Figure 4.32. IMAC & gel filtration chromatograms (i) and SDS PAGE gels for production of DV-1-1 Nuclease from *E. coli* Origami (ii). **A**) IMAC purification of DV-1-1-Nuc-$_{MBP}$ (i) Peak A represents flow through during IMAC purification, peak B represents proteins that eluted from the IMAC column during the 8% imidazole wash step, peak C represents where the protein of interest (80 kDa) eluted during the elution step of the IMAC purification. Lanes 1-3 are; insoluble (I), soluble (S) and flowthrough (F/T), lanes 4-8 are fractions eluted during a 8% imidazole wash step, lane 9-14 are fractions eluted during the imidazole gradient in the first IMAC step. The blue bar indicates fractions that were pooled and incubated overnight with Tev protease. **B**) Reverse IMAC purification of DV-1-1-Nuc. (i) Peak A represents flow through during IMAC purification, peak B represents fractions that contain the de-tagged protein of interest (36 kDa), peak C represents fractions eluted during the elution step of the IMAC purification. (ii) Lanes 1-2 are pooled IMAC fractions before the addition of TEV (1) and fractions after an overnight incubation with TEV (2), Lanes 3-12 are fractions from the flowthrough during the reverse IMAC purification. Lanes 13-14 are fractions eluted during the imidazole gradient step. Blue bar indicates fractions that were kept and up concentrated for size exclusion. **C**) Gel filtration purification of DV-1-1-Nuc. (i) Peak A contains protein of interest (DV-1-1-Nuclease). (ii) Lane 1; represents up concentrated pool fractions from Peak A (i). DV-1-1-Nuclease protein is indicated by red arrow (36 kDa). Chromatogram graph was designed in GraphPad Prism, version 9.0.0.

LC-MS/MS was used to analyze trypsin digested samples of DV-1-1-Nuc domain, from an SDS-PAGE (**Figure 4.33**, (**i**)), to confirm that the purified protein was in-fact the new DV-1-1-Nuc construct protein and that cloning steps had been successful. The resulting peptides were mapped to the DV-1-1-Nuc amino acid sequence and confirmed that the extracted protein band matched the protein sequence for DV-1-1-Nuc. The sequence coverage was 95 % and gave an average mass of 36.429 kDa. **Figure 4.33**, (**ii**) shows 95 % sequence coverage of the matched tryptic peptides.

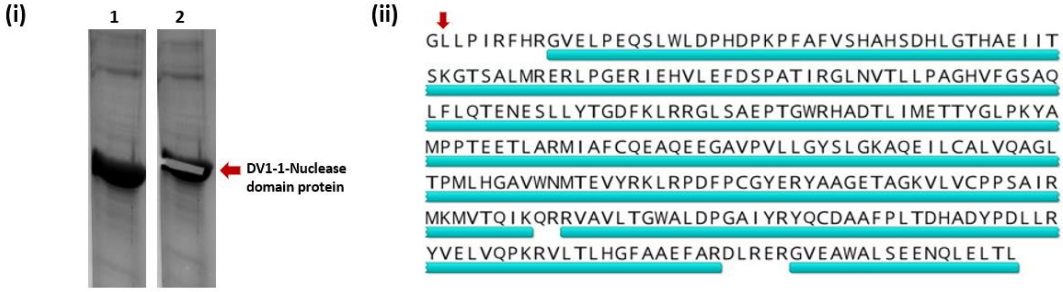


Figure 4.33. Mass spectrometry results for DV1-1-Nuclease domain protein. (i) 12 % SDS PAGE of DV1-1-Nuclease domain protein, lane 1 represents up concentrated DV1-1-Nuclease domain protein band (36 kDa), lane 2 represents the excised band, from SDS PAGE, sent for mass spectrometry. (ii) Sequence coverage of DV1-1-Nuclease. Blue bars indicate sequence coverage of DV1-1-Nuclease (95 %). A red arrow indicates the position of the annotated translational start site.

Removal of MBP-tag from DV-1-1-Nuc protein had varying levels of success from batch-to-batch, as often removal of this tag caused the protein to precipitate. Therefore, some of the biological characterisation experiments were performed with MBP-tagged protein, which did decrease its activity in comparison to un-tagged protein. Purifications for DV-1-1-Nuc with the MBP tag can be found in **Appendix C.4.1**. To allow direct comparison of DV-1-1-Nuc activity to the DV-Nuc active site mutant, an MBP-tagged version of the mutant was also purified and tested **Appendix C.4.4**.

4.2.10 Design of DV-1-1-Nuc mutant

To ensure any nuclease activity observed in activity assays was coming from DV-1-1-Nuc protein and not contaminants, a mutant protein was needed as a control. Following examples of mutant designs of MBL-β-CASP proteins, from the literature (Yosaatmadja et al., 2021) and (Silva et al., 2011), a double mutant

was designed for DV-1-1-Nuc (D36A-H37A) in an attempt to disrupt the metal binding site and inhibit nucleic acid catalysis (**Figure 4.34**). This mutation was made within Motif II, which is well conserved among other MBL-β-CASP nucleases, as described in **Section 4.2.2**.

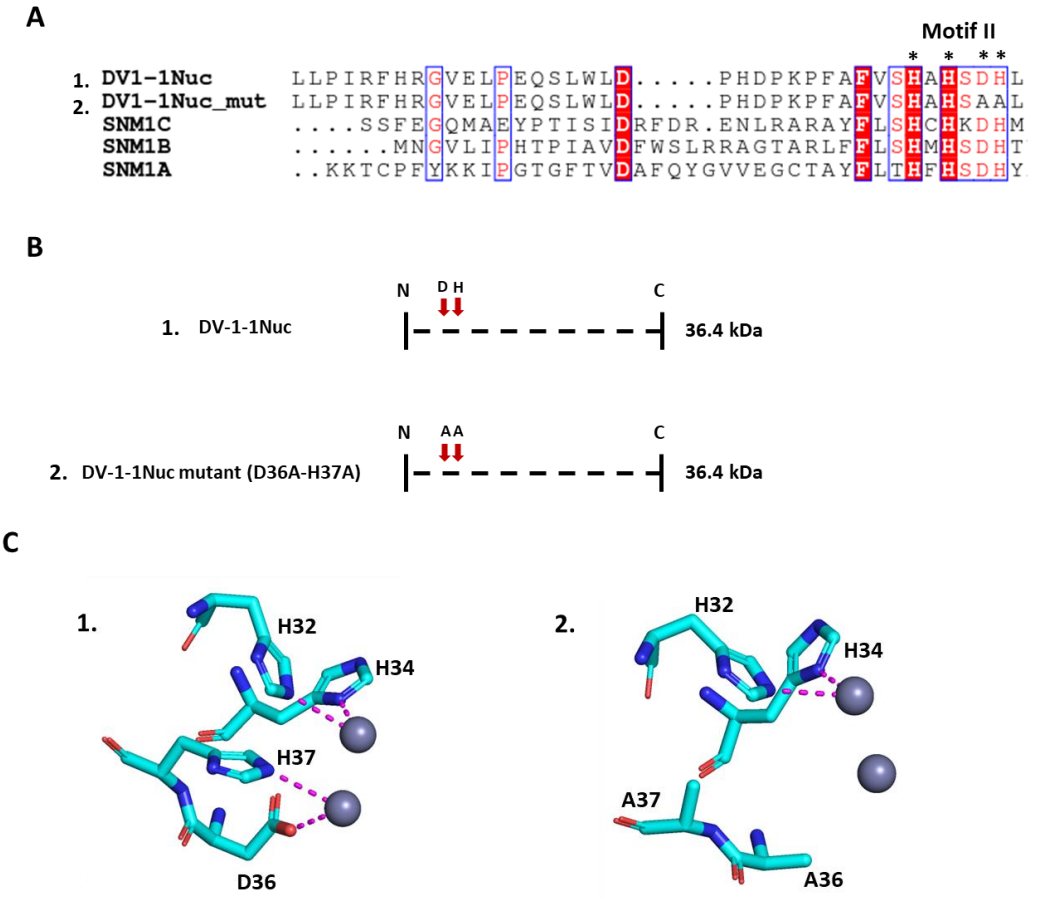


Figure 4.34. Double mutant design of DV-1-1-Nuc (D36A-H37A). **A**) Multiple sequence alignment of DV-1-1-Nuc (1.) and DV-1-1-Nuc mutant (2.), with SNM1A, SNM1B and SNM1C. Sequence alignment was created using Clustal Omega version 1.2.4. **Figure 4.23**, **A**, was created using ESPript 3 (Robert & Gouet, 2014). Motif II is indicated on graph by *. **B**) Schematic of amino acid change for DV-1-1-Nuc mutant. **C**) Pymol (Schrödinger, 2020) images of AlphaFold predicted DV-1-1-Nuc (1.) with residues Asp-36 and His-37, showing polar interactions with a zinc metal ion in the active site, superimposed onto structure from SNM1C crystal structure (7ABS) in comparison to DV-1-1-Nuc mutant (2.) with residues mutated to Ala-36 and Ala-37.

4.2.10.1 Large scale purification of DV1-1-Nuc mutant

Following on from successful small scale protein expression of the DV-1-1-Nuc mutant, in from the pHMGWA plasmid, expressed in *E. coli* (DE3) Origami (**Appendix C.4.3**), protein expression cultures were scaled up following methods from **Section 2.3.3**. A three-step purification via IMAC, reverse IMAC

and gel filtration chromatography (**Section 2.4**), produced soluble, folded protein, suitable for characterisation experiments (the chromatograms and corresponding SDS-PAGE gels in **Appendix C.4.3** depict the purification, column load and flow- through fractions).

4.2.11 Protein folding and stability of DV-1-1-Nuc

The folded structure of DV-1-1-Nuc protein was investigated using circular dichroism. Secondary structure predictions from both CD spectra and PDBsum analysis of the DV-1-1-Nuc AlphaFold model were similar with only slight differences between helix and other percentage contributions (**Figure 4.35**). These results provide additional confidence in the accuracy of the AlphaFold predicted structure and confirm that the purified protein is folded.

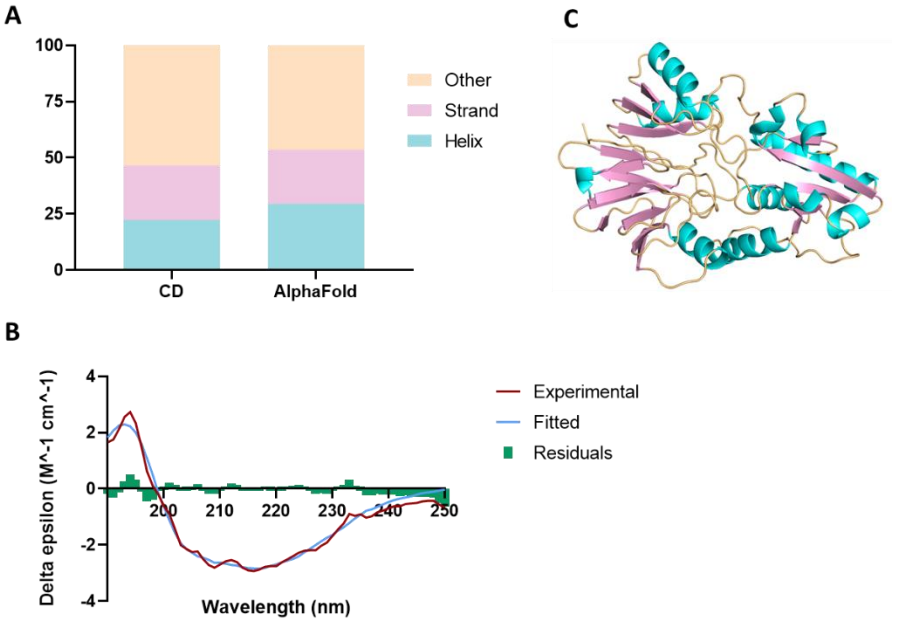


Figure 4.35. Circular dichroism (CD) and AlphaFold secondary structural composition of DV-1-1-Nuc protein. **A**) A graph showing comparison of secondary structural predictions from CD and AlphaFold prediction model. **B**) Single spectrum analysis of CD spectra, using BeStSel database (Micsonai et al., 2018). **C**) AlphaFold 3D structural prediction of DV-1-1-Nuc, coloured based on secondary structure (Helix in blue, strand in pink and other orange. (John Jumper, 2021). Graphs were produced using Prism version 8 (GraphPadSoftware). Wavelength range (190-250 nm) and scale factor (1). RMSD value (0.222). NRMSD value (0.03923).

Differential scanning fluorimetry (DSF) (**Section 2.8**) was used to compare the thermal stabilities of DV-1-1-Nuc wild-type and DV-1-1-Nuc mutant. The effect of metals (magnesium, manganese, zinc) and pH on T_m were also investigated. The following **Figure 4.6** shows the melt trace for both DV-1-1-

Nuc and DV-1-1-Nuc mutant, with the maximal T_m indicated on the melt trace or plotted on a separate graph. DV-1-1-Nuc and DV-1-1-Nuc mutant give similar T_m values, in standard conditions indicating that the introduced mutations did not negatively impact the overall stability (**Figure 4.36**, **A**). With the addition of manganese, the T_m increased for both proteins, while magnesium decreased the T_m in samples with DV-1-1-Nuc mutant (**Figure 4.36**, **B**). Melts with the addition of zinc, were attempted but did not produce suitable results as the zinc precipitated in the reactions. The thermal stability of the mutant varied with pH and was highest at lower pHs with a maximum stability of at pH 6.0 (**Figure 4.36**, **C**). No pH DSF was carried out with DV-1-1-Nuc, due to limited protein production of un-tagged DV-1-1-Nuc.

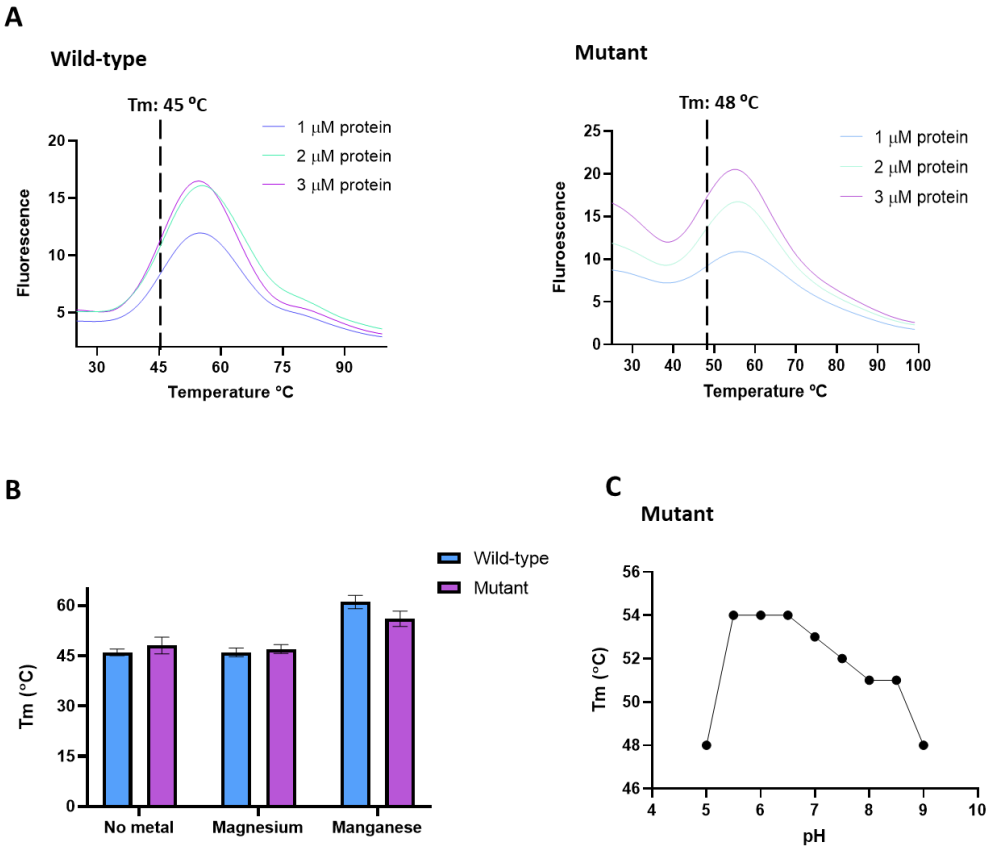


Figure 4.36. Results of Differential scanning fluorimetry (DSF), with SYPRO orange, with DV-1-1-Nuc wild-type and DV-1-1-Nuc mutant. Results of Differential scanning fluorimetry (DSF), with SYPRO orange, with DV-1-1-Nuc wild-type and DV-1-1-Nuc mutant. **A**) DSF with three different concentrations (1, 2 & 3 μM) of DV-1-1-Nuc wild-type and DV-1-1-Nuc mutant protein. T_m values were determined from the midpoint in the unfolding transition and are indicated on the graph, by a dotted line. Each concentration was carried out in triplicates. **B**) DSF with additions of magnesium and manganese metal ions to reactions with DV-1-1-Nuc wild-type and DV-1-1-Nuc mutant protein. Reactions for each metal ion group were carried out in replicates of three. T_m values were determined as above and are indicated in bar graphs. Proteins were at a final concentration of 1.5 μM. **C**) DSF thermal melts with DV-1-1-Nuc mutant, where reactions are of a different pH value. Reactions were carried out in replicates of three, for each pH value. T_m values were determined as

above and are indicated in a plot graph. Protein was at a final concentration of 2.5 μM. Graphs were generated using GraphPad Prism version 8 (GraphPadSoftware).

4.2.12 Biochemical characterisation of DV-1-1-Nuc

Initial biochemical activity assays were performed with untagged DV-1-1-Nuc, however due to difficulties in removal of the MBP tag from the protein, as described in **Section 4.2.9.4**, some of the remaining assays were performed with DV-1-1-Nuc$_{MBP}$ (with MBP tag) to complete biochemical activity characterisation. Because the MBP tag was left on the protein for some of the experiments, comparison assays with DV-1-1-Nuc mutant, were also completed with a MBP tagged version of the protein. Final comparison activity assays were performed on some of the DNA substrates, to highlight activity differences, firstly, between wild-type DV-1-1-Nuc and mutant DV-1-1-Nuc and secondly, between tagged and untagged versions of both wild-type and mutant DV-1-1-Nuc.

The following section details the binding ability of DV-1-1-Nuc to DNA substrates in electrophoretic mobility shift assays and describes its nuclease activity on un-modified as well as damaged DNA substrates, under a range of different conditions, using gel-based activity assays.

4.2.12.1 DNA binding by DV-1-1-Nuc

DV-1-1-Nuc protein was used in a DNA binding experiment (EMSA) with double stranded (Ds), single stranded (Ss) and Abasic (dSpacer) DNA substrates. Incubation of the DNA substrates with DV-1-1-Nuc resulted in a small shift and smearing of the band relative to no-protein controls, however this was smaller than expected for a bound complex. It is possible that there was some binding of the DNA substrate, by protein, but it has not stayed bound, hence the slight shift in size of these bands above the substrate (**Figure 4.37**). No band shift was seen with Ss DNA substrates (data not shown).

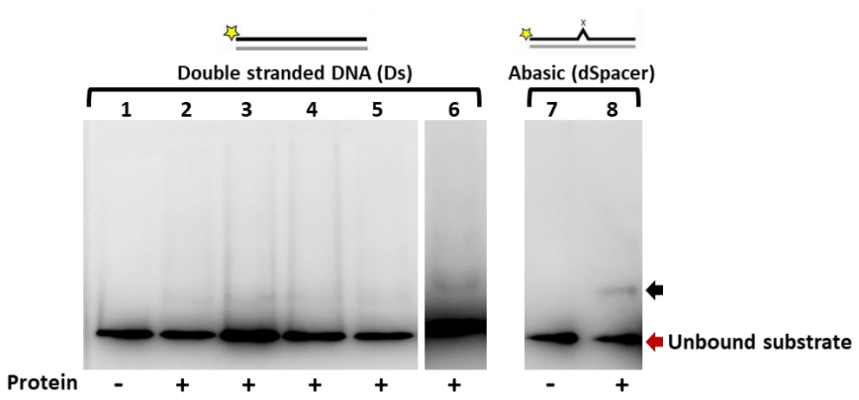


Figure 4.37. Electrophoretic mobility shift assay (EMSA) of DV-1-1-Nuclease domain protein with DNA substrates, run on a native TBE gel. Lanes 1 and 7 don't contain any protein (-). Lanes 2-6, contain double stranded (Ds) DNA, with 1 µM DV-1-1-Nuclease protein and represent a time series from 20 minutes (2), 30 minutes (3), 45 minutes (4) and 50 minutes (5). Lane 6 contains protein incubated with zinc, for 30 minutes. Lane 8 contains abasic (dSpacer) damaged DNA substrate, with 1 µM DV-1-1-Nuclease protein, incubated for 30 minutes, with zinc. EDTA was included in all reactions, except for lanes 6 and 8. All reactions were incubated at 20 ^{o}C. A Red arrow indicates unbound substrates, while the black arrow indicates DNA bands above the unbound substrates. Results of EMSA were visualized using iBright™ CL750 Imaging System, Invitrogen™.

4.2.12.2 Activity of DV-1-1-Nuc on un-modified DNA

Initial activity of DV-1-1-Nuc protein was tested on un-modified DNA substrates (double stranded (Ds), single stranded (Ss), 20 mer single stranded, 3'-tail and 5'-tail), to determine if the protein had specificity towards a certain type of DNA substrate and whether the nuclease activity on those substrates was specific or non-specific. The following **Figure 4.38** represents a schematic of a typical activity assay setup, with results assessed by TBE urea PAGE (**A**) and the un-modified DNA substrates to be tested (**B**).

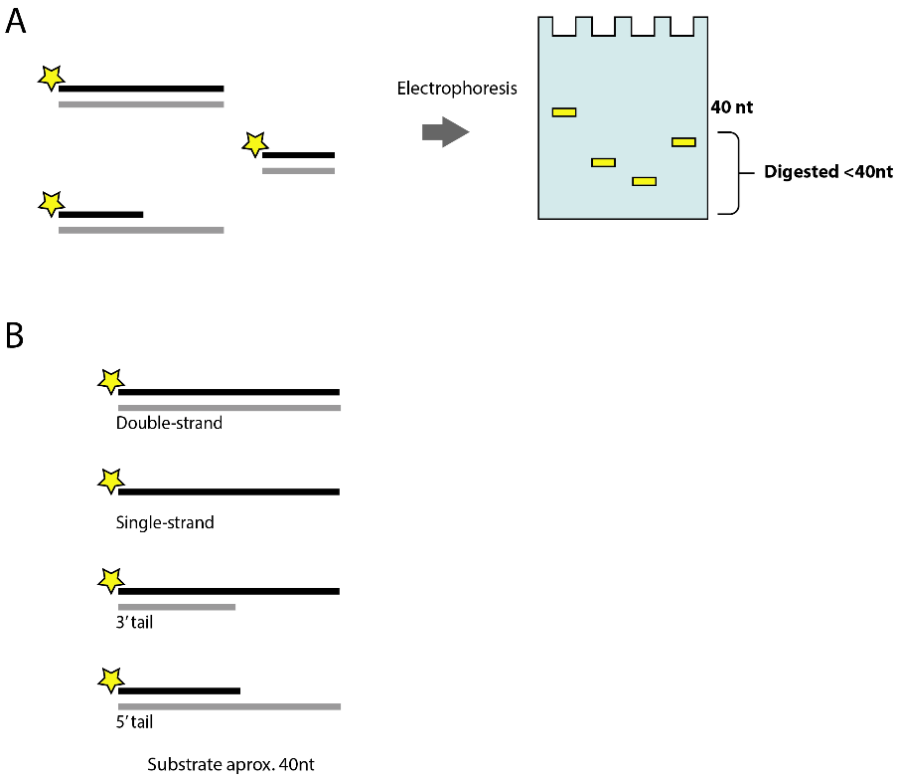
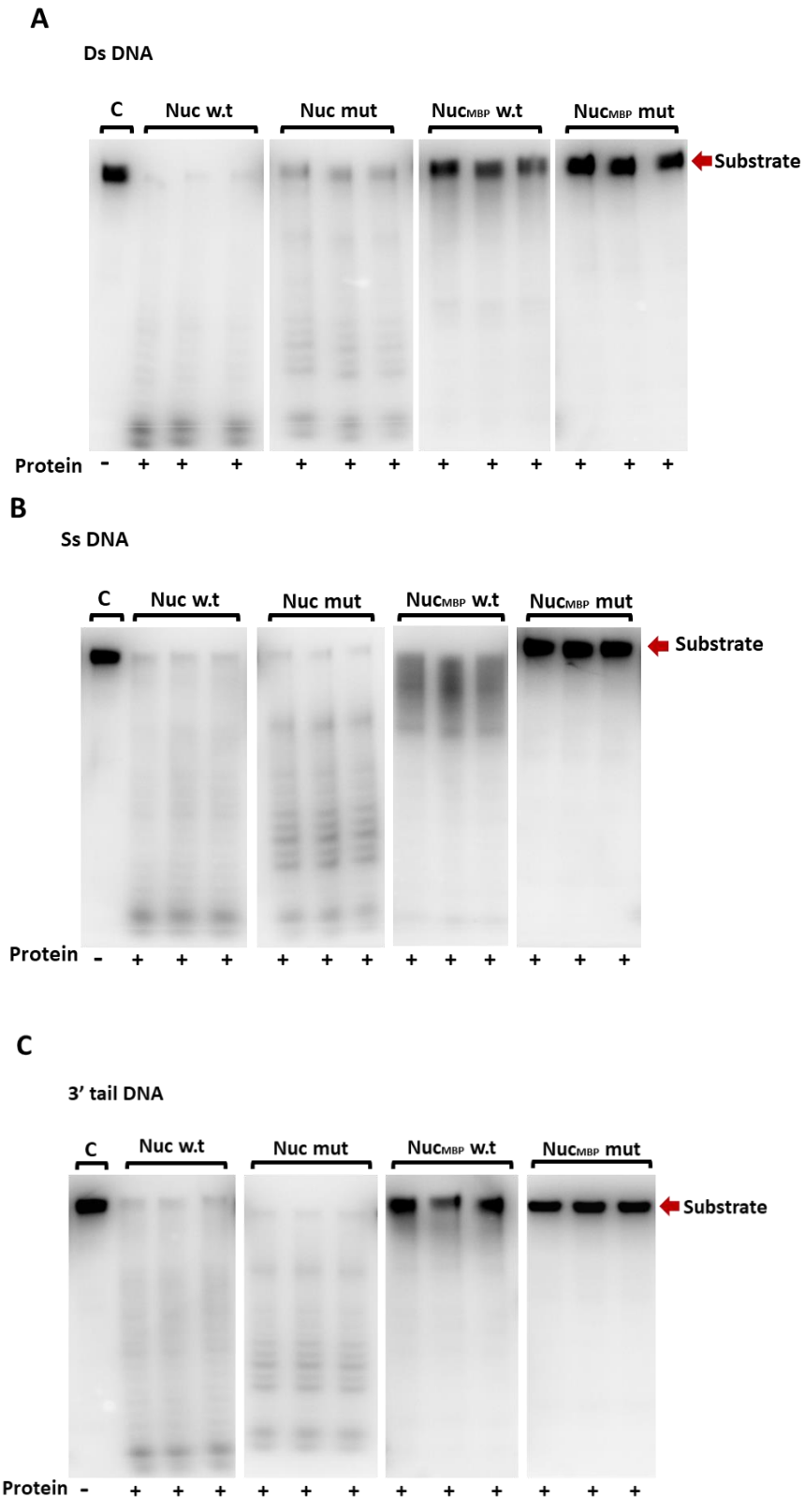


Figure 4.38. Schematic of enzyme assays for nuclease activity on DNA substrates with carrying single-stranded portions. Stars represent labelling with the 6-carboxyfluorescein at the 5' terminus (5'FAM). Labelled strands are indicated by a black line while unlabelled portions of substrate duplexes are not visible during analysis are indicated by grey lines. **A**) Analysis of assay products by urea PAGE indicating a size-shift based on degradation of any part of the duplex (yellow boxes). **B**) Design of substrates with double and single-stranded portions.

Nuclease activity is observed on all four DNA substrates (Ds, Ss 3' tail and 5'tail DNA). DV-1-1-Nuc protein is more active on 5'-tail DNA in comparison to 3'-tail DNA, with 5'-tail DNA almost completed degraded with the addition of DV-1-1-Nuc protein. Both wild-type (w.t) and mutant (mut) DV-1-1-Nuc showed nuclease activity on Ds and Ss DNA, although the extent was greater for the wild-type enzyme. This suggests that although mutation of the putative Zn binding site residues decreased activity, it did not completely abolish it. A greater impact was seen for both wild-type and mutant, from leaving the MBP tag on (**Figure 4.39**).


A
Ds DNA
C
Nuc w.t
Nuc mut
NucMBP w.t
NucMBP mut
Substrate
Protein - + + + + + + + + + + +
B
Ss DNA
C
Nuc w.t
Nuc mut
NucMBP w.t
NucMBP mut
Substrate
Protein - + + + + + + + + + + +
C
3' tail DNA
C
Nuc w.t
Nuc mut
NucMBP w.t
NucMBP mut
Substrate
Protein - + + + + + + + + + + +

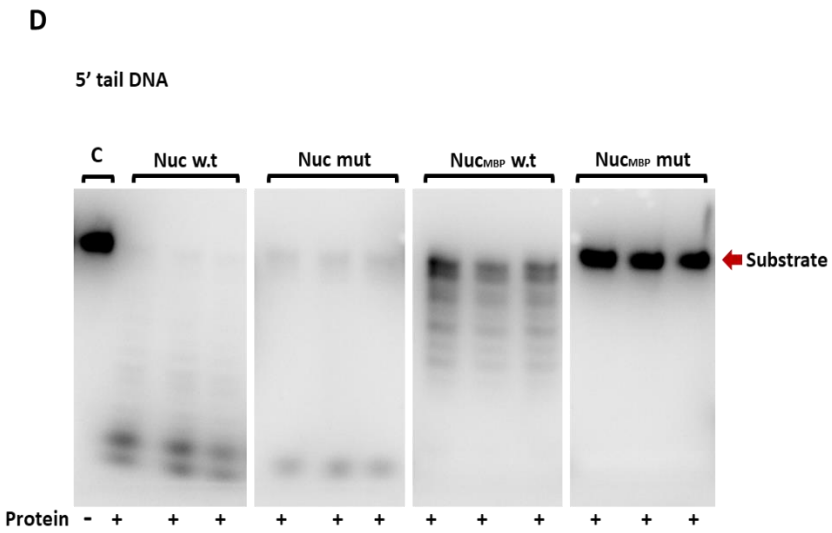
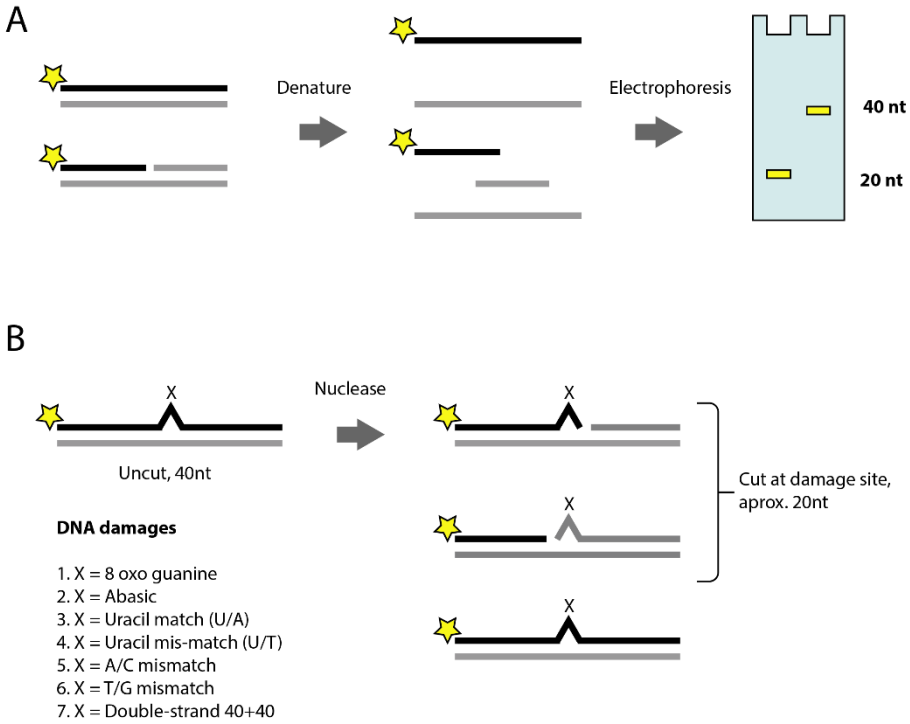


Figure 4.39. TBE urea PAGEs show results of nuclease activity by DV-1-1-Nuclease domain protein (Nuc w.t and NucMBP w.t) and DV-1-1-Nuclease mutant protein (Nuc mut and Nuc_{MBP} mut) on: double stranded (Ds), single stranded (Ss), 3'-tail and 5'-tail DNA substrates. **A**) represents activity on Ds DNA substrates. **B**) represents activity on Ss DNA substrates. **C**) represents activity on 3'-tail DNA substrates and **D**) represents activity on 5'-tail DNA substrates. Control lanes with no protein (-) have a C, symbol above them. All four proteins have 3 replicates for each DNA substrate. Substrates are indicated by red arrows. Reactions were carried out at 20 ^{o}C, for 8 hours, with 1.2 µM final protein concentration and 10 mM final of magnesium in all reactions. Results of activity assays were visualized using iBright™ CL750 Imaging System, Invitrogen™.

4.2.12.3 Activity of DV-1-1-Nuc on damaged or mis-matched DNA duplexes

Following the activity results of DV-1-1-Nuc on un-modified DNA substrates, further activity assays were performed with several damaged or mis-matched DNA substrates to determine if the protein acted on a certain type of DNA damage and whether the nuclease activity on those substrates was specific or nonspecific. Substrates tested included a centrally-placed 8 oxo guanine, abasic (dSpacer), uracil match (A/U), uracil mis-match (U/T), A/C mismatch and T/G mismatch. The following **Figure 4.40** represents a schematic of a typical activity assay setup, with results run on a TBE urea PAGEs and the damaged or mis-matched DNA substrates to be tested.



Types of DNA damages:
8 oxo dG: Marker of oxidative stress
Abasic/dSpacer: Analog of an abasic site
Uracil match (U/A): Result of mis-incorporation
Uracil mis-match (U/T): Pyrimidine-pyrimidine
Regular mis-match (A/C): Purine-pyrimidine mis-match
Regular mis-match (T/G): Purine-pyrimidine mis-match

Figure 4.40. Schematic of enzyme assays for nuclease activity on damage/mismatch DNA substrates activity using fluorescently labelled oligonucleotide substrates. Stars represent labelling with the 6-carboxyfluorescein at the 5' terminus (5'FAM). Labelled strands are indicated by a black line while unlabelled portions of substrate duplexes are not visible during analysis are indicated by grey lines. **A**) Analysis of assay products by denaturing TBE-urea PAGE indicating separation of oligonucleotide strands and electrophoretic detection of the labelled strands on the gel (yellow boxes). **B**) Design of double-stranded substrates incorporating damaged bases or mismatches at a central position and the predicted outcomes of endonuclease activity. **C**) Schematic showing chemical modification of damage and table description of what the damages/mis-matches refer to.

Activity on damaged and mismatched DNA substrates was initially tested with just DV-1-1-Nuc, to determine what substrates it was most active on. There is obvious nuclease activity on abasic and uracil mismatch DNA substrates, with a very specific single band observed in reactions with abasic and less specific banding observed in reactions with uracil mismatch. No nuclease activity was observed with 8-Oxo-dG, uracil match, or regular mismatch T/G DNA substrates. 8-Oxo-dG substrate shows residual low molecular weight bands, across all reactions, including the control, indicating that this band was not produced by nuclease activity from DV-1-1-Nuc. There is some smearing in reactions containing protein and regular mismatch T/G DNA substrate, however it is difficult to conclude that this is due to nuclease activity from DV-1-1-Nuc protein (**Figure 4.41**).

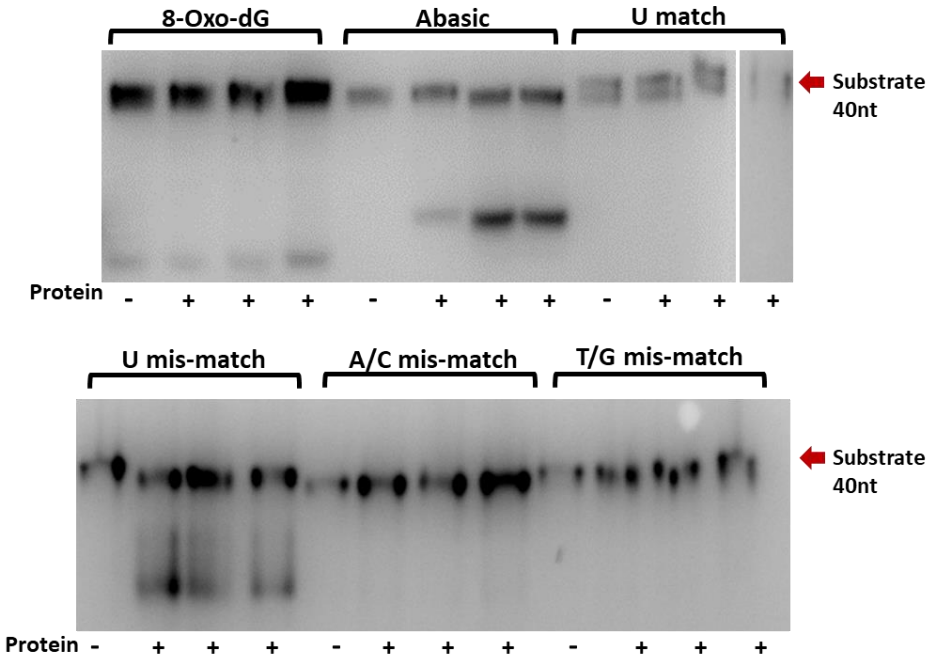


Figure 4.41. TBE Urea PAGEs showing DV-1-1-Nuc protein activity on damaged and mis-matched DNA substrates. Reactions containing protein are indicated by a plus (+) and ones with no protein, by (-). Reactions were run for 4 hours at 15 ^{o}C. Protein was at a final concentration of 1.8 µM and magnesium at 10 mM. Substrates are fluorescently labelled, and these bands were visualized using the iBright™ CL750 Imaging System, Invitrogen™.

Abasic and mismatch DNA substrates were used in final comparison activity assays with, DV-1-1-Nuc wild-type, DV-1-1-Nuc mutant and MBP-tagged versions of both proteins, with magnesium as a metal ion cofactor. In reactions with abasic and mismatch DNA substrates, nuclease activity was seen

with the addition of DV-1-1-Nuc wild-type, DV-1-1-Nuc mutant and MBP-tagged DV-1-1-Nuc. No activity was seen in reactions containing MBP-tagged DV-1-1-Nuc mutant. Activity was reduced in reactions containing mutant protein, in comparison to wildtype. A 20 nt marker (N) was also run on the gel, alongside activity reactions, to determine if DV-1-1-Nuc exhibits specific endonuclease activity at the damage site, which would generate a 20 nt product. Reactions containing MBP-tagged DV-1-1-Nuc protein, show products beginning at a 20 nt position indicating an initial specific cut could be made here, followed by more general exonucleolytic degradation (**Figure 4.42**).

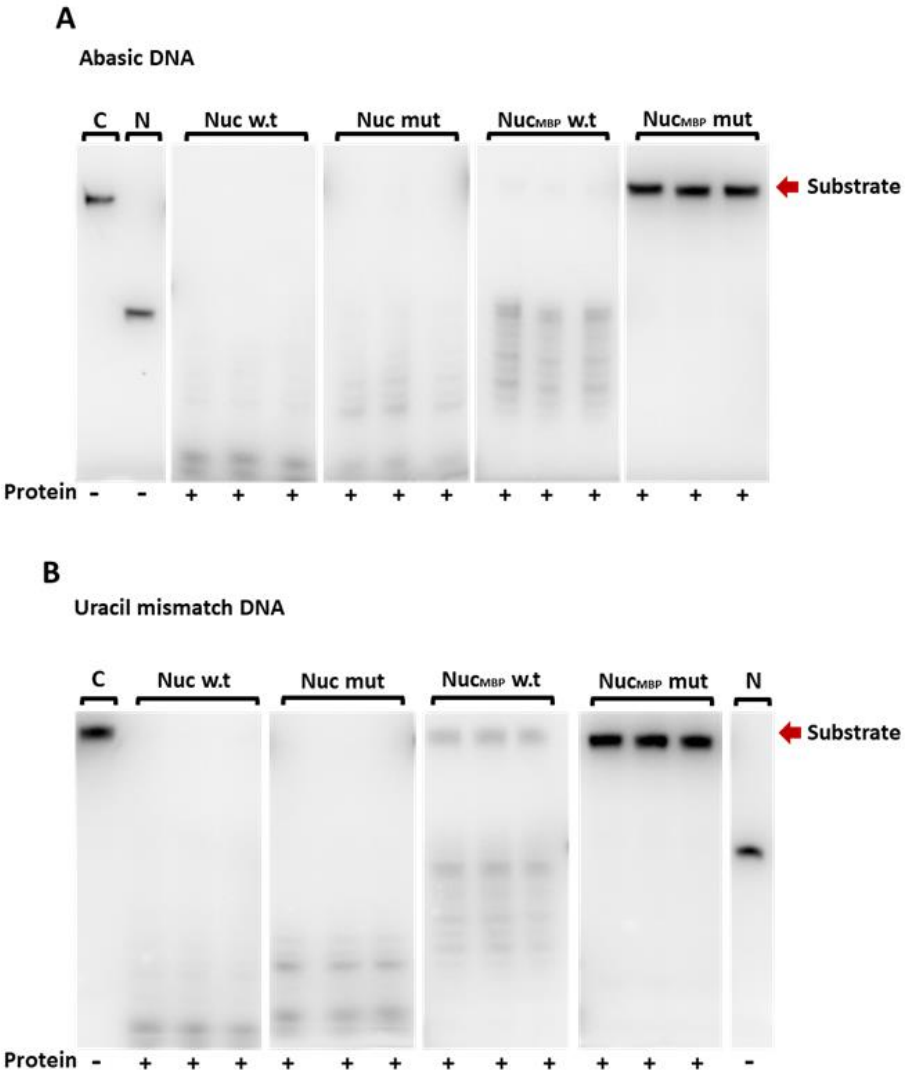


Figure 4.42. TBE urea PAGEs show results of nuclease activity by DV-1-1-Nuclease domain protein (Nuc w.t and NucMBP w.t) and DV-1-1-Nuclease mutant protein (Nuc mut and Nuc_{MBP} mut) on: abasic and uracil mis-match DNA substrates. **A**) represents activity on abasic DNA substrates. **B**) represents activity on uracil mis-match DNA substrates. Control lanes with no protein (-) have a C, symbol above them. A 20 nt marker (N), with no protein (-) was included in the gel. All four proteins have 3 replicates for each DNA substrate. Substrates are indicated by red arrows. Reactions were carried out at 20 ^{o}C, for 8 hours, with 1.2 µM final

protein concentration and 10 mM final of magnesium in all reactions. Results of activity assays were visualized using iBright™ CL750 Imaging System, Invitrogen™.

In summary, DV-1-1-Nuc wild-type protein shows nuclease activity on abasic and uracil mis-match DNA substrates. Reactions with shorter incubation times (4 hours), produced a single band product on the gel, in comparison to longer incubation times (8 hours), that produced multiple bands on the gel. This indicates that the protein specifically cuts the DNA substrate, at a specific position and then continues to cut showing a non-specific pattern. While DV-1-1-Nuc mutant still show nuclease activity on DNA substrates, this activity is reduced compared to wild-type.

4.2.12.4 Activity of DV-1-1-Nuc on flapped and splayed DNA substrates

Alongside un-modified, damage and mis-match DNA substrates, the nuclease activity of DV-1-1-Nuc was also investigated with flapped (3’ Flap, 5’ Flap) and splayed DNA substrates (**Figure 4.43**). These substrates were designed to mimic biological flaps that can form after DNA polymerases, with strand displacement activity, have replicated past a damage. If the nuclease domain showed activity on these DNA substrates, then it might suggest a role for the protein in a DNA repair pathway that requires a nuclease to cleave flaps produced by a DNA polymerase.

These three DNA substrates were used in comparison activity assays with DV-1-1-Nuc wild-type, DV-1-1-Nuc mutant and MBP-tagged versions of both proteins and magnesium metal ion as a cofactor.

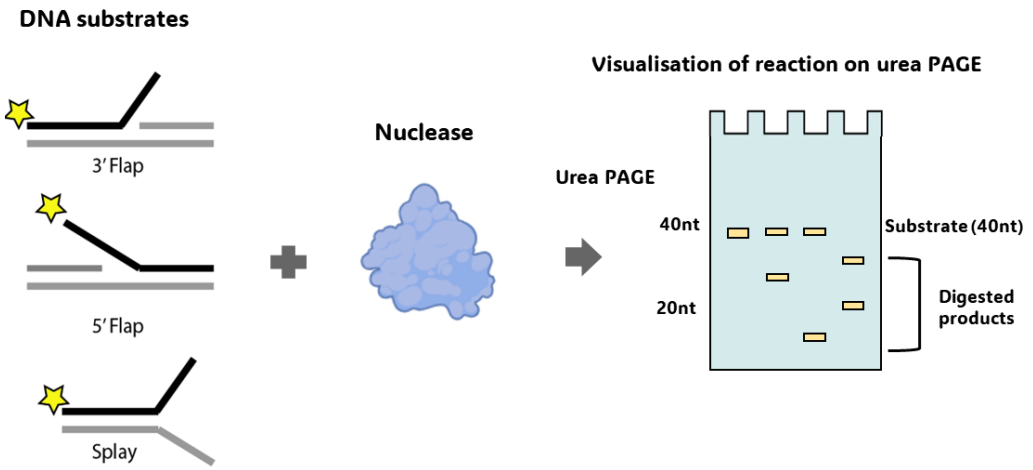
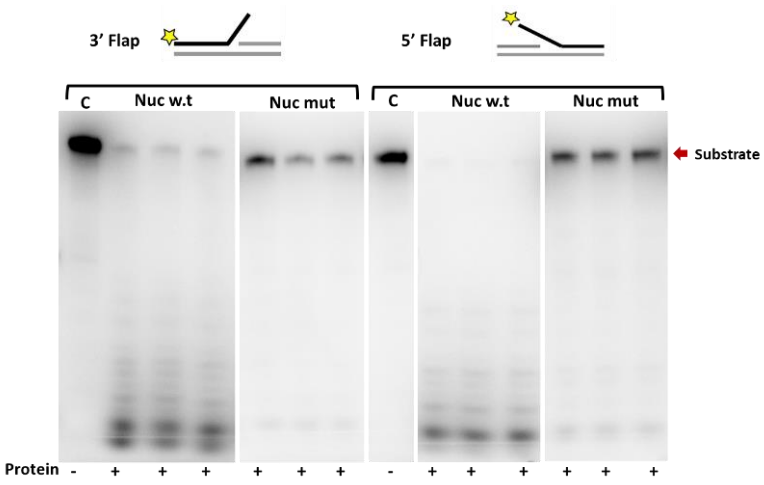


Figure 4.43. Schematic of enzyme assays for nuclease activity on DNA substrates with 5′ or 3′ flaps and splayed substrates. Stars represent labelling with the 6-carboxyfluorescein at the 5′ terminus (5′FAM). Labelled strands are indicated by a black line while unlabelled portions of substrate duplexes are not visible during analysis are indicated by grey lines. Analysis of assay products by denaturing urea PAGE indicating a size-shift based on degradation of any part of the duplex (yellow boxes).

Results of these activity assay indicate DV-1-1-Nuc wild-type is active on all three DNA substrates, with the best activity seen on 5’ Flap DNA substrate. DV-1-1-Nuc mutant is less active on these DNA substrates, compared to wild-type. In summary DV-1-1-Nuc wild-type protein shows nuclease activity towards flapped and splayed DNA substrates, with best activity ordered by 5’ Flap, 3’ Flap and Splayed. DV-1-1-Nuc mutant has minimal activity on these DNA substrates, compared to wild-type, which indicates that the mutated residues in DV-1-1-Nuc mutant (D36A-H37A) may be important for activity on these type of DNA substrates (**Figure 4.44**).



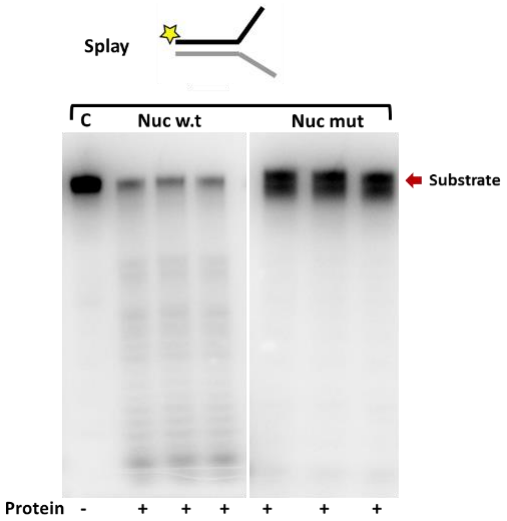


Figure 4.44. TBE urea PAGEs show results of nuclease activity by wild-type DV-1-1-Nuc and mutant DV-1-1-Nuclease protein on flapped (3' Flap, 5' Flap) and splayed DNA substrates. Control lanes with no protein (-) have a C, symbol above them. Reactions are carried out triplicate, for each DNA substrate. Substrates are indicated by red arrows and diagrams of each DNA substrate is shown above each gel. Reactions were carried out at 20 ^{o}C, for 8 hours, with 1.2 µM final protein concentration and 10 mM final of magnesium in all reactions. Results of activity assays were visualized using iBright™ CL750 Imaging System, Invitrogen™.

4.2.12.5 Metal ion preference of DV-1-1-Nuc

Activity assays testing the DNA substrate preference of DV-1-1-Nuc contained magnesium as the metal ion cofactor, which was necessary for nuclease activity on DNA substrates. To determine metal ion cofactor specificity of DV-1-1-Nuc, activity was tested with a range of divalent metal ion additives known to be used by other nucleases.

Un-modified Ds and Ss DNA substrates were tested first with DV-1-1-Nuc, with magnesium, manganese, and zinc metal ions. Activity on Ds DNA is supported by the addition of magnesium, manganese, and zinc. Activity with magnesium and manganese, shows a similar degree of nonspecific activity, with slightly higher activity with manganese. Activity with the addition of zinc, shows a single band product. Activity on Ss DNA is supported with the addition of magnesium and manganese and these reactions show a similar degree of non-specific activity, again with slightly more activity in reactions containing manganese. No activity is observed in reactions containing zinc. For both Ds and Ss activity assays, no activity was observed in reactions without a metal cofactor or in reactions with the addition of EDTA and magnesium (**Figure 4.45**).

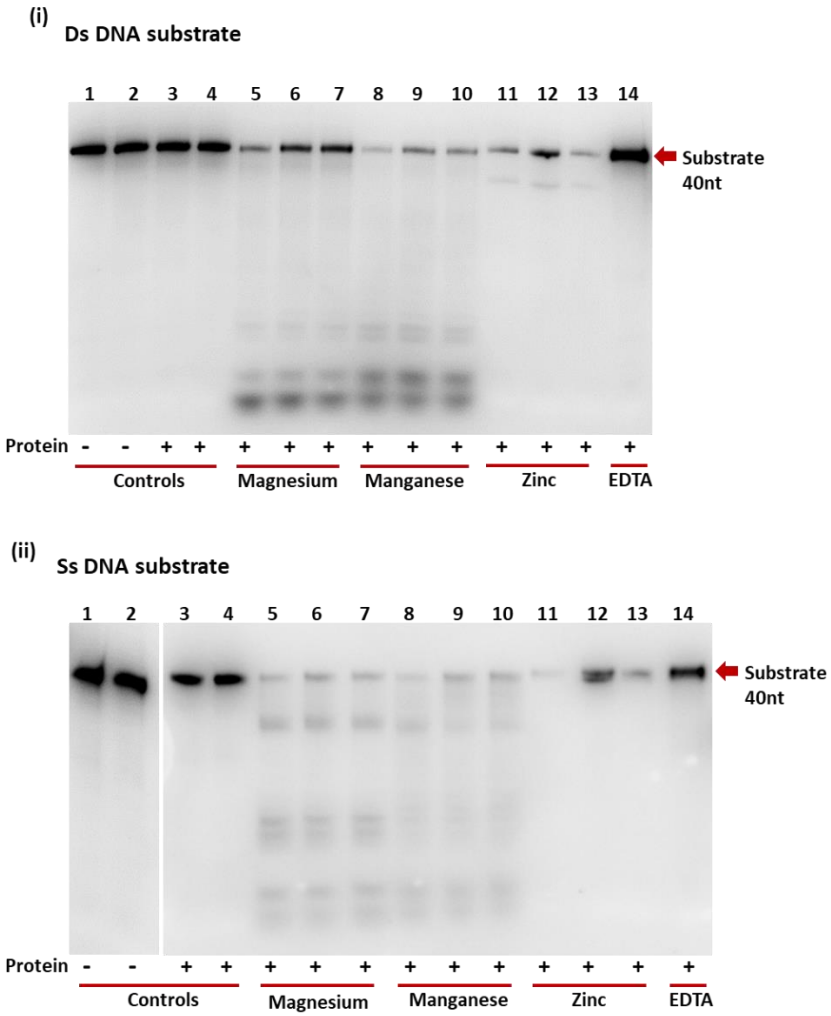


Figure 4.45. TBE urea PAGEs show results of nuclease activity by wild-type DV-1-1-Nuc on Ds and Ss DNA substrates, with different metal cofactors (magnesium, manganese & zinc). Control lanes contain DNA substrate with no protein (-) and DNA substrate with protein (+) and no cofactor. EDTA reactions contain protein (+), 10 mM magnesium and 40 mM EDTA. Reactions are carried out in replicates of 3, for each metal ion cofactor. Substrates are indicated by red arrows. Reactions were carried out at 20 ^{o}C, for 4 hours, with 1.5 µM final protein concentration and 10 mM final of metal ion in all reactions. Results of activity assays were visualized using iBright™ CL750 Imaging System, Invitrogen™.

Following activity assays with un-modified DNA substrates, further activity assays were performed with DV-1-1-Nuc wild-type, on damaged (abasic) and mis-matched (uracil mis-match, A/C mis-match and T/G mis-match) DNA substrates using the three metal ion cofactors, from above. MBP-tagged DV-1-1-Nuc protein was used for these reactions, due to low protein stock of un-tagged DV-1-1-Nuc.

In reactions with abasic DNA substrate, nuclease activity was seen in reactions containing magnesium (Mg), and manganese (Mn). No nuclease activity was observed in reactions containing zinc (Zn). In reactions with uracil mis-match

(U mis-match) DNA substrate, there was some nuclease activity observed with Mg. There was also nuclease activity seen in reactions with no-cofactor, which might indicate activity from contaminating nucleases, that don't require a metal cofactor, or that a tightly-bound metal ion might be present in the active site, of DV-1-1-Nuc$_{MBP}$ wild-type. No nuclease activity, with any of the metal cofactors, was observed for A/C and T/G mis-match DNA substrates (**Figure 4.46**).

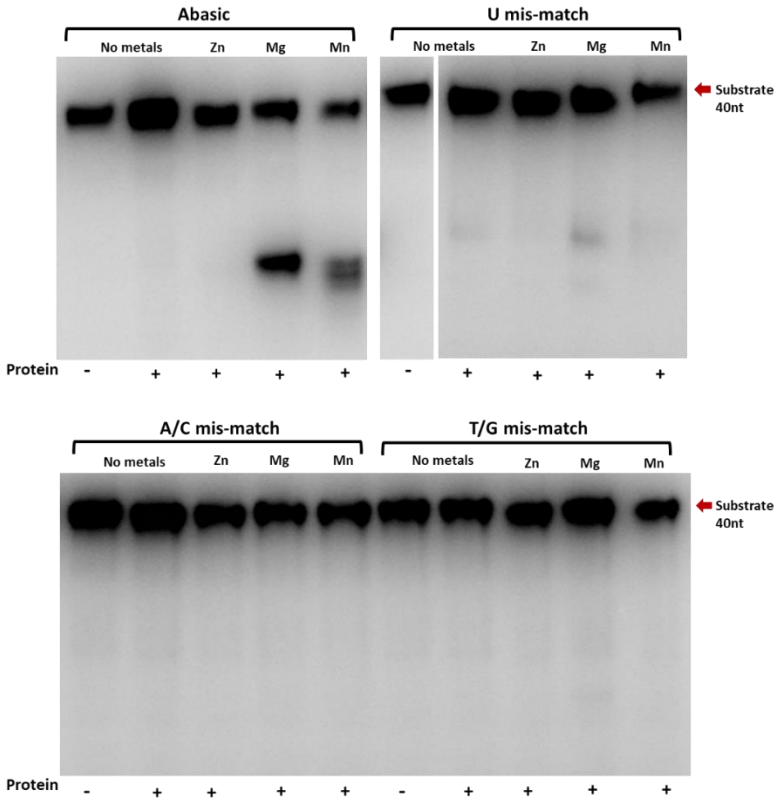


Figure 4.46. TBE urea PAGEs show results of nuclease activity by wild-type DV-1-1-NucMBP protein on damaged and mis-matched DNA substrates, with different metal cofactors. Control lanes are labelled no metal and contain no protein (-) or protein (+). Reactions with metals are indicated above the gel, with Zn (zinc), Mg (magnesium) or Mn (manganese). Substrates are indicated by red arrows (40 nt) Reactions were carried out at 20 ^{o}C, for 6 hours, with 1.2 µM final protein concentration and 10 mM final of each metal cofactor, in all reactions. Results of activity assays were visualized using iBright™ CL750 Imaging System, Invitrogen™.

Further testing with different concentrations of zinc was carried out to determine if the current zinc concentration was inhibiting nuclease activity. The activity of DV-1-1-Nuc$_{MPB}$ was tested on Ss and abasic DNA substrates with different concentrations of zinc metal ions. Reactions with Ss DNA substrate, showed that some nuclease activity does occur, but only at low zinc concentrations (0.25 and 0.5 mM). Increasing the concentration of zinc, in the reactions, appeared to inhibit nuclease activity. Testing on abasic DNA substrate

saw some nuclease activity in the no-cofactor reactions, however with the addition of low concentrations of zinc (0.15 mM) there was an increase in nuclease activity (**Figure 4.47**).

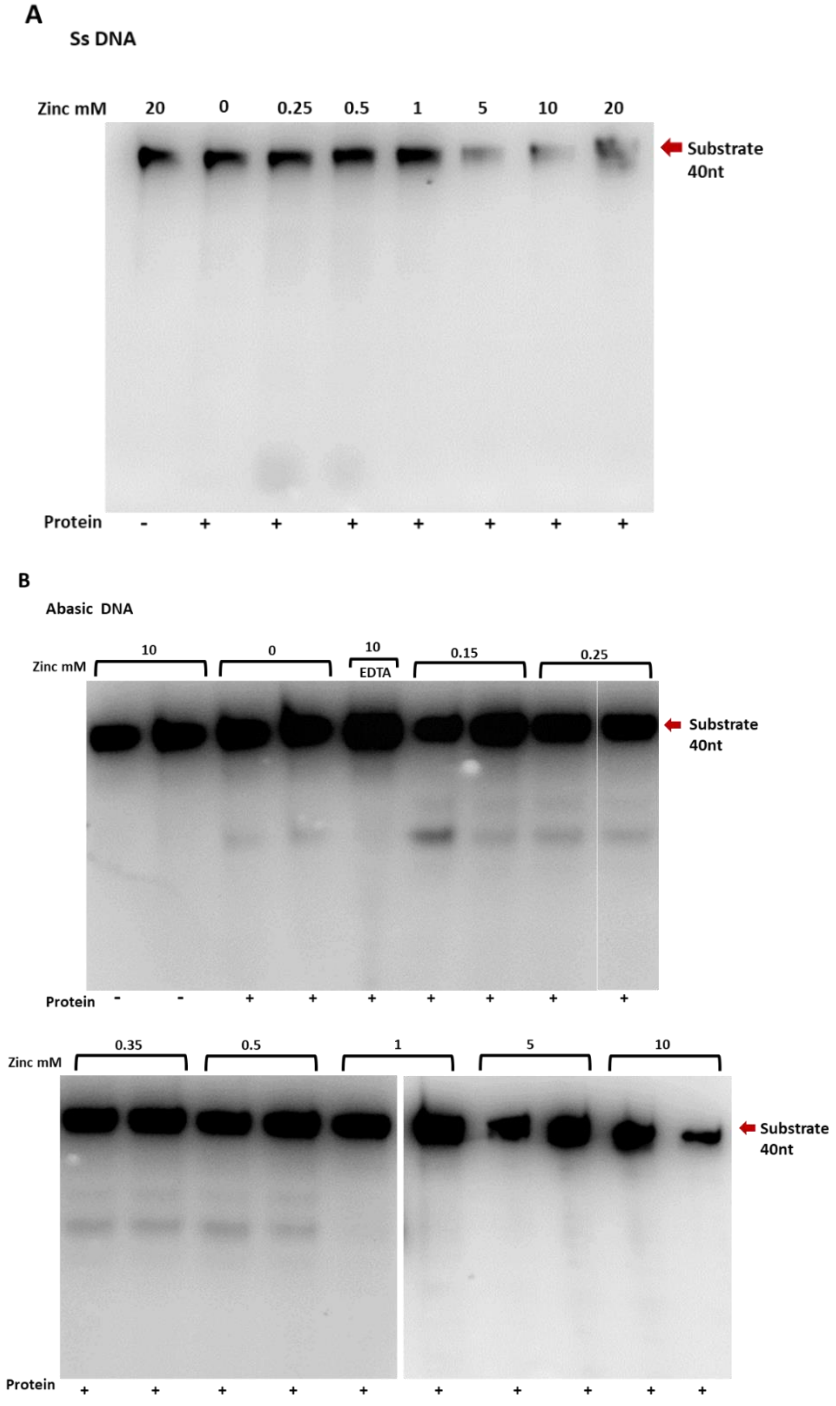


Figure 4.47. TBE urea PAGEs show results of nuclease activity by MBP-tagged DV-1-1-Nuc protein on Ss and abasic DNA substrates, with zinc at varying concentrations. **A**) Reactions with MBP-tagged DV-1-1-Nuc with Ss DNA substrate, with different zinc concentrations added to each reaction (0, 0.25, 0.5, 1, 5, 10 & 20 mM). The control lane, with no protein (-), contains 20 mM zinc. **B**) Reactions with MBP-tagged DV-1-1-Nuc with abasic DNA substrate, with different zinc concentrations (mM) added to each reaction (0, 0.15, 0.25, 0.35, 0.5, 1, 5 & 10). The control lane with no protein (-), contains 10 mM zinc. Substrates are indicated by red arrows (40 nt). Reactions were carried out in replicates of two, for each zinc concentration. Both assays contain controls with protein (+), but no metal cofactor. Reactions were carried out at 20 ^{o}C, for 6 hours,

with 1.2 µM final protein concentration and varying concentrations of zinc metal ion cofactor, in all reactions. Results of activity assays were visualized using iBright™ CL750 Imaging System, Invitrogen™.

Final comparison activity assays were performed with DV-1-1-Nuc wild-type and DV-1-1-Nuc mutant, on abasic DNA substrate, with magnesium, manganese and zinc metal ion cofactors and no cofactor controls.

Nuclease activity was observed on abasic DNA, with addition of all three metal ion cofactors. The best nuclease activity was seen in reactions containing magnesium, followed by manganese and zinc. Reactions with zinc, showed only one product band, in the gel. In comparison between DV-1-1-Nuc wild-type and DV-1-1-Nuc mutant, there was more activity seen in reactions with wildtype, except for reactions that contained zinc. No cofactor reactions showed no nuclease activity occurring, with wild-type or mutant protein (**Figure 4.48**).

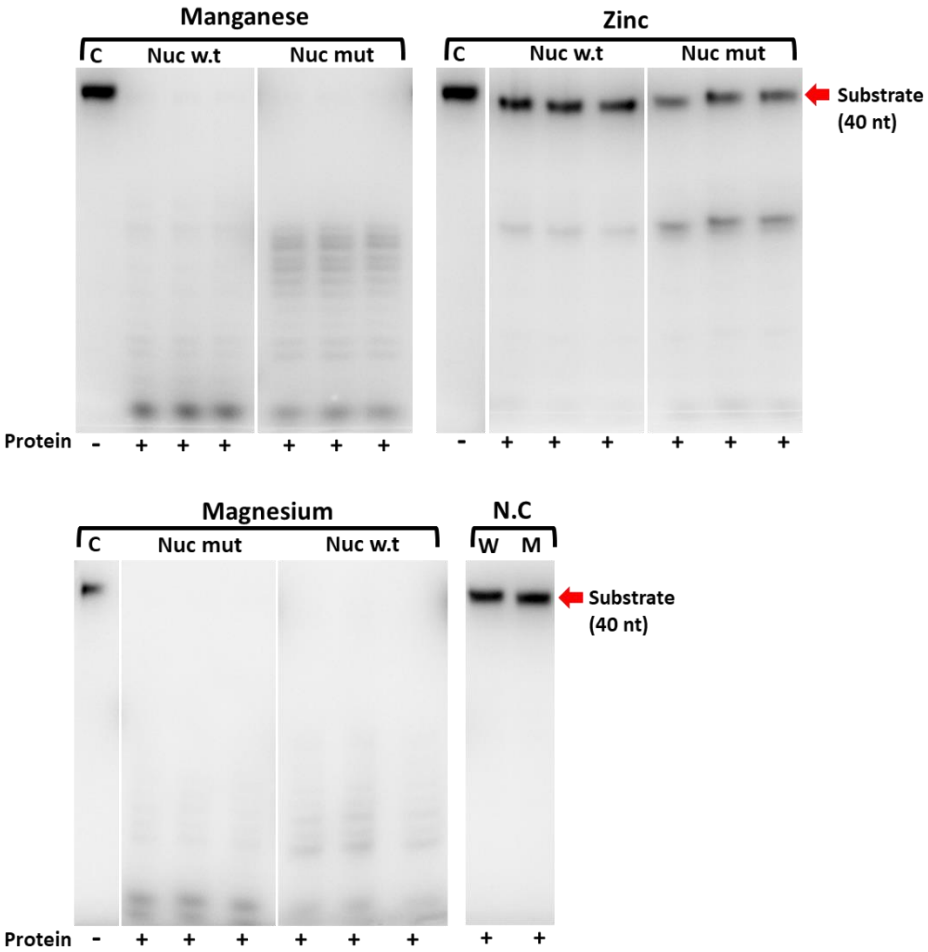


Figure 4.48. TBE urea PAGEs show results of nuclease activity by DV-1-1-Nuc wild-type and DV-1-1-Nuc mutant proteins on abasic DNA substrate, with magnesium, manganese and zinc. The control lanes (C) don't contain any protein (-). No cofactor (N.C) reactions contain protein (+), but no metal ion cofactor. Reactions were carried out in replicates of three. Substrates are indicated by red arrows (40 nt). Reactions were

carried out at 20 ^{o}C, for 8 hours, with 1.2 µM final protein concentration and 10 mM final magnesium and manganese and 0.25 mM zinc metal cofactors. Results of activity assays were visualized using iBright™ CL750 Imaging System, Invitrogen™.

In summary nuclease activity by DV-1-1-Nuc wild-type on DNA substrates is supported by the addition of magnesium, manganese, and low concentrations of zinc ion metal cofactors. DV-1-1-Nuc shows variability in its activity with the addition of magnesium and manganese, depending on the DNA substrate it is acting on. DV-1-1-Nuc is less active with the addition of zinc, compared to the other two metals and shows specific activity on abasic DNA substrate. Interestingly, the DV-1-1-Nuc mutant was slightly more active on abasic DNA with the addition of zinc, compared to DV-1-1-Nuc wild-type which is surprising given that the mutation targeted the metal binding site.

4.2.12.6 Temperature dependence of DV-1-1-Nuc activity

A series of temperature dependence activity assays were carried out on double stranded (Ds), single stranded (Ss), and abasic (dSpacer) DNA substrates with DV-1-1-Nuc protein. These DNA substrates were chosen, as previous activity assays, from above, showed that DV-1-1-Nuc protein was most active on them, compared to some of the other DNA substrates.

The following **Figure 4.49** shows the results of an activity assay on Ds and Ss DNA substrates, from 1 ºC to 15 ºC. These results show that the protein is more active on Ss DNA substrates than Ds DNA substrates, across all temperature ranges. The best nuclease activity, on both DNA substrates, is seen at 15 ºC. The nuclease cutting activity on these DNA substrates appears to be non-specific.

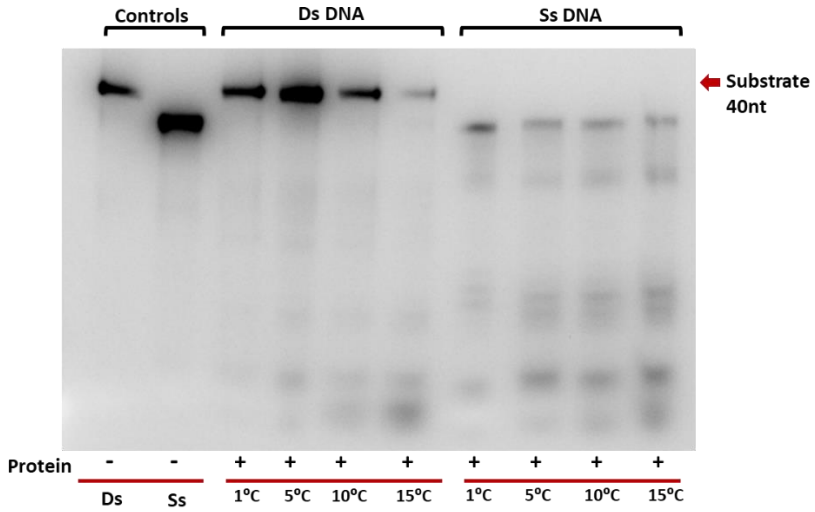
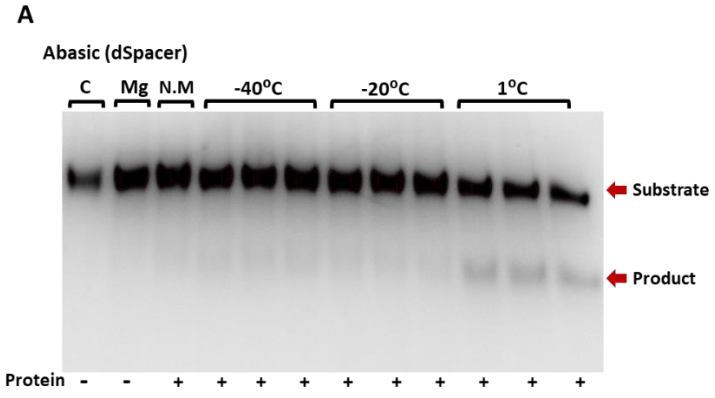


Figure 4.49. TBE urea PAGE showing DV-1-1-Nuc protein activity on double stranded (Ds) and single stranded (Ss) DNA substrates, from 1 to 15 °C. Lanes 1 (Ds) and 2 (Ss) contain no protein. Lanes 3-6 contain protein (+) and Ds DNA substrates. Lanes 7-10 contain protein (+) and Ss) DNA substrates. Substrate bands are indicated by a red arrow. Reactions were run for 4 hours at 1, 5, 10 °C and 15 °C. Protein was at a final concentration of 3.3 µM. Magnesium was used in all reactions, at a final concentration of 10 mM. Substrates are fluorescently labelled, and these bands were visualized using iBright™ CL750 Imaging System, Invitrogen™.

The following **Figure 4.50** shows the results of an activity assay with MBP-tagged DV-1-1-Nuc on abasic DNA substrate, from -40 ºC to 80 ºC. These results show that DV-1-1-Nuc has nuclease activity on abasic DNA across a wide range of temperatures. The low temperatures (-20 and -40 ºC) showed faint product bands, across both temperatures. Nuclease activity was most obvious between 1 and 50 ºC, with best activity observed at 35 ºC. No nuclease activity was seen in reactions 65 ºC and above.



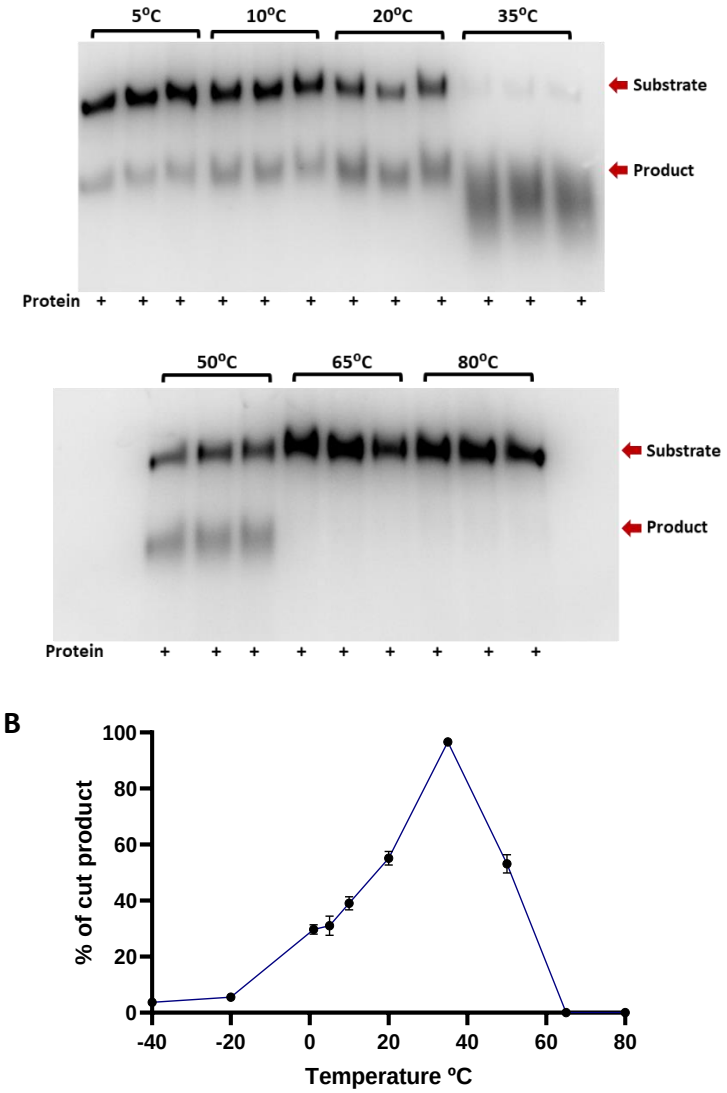


Figure 4.50. MBP-tagged DV-1-1-Nuc protein activity on abasic DNA substrate, from -40 to 80 °C. **A**) results of activity assay run on TBE urea PAGEs. Control reactions contain no protein or magnesium (C), contain DNA substrate and magnesium, but not protein (Mg) and contain DNA substrate and protein but no metal ion (N.M). Substrate and product bands are indicated by red arrows. Each temperature reaction was run in replicate of three. All reactions contain 20 % glycerol, to stop lower temperature reactions from freezing. **B**) quantitative results from analysis of product vs substrate band intensity, presented as a graph plot in GraphPad prism version 8 (GraphPadSoftware). Reactions were run for 4 hours at varying temperatures. Protein was at a final concentration of 2 µM. Magnesium was used in all reactions, at a final concentration of 10 mM. Substrates are fluorescently labelled, and these bands were visualized by iBright™ CL750 Imaging System, Invitrogen™.

4.2.13 Characterisation of the complete DV-1-1-Lig-Nuc protein

4.2.13.1 Re-design, expression, and purification of DV-1-1-Lig-Nuc protein

4.2.13.1.1 Design of new DV-1-1-Lig-Nuc construct

In light of the successful soluble production of the DV-1-1-Nuc construct 1, which had a new start site based on alignments with homologous proteins (**Figure 4.51, A**), expression of full-length DV-1-1-Nuc-Lig was re-visited. A new construct was designed for the full-length protein DV-1-1-Lig-Nuc to use the same start as DV-1-1-Nuc construct 1, removing the region modelled as unstructured by AlphaFold (**Figure 4.51**, **B**). The synthetic gene was ordered and was cloned into pHMGWA and pDEST17 expression vectors, as described in **Section 2.2.1**.

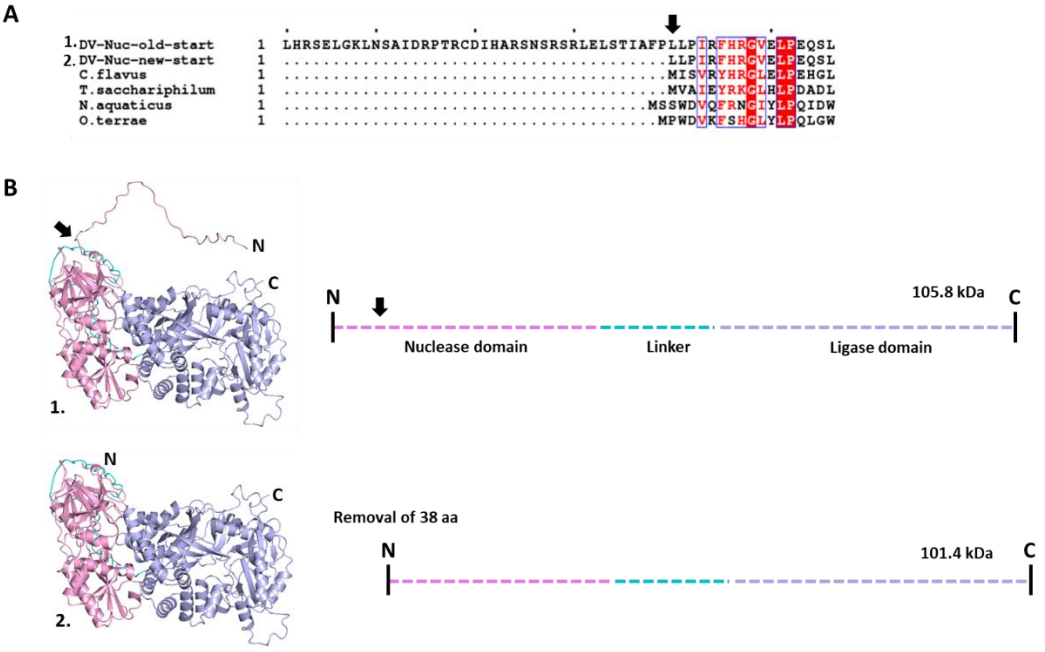


Figure 4.51. Design of N-terminally truncated DV-1-1-Lig-Nuc construct. **A**) Multiple sequence alignment of DV-1-1-Lig-Nuc with original N-terminus (1.) and N-terminal truncation (2.) against other ligase nuclease fusion proteins from *C. flavus, T. saccharophilum, N. aquaticus* and *O. terrae.* Sequence alignment was created using Clustal Omega version 1.2.4. and visualised using ESPript 3 (Robert & Gouet, 2014). **B**) Schematic of new N-terminally truncated construct for DV-1-1-Lig-Nuc. Pymol (Schrödinger, 2020) images of AlphaFold predicted DV-1-1-Lig-Nuc, show the removal of 38 aa from the N-terminus of the new construct (2.).

4.2.13.1.2 Small scale expression testing

DV-1-1-Lig-Nuc protein was recombinantly expressed in *E. coli* Origami cells and used in small scale expression screens, to determine optimal conditions for soluble protein expression. Two different temperatures (15 ºC and 20 ºC) and two different $O.D_{600}$ induction points ($O.D_{600}$ 0.9 & $O.D_{600}$ 0.5) were trialled, and the condition that produced the best soluble expression of DV-1-1-Lig-Nuc protein was 15 ºC at $O.D_{600}$ 0.9. MBP-tagged protein

(pHMGWA plasmid) showed highly expressed soluble protein in the Ni beads sample. No expression was seen for His-tagged (pDEST17 plasmid) protein, in any of the conditions (**Figure 4.52**).

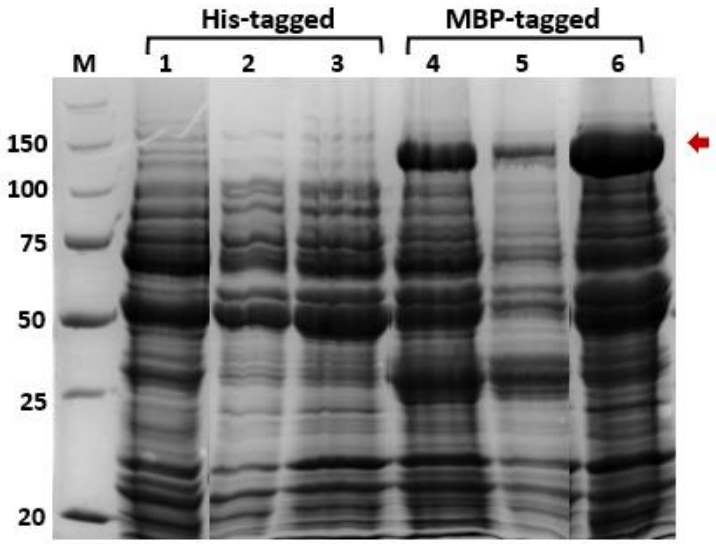


Figure 4.52. SDS PAGE of small scale protein expression results for His-tagged and MBP-tagged DV-1-1-Lig-Nuc. Lanes 1 & 4 represent insoluble protein, lanes 2 & 5 represent soluble protein and lanes 3 & 6 represent soluble protein bound to Ni beads. Red arrow indicates expression of of DV-1-1-Lig-Nuc protein, at the expected size for MBP tagged protein (145.3 kDa). No protein expression is seen in His tagged protein (104.9 kDa). A precion plus protein ladder was used as a molecular weight marker (M).

4.2.13.1.3 Large scale expression testing

Following on from results of soluble protein expression of DV-1-1-Lig-Nuc protein in small scale screens, protein expression cultures were scaled up following methods from **Section 2.3.3**. DV1-1-Lig-Nuc was put through IMAC, followed by a reverse IMAC and gel filtration chromatography. Several attempts were made to remove the MBP tag from the protein, but this was unsuccessful, as tagged, and un-tagged protein eluted in the same fractions at all purification steps.

During the first IMAC purification, the DV-1-1-Lig-Nuc protein eluted off the column in both the wash step and the imidazole gradient. DV-1-1-Lig-Nuc is the most abundant protein, however many contaminating *E. coli* proteins remain present (**Figure 4.53**, **A**). Overnight TEV cleavage of the MBP tag, was 50 % successful, as seen by the before and after TEV samples and DV-1-1-Lig-Nuc protein eluted off the IMAC column as MBP-tagged (145.3 kDa) and un-tagged (101.5 kDa) samples in the same fractions (**Figure 4.53**, **B**). Attempts to separate

tagged and un-tagged protein by gel filtration were likewise unsuccessful, despite the chromatogram showing two peaks, both tagged and untagged DV-1-1-Lig-Nuc protein were present in Peak A (**Figure 4.53**, **C**).

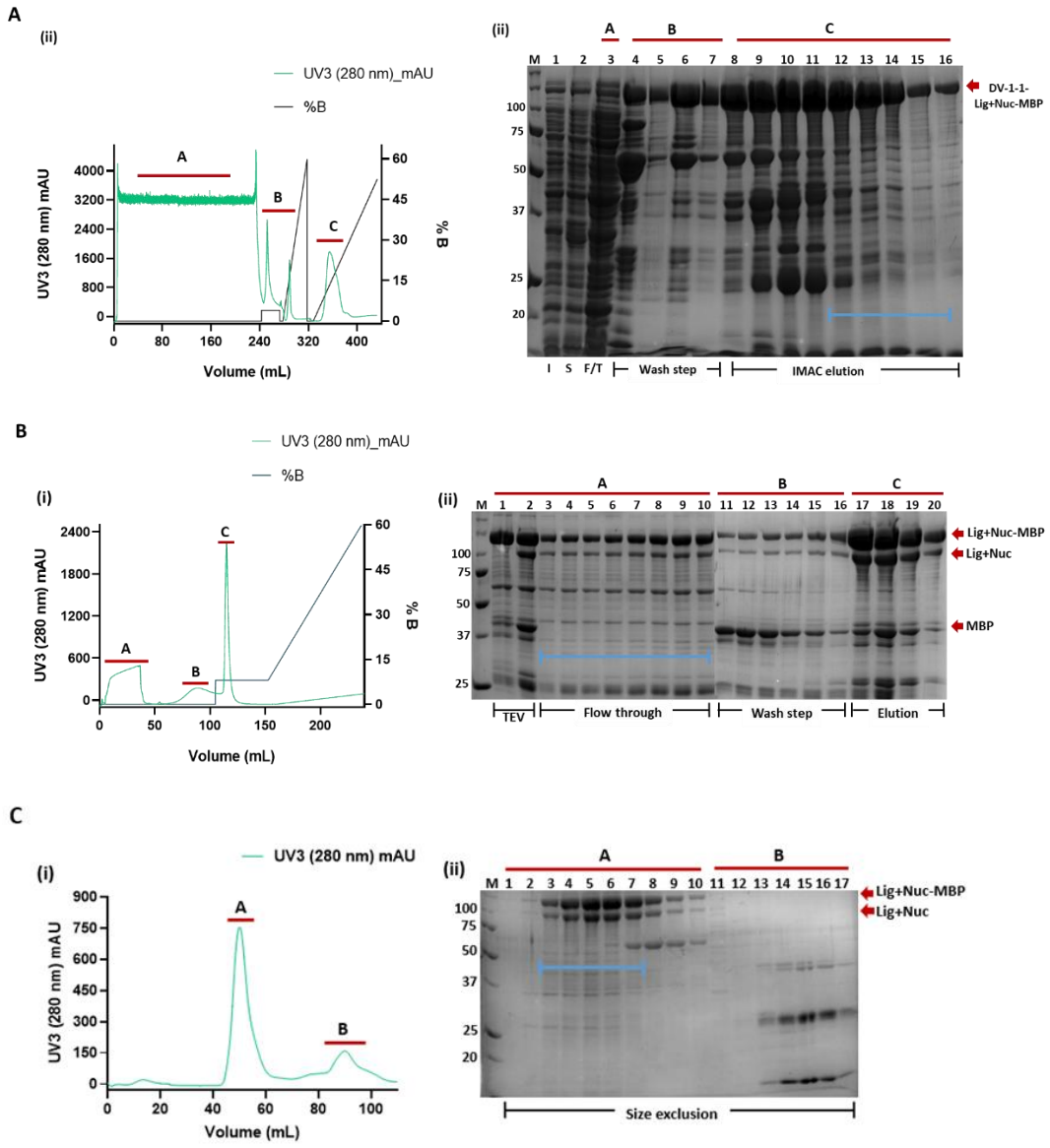


Figure 4.53. IMAC, gel filtration and MBP chromatograms (i) and SDS PAGE gels for production of DV-1-1-Lig-Nuc expressed in *E. coli* Origami (ii). **A**) IMAC purification of DV-1-1-Lig-Nuc. (i) Peak A represents flow through during IMAC purification, peak B represents fractions of proteins that eluted during the wash step. Peak C represents fractions of proteins that eluted during the elution step of the IMAC purification, including DV-1-1-Lig+Nuc protein (145.3 kDa). DV-1-1Lig+Nuc protein is indicated by a red arrow. **B**) Reverse IMAC purification of DV-1-1-Lig-Nuc. (i) Peak A represents flow through during IMAC purification, peak B represents proteins that eulted off the IMAC column during the wash step. Peak C represents proteins that eluted off the column during the imidazole concentration gradient, aorund 10 % imidazole. (ii) Lanes 1-2 are pooled IMAC fractions before the addition of TEV (1) and fractions after an overnight incubation with TEV (2), Lanes 3-10 are fractions from the flowthrough during the reverse IMAC purification. Lanes 11-16 are fractions that eluted off the IMAC column as a small peak, during the wash step. Lanes 17-20 are fractions that eluted during the imidazole gradient step, at 10 % imidazole. DV-1-1-Lig-Nuc protein eluted off the column as tagged (145.3 kDa) and un-tagged (101.5 kDa) samples. DV-1-1-Lig-Nuc proteins are indicated by

red arrows. **C**) Gel filtration purification of DV-1-1-Lig-Nuc. (i) Peak A contains fractions of MBP tagged (145.3 kDa) and un-tagged (101.5 kDa) DV-1-1Lig+Nuc protein. Peak B contains fractions of contaminating *E. coli* proteins and potentially TEV protease and MBP tag protein. (ii) Lanes 1-10 represent protein fractions from peak A and lanes 11-17 represents protein fractions from peak B. DV-1-1Lig+Nuc proteins are indicated by red arrows Blue bars indication fractions that were pooled, up concentrated and used in next purification step. Chromatogram graphs were designed in GraphPad Prism, version 9.0.0.

In a further attempt to separate out MBP-tagged and un-tagged DV-1-1-Lig-Nuc protein, fractions from the gel filtration purification indicated by the blue bar, in **Figure 4.53**, **C**, were pooled and put through a MBP column purification. This additional step did not separate out MBP tagged and un-tagged DV-1-1-Lig-Nuc protein, however, it did remove a significant *E. coli* contaminant, which eluted in the flow through (**Figure 4.54**).

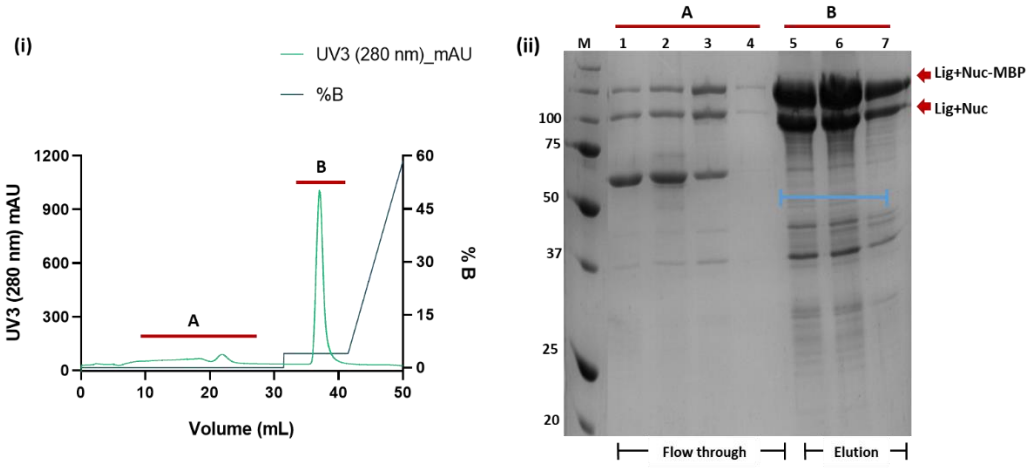


Figure 4.54. MBP purification chromatogram (i) and SDS PAGE gel for production of DV-1-1-Lig-Nuc protein expressed in *E. coli* (DE3) Origami(ii). (i) Peak A represents protein fractions that came through the flow through. Peak B represents protein fractions that eluted off the MBP column, which the addition of 10 % maltose. (ii) Lanes 1-4 are protein fractions that came through the flow through, lanes 5-7 represents protein fractions that eluted of the MBP during the maltose gradient step. DV-1-1-Lig-Nuc protein eluted off the MBP column at 10 % maltose, as MBP tagged (145.3 kDa) and un-tagged (101.5 kDa) fractions. DV-1-1-Lig-Nuc proteins are indicated by red arrows. The blue bar indicates fractions that were pooled, up concentrated, frozen and stored at -80 ^{o}C. Chromatogram graph was designed in GraphPad Prism, version 9.0.0.

Several attempts were made to separate out MBP tagged and untagged DV-1-1-Lig-Nuc protein, through further purifications. However, these attempts were also unsuccessful.

4.2.13.2 Biochemical characterisation of DV-1-1-Lig-Nuc protein

Following purification of DV-1-1-Lig-Nuc protein, biochemical activity assays were carried out to determine if the protein showed the nuclease and ligase

activities that was seen with the separate domain, in the sections above. Owing to the protein eluting as tagged and un-tagged in the same fractions, protein concentration was based on nanodrop readings, as described in **Section 2.4.8.**

4.2.13.2.1 Time dependence of DV-1-1-Lig-Nuc nuclease and ligase activities

A series of time dependence activity assays was carried out with DV-1-1-Lig-Nuc protein, on abasic DNA substrate and nick DNA substrate.

The results below (**Figure 4.55**) show that DV-1-1-Lig-Nuc protein has nuclease activity on abasic DNA, after 20 minutes of incubation, with magnesium. Activity was slower in reactions containing manganese, taking an hour before any nuclease activity was observed. No activity was observed with zinc, except in the 1-hour time point. For future activity assays, protein was incubated with abasic DNA substrate for four hours, to ensure detectable nuclease activity on this substrate.

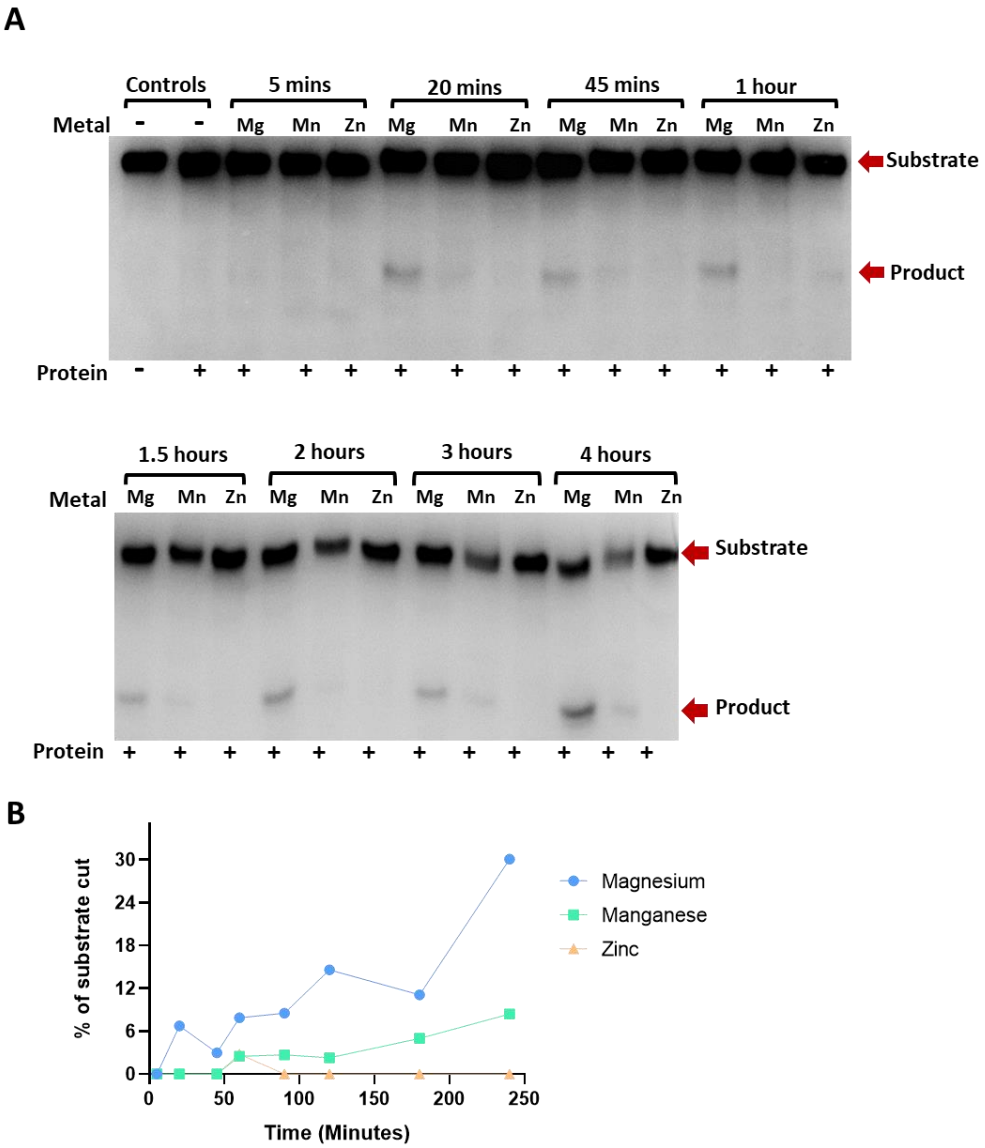


Figure 4.55. Time dependence activity assay of DV-1-1-Lig-Nuc protein, showing activity on abasic DNA substrate. Abasic DNA substrate was used to see if DV-1-1-Lig-Nuc was active on this substrate and whether it showed similar activity to that seem seen by the nuclease domain alone, in Section 4.2.10.3. Here DV-1-1-Lig-Nuc was incubated with abasic DNA substrate, with magnesium, manganese, and zinc, at different time periods (5, 20, 45 mins, 1, 1.5, 2, 3 & 4 hours) to determine what time point would show significant nuclease activity on abasic DNA substrate. **A**) TBE urea PAGE showing results of DV-1-1-Lig-Nuc activity assay on abasic DNA substrate at different time points. Addition of protein to the reaction is indicated by a plus symbol (+). Control reactions are indicated by (Controls) and don't contain protein (-). Product and substrate are indicated by red arrows. **B**) Quantification of nuclease activity, on abasic DNA, at different time points. The percentage of products were calculated from Image J software and presented as graphs, using GraphPad Prism. Reactions were incubated at different time points, at 25 ^{o}C, with 1.5 mg/ml protein concentration, and 10 mM final metal ion concentrations. Results of activity assays were visualized using iBright™ CL750 Imaging System, Invitrogen™.

The results below (**Figure 4.56**) show that DV-1-1-Lig-Nuc protein does have ligation activity on nick DNA, but the ligation activity is low compared to that seen by the ligase domain alone. There are inconsistencies in ligated product observed on the gel, where the product band intensity decreases at the 2 hour

incubation period and increases again after 3 hours. This is not an expected result and will require further experiments to investigate why the formation of ligated product was lower at the 2 hour time point.

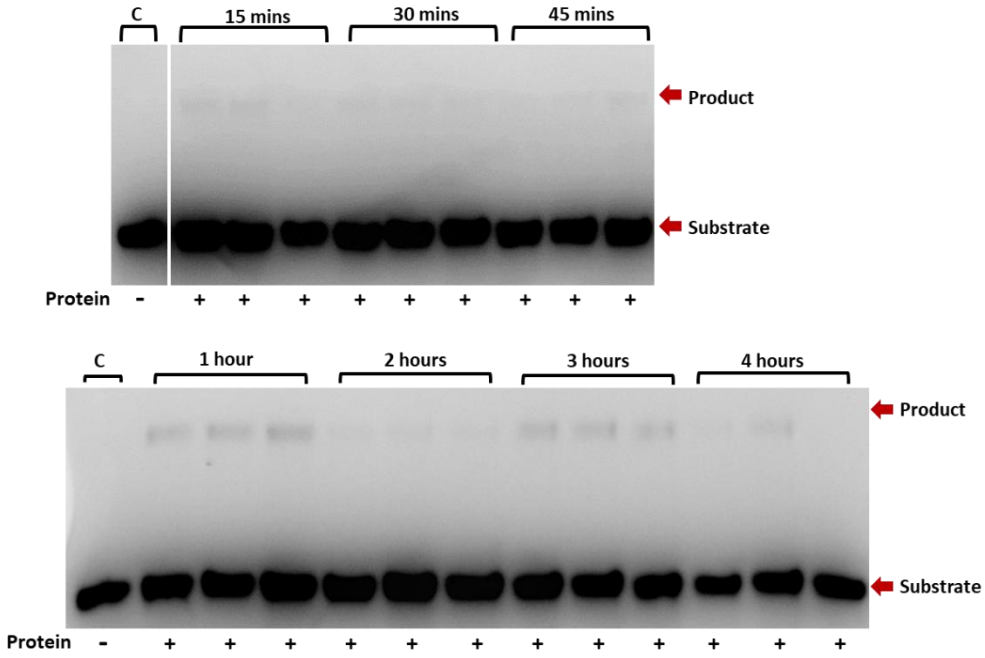


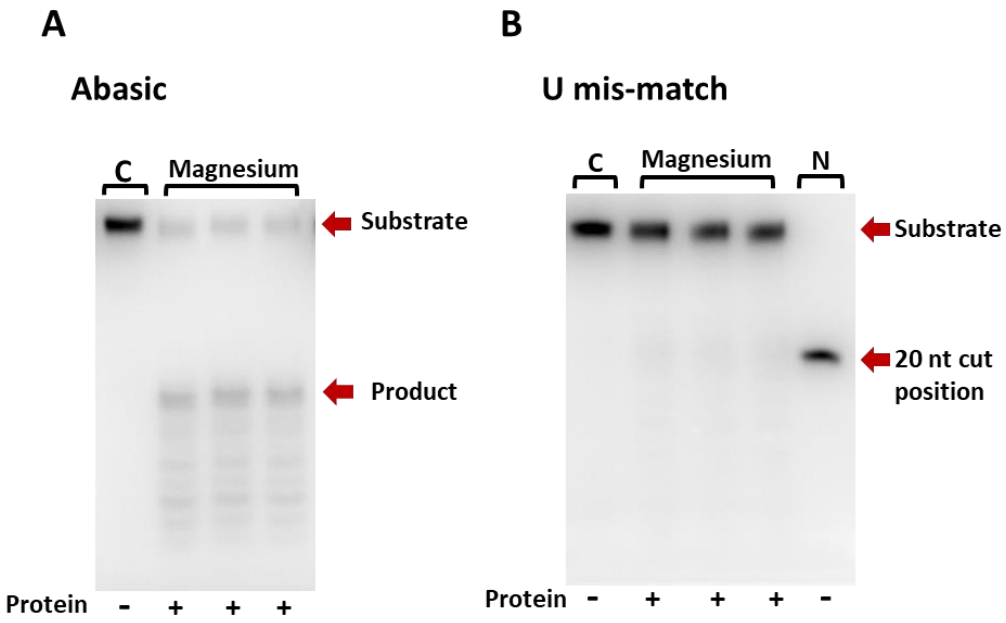
Figure 4.56. Time dependence activity assay of DV-1-1-Lig-Nuc protein, showing activity on nick DNA substrate. A TBE urea PAGE showing results of DV-1-1-Lig-Nuc activity assay on nick DNA substrate at different time points (15, 30, 45 mins, 1, 2, 3 & 4 hours). Addition of protein to the reaction is indicated by a plus symbol (+). Control reactions are indicated by (C) and don't contain protein (-). Product and substrate are indicated by red arrows. Reactions were incubated at different time points, at 25 ^{o}C, with 1.5 mg/ml protein concentration, 1 mM final concentration of ATP and 10 mM final magnesium ion concentration. Results of activity assays were visualized using iBright™ CL750 Imaging System, Invitrogen™.

4.2.13.2.2 Substrate specificity for DV-1-1-Lig-Nuc protein

Different DNA substrates were used to determine if DV-1-1-Lig-Nuc protein was active on these substrates and if the activity seen matched that of the ligase and nuclease domains separately.

Nuclease activity by DV-1-1-Lig-Nuc was tested with un-modified (Ds, Ss, 5' tail, 3'-tail), damaged (abasic) and mis-matched (uracil match) DNA substrates, along with flapped and splayed DNA substrates, with magnesium ion as the metal cofactor. The results below (**Figure 4.57**) show that there is some nuclease activity observed across most substrates, with the best activity seen on the abasic DNA substrate. Activity on abasic DNA is similar to what is observed

with the nuclease domain alone, as described in **Section 4.2.12.3**, where a specific cut is made on the substrate, followed by further exonuclease activity on this product. This specific cutting pattern is also observed on 3' and 5' flapped, 5' tail (un-modified) and very slightly with uracil match DNA substrates. Activity on splayed DNA was only observed in one replicate reaction and would need to be repeated to confirm activity on this substrate. In reactions with un-modified DNA substrates, nonspecific cutting is observed on Ds and Ss DNA, seen by smearing of the band. DV-1-1-Lig-Nuc protein, shows a specificity towards 5' tail DNA, compared to 3' tail DNA, observed by specific sized products with the 5' tail DNA.



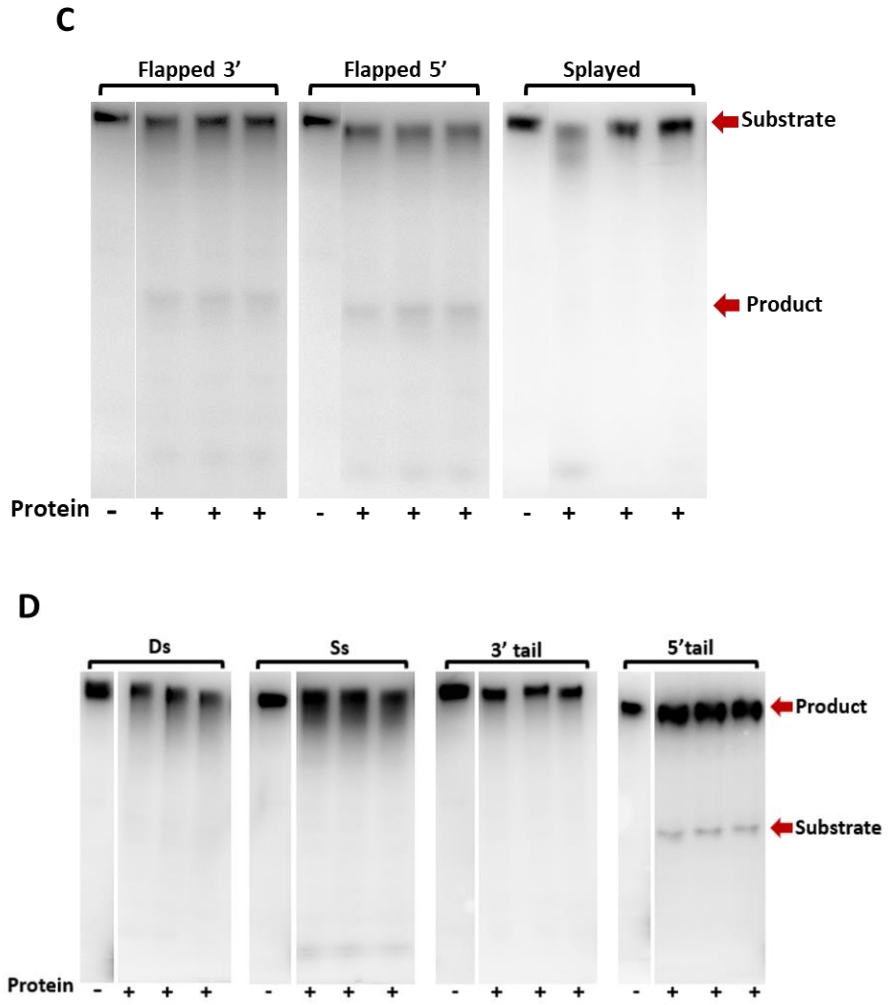


Figure 4.57. Nuclease activity by DV-1-1-Lig-Nuc on different DNA substrates. **A**) Results of activity assay with abasic DNA substrate, visualized on a urea PAGE. **B**) Results of activity with uracil match DNA substrate, visualized on a urea PAGE. A 20 nt control (N) was used here, to indicate size of product. **C**) Results of activity assay with flapped and splayed DNA substrates, visualized on urea PAGEs. **D**) Results of activity assay with un-modified DNA substrates, visualized on urea PAGEs. Addition of protein to reactions is indicated by a plus symbol (+). Control reactions don't contain protein (-). Product and substrate are indicated by red arrows. Reactions were incubated for 8 hours, at 25 ^{o}C, with 1.5 mg/ml protein concentration, and 10 mM final magnesium ion concentration. Results of activity assays were visualized using iBright™ CL750 Imaging System, Invitrogen™.

The ligation ability of DV-1-1-Lig-Nuc was also tested on the same seven non-canonical DNA substrates described previously (**Section 1.10 and Chapter 3**). Ligation on these DNA substrates was minimal, with some ligation seen on UB_duplex 14, UBSB_duplex 14 and UBSB_duplex 15. The best ligation was observed on UB_duplex 13, which is comparable to the ligation seen on nick DNA substrate (**Figure 4.58**).

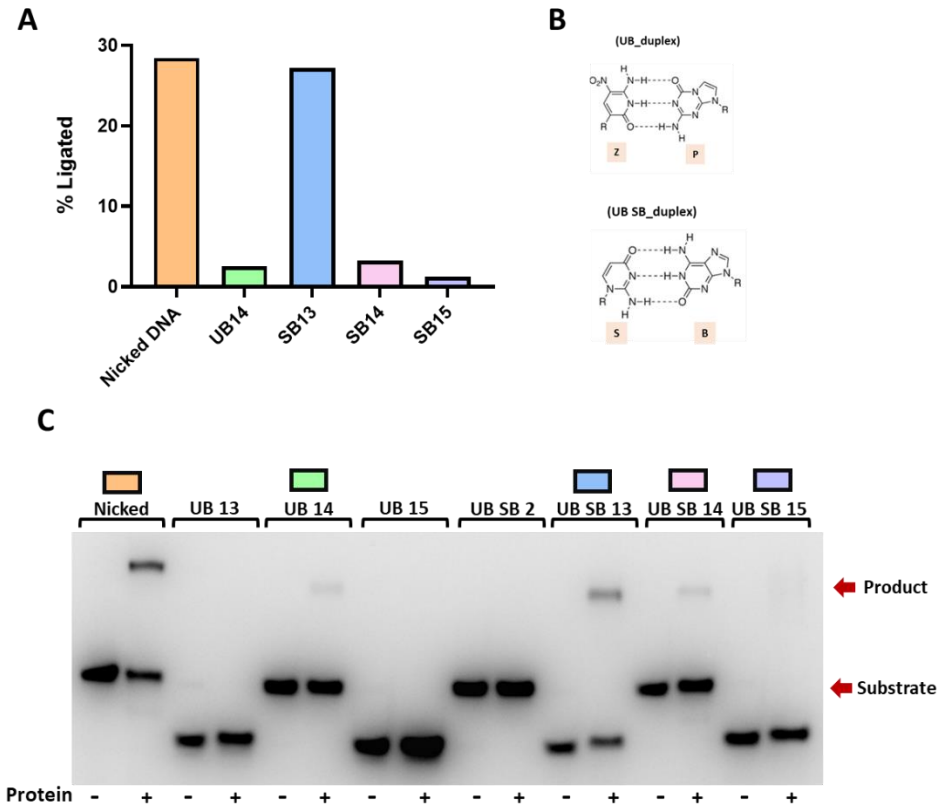


Figure 4.58. Represents the ligation ability of DV-1-1-Lig-Nuc on a range of substrates with 3-6 non-canonical expanded base-pair substrates, with magnesium as the divalent metal cofactor. The control, fully-natural nicked substrate, is indicated in orange. **A**) represents the quantitative summary of ligation by DV-1-1-Lig-Nuc on nicked DNA and four different non-canonical substrates. **B**) represents chemical modification of DNA to generate UB and SB DNA duplexes. **C**) represents the results of these ligation activity assays shown on urea PAGE gels. Controls contain no protein (-). Product and substrate bands are indicated on the gel, by red arrows. Reactions were carried out for 2 hours, at 25 ^{o}C, with 1.5 mg/ml protein concentration, 1 mM final concentration of ATP and 10 mM final concentration of magnesium. Results of activity assays were visualized using iBright™ CL750 Imaging System, Invitrogen™.

4.2.13.2.3 Metal ion preference of DV-1-1-Lig-Nuc

In previous activity assays with DV-1-1-Nuc protein, nuclease activity on DNA substrates was activated by the addition of magnesium, manganese and zinc metal ions, **Section 4.2.12.5.** These same activity assays were carried out with DV-1-1-Lig-Nuc protein, on abasic DNA substrate, to determine if the full-length protein also showed activity with the addition of these metal ions.

Here reactions with the addition of magnesium show the greatest nuclease activity on abasic DNA, in comparison to reactions with manganese and zinc. All reactions with the different metal ions result in a specific product band. Different concentrations of zinc metal ion were used to determine the best concentration

range for optimum nuclease activity on substrate. Results from this activity assay, show that a lower zinc concentration, supports nuclease activity on abasic DNA by DV-1-1-Lig-Nuc protein (**Figure 4.59**).

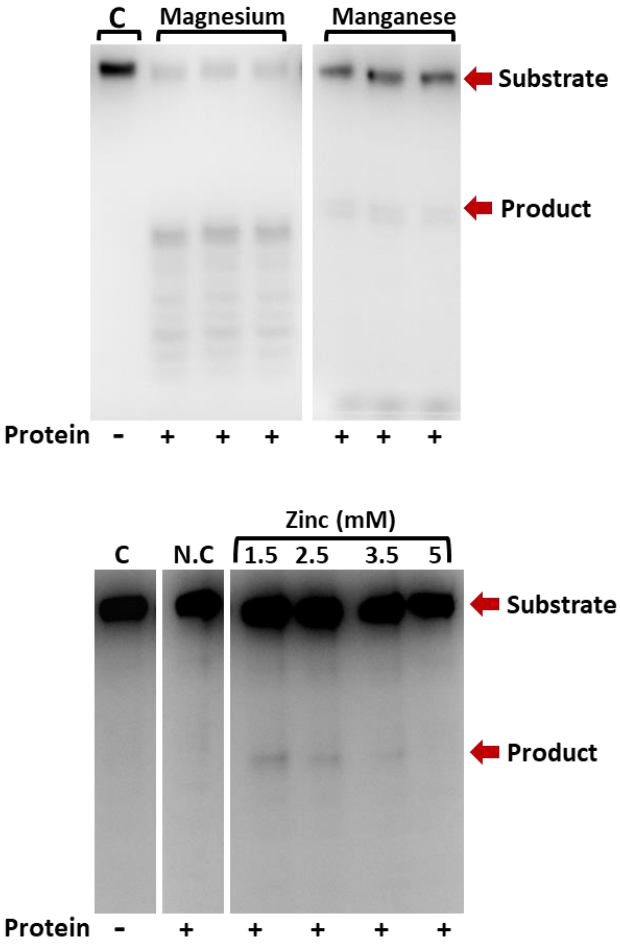


Figure 4.59. Results of nuclease activity by DV-1-1-Lig-Nuc on abasic DNA substrate, with magnesium, manganese, and zinc metal ions. Activity assays are visualized on urea PAGEs. Control reactions are indicated by (C). A no-cofactor reaction was also included (N.C), that contains protein and DNA substrate. Substrate and product are indicated by red arrows. Protein in reactions is indicated by a plus symbol (+). Magnesium and manganese were added to the reactions, with a final concentration of 10 mM. Zinc was added to the reactions with different concentrations (1.5, 2.5, 3.5 & 5 mM). Reactions were carried out for 8 hours at 25 ^{o}C, with 1.5 mg/ml protein concentration. Results of activity assays were visualized using iBright™ CL750 Imaging System, Invitrogen™.

4.2.13.2.4 Temperature dependence of DV-1-1-Lig-Nuc activity

To determine the optimal temperature for DV-1-1-Lig-Nuc protein activity, a temperature gradient activity assay, from 5-50 ºC was carried out with nick and abasic DNA substrates.

Nuclease activity is seen between 5 and 30 ºC, with the best activity observed at 30 ºC. No nuclease activity is seen in reactions that were incubated at 50 ºC. Nick DNA substrate was used as a marker to show that the resulting product, from nuclease activity, is around 20 nt (**Figure 4.60**).

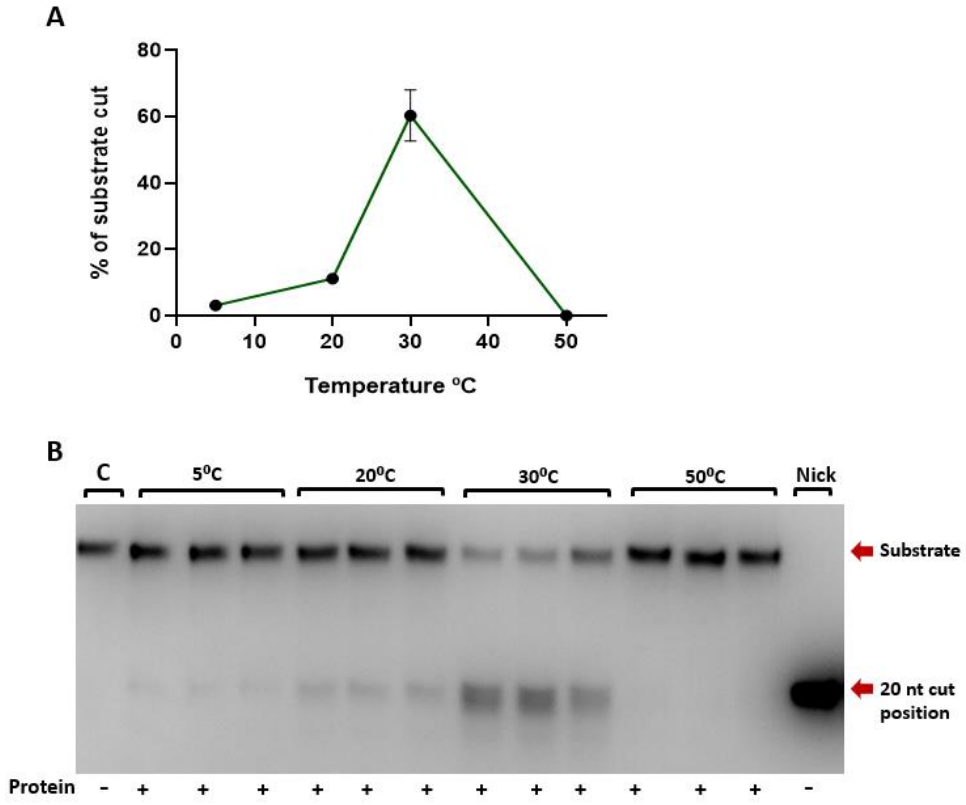


Figure 4.60. Results of nuclease activity by DV-1-1-Lig-Nuc protein, on abasic DNA substrate, at different reaction temperatures. **A**) a graph representing the quantitative summary of nuclease activity by DV-1-1-Lig-Nuc on abasic DNA, with different reaction temperatures (5, 20, 30 & 50 ºC). Plots on graph represent averages of each reaction temperature. Standard deviation error bars, are shown on graph. **B**) urea PAGE showing results of nuclease activity by DV-1-1-Lig-Nuc, at different temperatures. Addition of protein to the reaction is indicated by a plus symbol (+). Control reaction (C) dosent contain any protein (-). Nick DNA substrate, with no protein, was used as a 20 nt marker on the gel. The 20 nt cut position is indicated by a red arrow, which also indicates product bands. Substrate is also indicated by a red arrow. Reactions were carried out for 8 hours, at varying temperatures, with 1.5 mg/ml of protein added to reactions and 10 mM final magnesium ion concentration. Results of activity assays were visualized using iBright™ CL750 Imaging System, Invitrogen™.

Ligation activity is seen in all reactions between 5 and 50 ºC, with the best activity observed at 50 ºC (**Figure 4.61**).

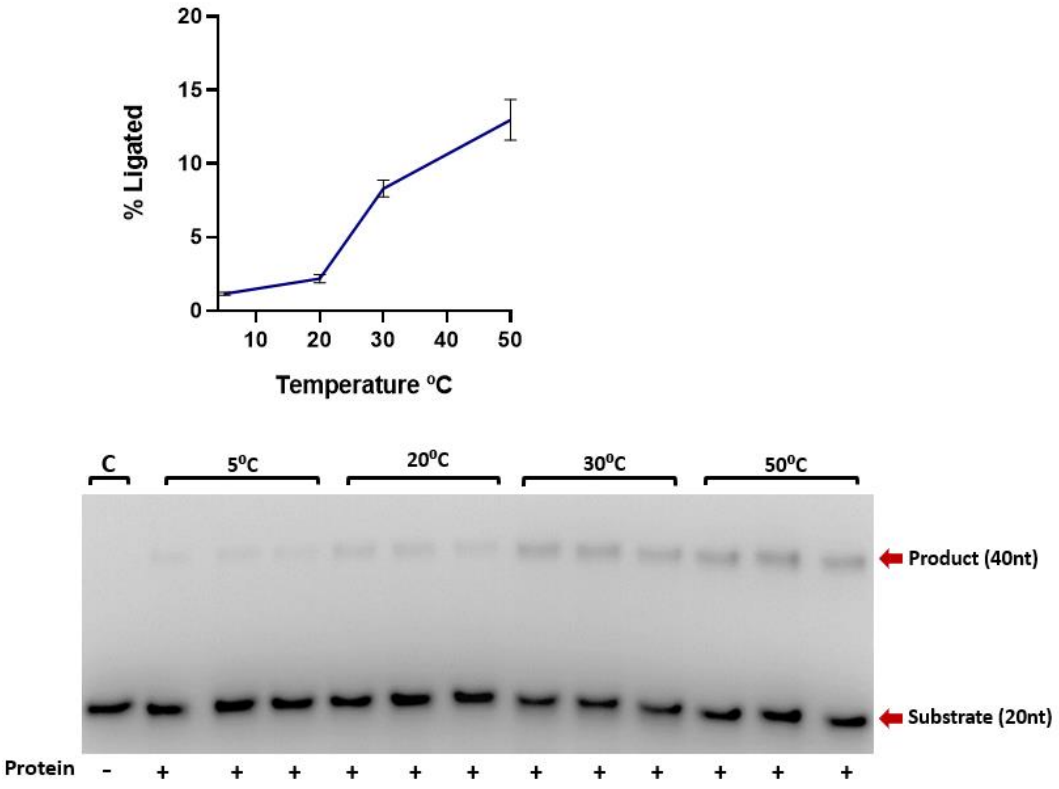


Figure 4.61. Results of ligation activity by DV-1-1-Lig-Nuc protein, on nick DNA substrate, at different reaction temperatures. **A**) a graph representing the quantitative summary of nuclease activity by DV-1-1-Lig-Nuc on abasic DNA, with different reaction temperatures (5, 20, 30 & 50 ^{o}C). Points on graph represent averages of each reaction temperature. Standard deviation error bars, are shown on graph. **B**) urea PAGE showing results of ligation activity by DV-1-1-Lig-Nuc, at different temperatures. Addition of protein to the reaction is indicated by a plus symbol (+). Control reaction (C) dosent contain any protein (-). Substrate and product are indicated by red arrows. Reactions were carried out for 8 hours, at varying temperatures, with 1.5 mg/ml of protein added to reactions, 1 mM ATP final and 10 mM final magnesium ion concetration. Results of activity assays were visualized using iBright™ CL750 Imaging System, Invitrogen™.

4.3 Discussion

DV-1-1-Lig-Nuc is a ligase nuclease fusion protein identified from the DV-metagenome. The gene encoding this ligase nuclease fusion protein belongs to a unique gene cluster, with several neighboring genes also encoding DNA modifying proteins. This gene cluster has not been previously described in other organisms. Attempts were made to recombinantly produce all proteins from this gene cluster, however only DV-1-1-Lig-Nuc was pursued further due to complications in expression and purification of the additional neighbouring proteins. Initial purification of the full-length ligase nuclease fusion protein (DV-1-1-Lig-Nuc) was unsuccessful, however expression and purification of the ligase (DV-1-1-Lig) and nuclease (DV1-1-Nuc) domains separately, resulted in soluble, active protein for use in characterisation studies. Mass spec confirmed expression

and purification of DV-1-1-Nuc and correct folding of both separate domain was confirmed through CD and SYPRO thermal melts. Full length DV-1-1-Lig-Nuc was eventually obtained by re-designing with an N-terminal truncation to match the starting position of DV-1-1-Nuc. This new construct resulted in soluble expression of DV-1-1-Lig-Nuc protein.

Attempts to crystalize each domain separately were ultimately unsuccessful. DV-1-1-Lig forms protein crystals when bound to nicked DNA substrate; however, these crystals were unsuitable for X-ray diffraction experiments. In lieu of a crystal structure, AlphaFold predicted structural models were generated for both the fused protein (DV-1-1-Lig-Nuc), the ligase (DV-1-1-Lig) and nuclease (DV-1-1-Nuc) domains separately.

The ligase domain is made up of a DB, an AD and an OB domain. This arrangement of domains is common to the Lig B class of ATP-dependent DNA ligases. The conserved motifs of these Lig B type ligases are also found in the sequence of DV-1-1-Lig. This domain arrangement is also found in human DNA ligase I (Pascal et al., 2004). The structure of human DNA ligase I reveals that the protein's DB domain enables it to surround its DNA substrate, stabilize the distorted DNA structure, and position the catalytic core on the nick, thereby creating a base for DNA ligation (Pascal et al., 2004). In h-LigI a salt bridge forms between residues Asp-570 and Arg-871 which stabilizes the AD-OB domain interface. DV-1-1-Lig also forms a salt bridge between the AD and OB domains, as well as the AD and OB domains. DV-1-1-Lig was overlayed onto the crystal structure of h-LigI, bound to nicked DNA duplex. DV-1-1-Lig forms similar interactions with the DNA duplex, as h-LigI, with all domains forming polar contacts with the DNA duplex. The catalytic lysine (Lys-286) was identified in the AD domain and forms contacts with the AMP group.

The DV-1-1-Nuc domain is a member of the metallo-β-lactamase (MBL) structural superfamily, of which the β-CASP (CPSF, Artemis, SMN1, PSO2) nucleic acid processing nucleases are a subfamily (Baddock et al., 2021; Callebaut et al., 2002; Schmiester & Demuth, 2017). Members of this family and DV-1-1-Nuc possess a characteristic α/β/β/α MBL core fold, containing five conserved

sequence motifs. The β-CASP domain is inserted within the MBL domain and contains three conserved motifs (Callebaut et al., 2002). Between these domains a salt bridge is formed between Glu177 from the β-CASP domain to residues Lys110 and Arg-112 from the MBL domain, stabilising interactions between domains at the active site. The active site is located at the interface between the MBL and β-CASP domains which contains key residues important for metal binding. DV-1-1-Nuc was overlayed onto the structure of SNM1C, containing two zinc ions in the active site. In both structures the same residues form polar contacts to the zinc ions. Identification of these conserved residues involved in zinc coordination were important in the design of DV-1-1-Nuc mutant. Other structures of homologous proteins have been solved with different metal ions in the active site, but with same active site residues forming contacts with the metal ions (Baddock et al., 2021; Baddock et al., 2020).

There are no other existing crystal structures of homologous ligase nuclease fusion proteins. Structural comparisons were based on other predicted models of these ligase nuclease fusion proteins identified in *C. flavus Ellin428, T. sacchariphilum*, *N. aquaticus* and *O. terrae* in the AlphaFold database. These ligase nuclease proteins all share structural and some sequence homology. Several interactions were observed between the nuclease and ligase domain, constituents of DV-1-1-Lig-Nuc. Within the nuclease domain, residues from both MBL and β-CASP sub-domains form polar contacts to several residues from the OB domain of the ligase. A polar contact is also observed between the β-CASP and the DB domain of the ligase. Several of these contacts were revealed to form salt bridges that can contribute to protein stability (Donald et al., 2011). The ligase nuclease fusion protein from *O. terrae* belongs to a well characterized gene cluster, with genes encoding for proteins involved in DNA repair. It is likely that DV-1-1-Lig-Nuc and its additional neighbouring proteins are involved in a novel DNA repair pathway.

In vitro characterization indicates DV-1-1-Lig domain alone exists as a monomer in solution. It can bind to a nicked DNA substrate, as observed in DNA binding experiments, and readily ligates this DNA substrate, with the addition of magnesium or manganese. Generally, there was an increase in ligated product

with the addition of manganese over magnesium. This result is interesting, as with DV-Lig5 and DV-Lig2 (**Section 3.3**) more ligation of product is observed with magnesium compared to manganese. As discussed in Section **3.3** most phosphoryl transfer enzymes appear to use magnesium as the physiological cofactor, however many are also able to utilize manganese. H-Lig I can ligate nicked DNA substrate in the presence of manganese, as a substitute for magnesium, displaying similar binding affinities and resulting in similar maximal rate constants between these metal ions (Taylor, 2014). Several other DNA ligases are also capable of substituting manganese, in the place of magnesium, such as those found in; *Chlorella* virus, *Methanobacterium thermoautotrophicum* and *Haemophilus Influenzae* (Ho et al., 1997; Sriskanda et al., 2000). In DV-1-1-Lig increasing the concentration of both metal ions causes inhibitory effects on the ligation activity of nicked DNA. Interestingly, addition of manganese to reactions across a range of different incubation temperatures results in higher percentage of ligated product compared to magnesium. At incubations of 25 ºC the percentage of ligated product is comparable between reactions with magnesium or manganese, while incubations at high or low temperature ranges show greater ligation with manganese. Overall DV-1-1-Lig can ligate nicked DNA from a temperature range of -5 ºC up to 80 ºC, with a temperature optimum of 25 ºC.

Ligation of nicked DNA is slower, comparatively to other homologous DNA ligases, suggesting it may require the nuclease domain for increased ligation ability. This is supported by the several interactions between residues of the ligase and nuclease domain, which may be important for optimal binding and activity on DNA substrates. Most DNA ligases are capable of ligating nicked DNA substrates, with a select number also being able to ligate mismatch, cohesive and blunt DNA substrates, such as the T4 DNA ligases. It has been speculated that the addition of a DBD allows these ligases to ligate additional substrates, without the requirement for accessory proteins (Bilotti et al., 2022). DV-1-1-Lig contains this DBD and can also ligate cohesive and mismatched DNA substrates, but this ligation ability is not comparable to that on nicked DNA substrate. In the literature it has been noted that the addition macromolecular crowding reagents, such as polyethylene glycol (PEG), are commonly used as ligation enhancers to boost

yield of ligation product and increase ligation efficiency. (Bilotti et al., 2022; Wang et al., 2019).

As observed with DV-Lig2 and DV-Lig5, DV-1-1-Lig can ligate non-canonical DNA substrates, although ligation on these substrates is minimal compared to that with nicked DNA substrate. As described in **Sections 1.10** and **3.3**, non-canonical DNA substrates are examples of unnatural base pairs, where the traditional nucleotides are expanded by the inclusion of two unnatural pyrimidine analogs and their complementary partners. DV-1-1-Lig showed minimal ligation on UB_duplex 13, UB_duplex 14, UBSB_duplex2 and UBSB-duplex 14. DV-1-1-Lig was most active on UBSB_duplex2 and UBSB-duplex 14, with a higher percentage of ligated product observed in reactions containing magnesium.

Purified DV-1-1-Lig, like DV-Lig5, is pre-adenylated from recombinant expression in *E. coli*, and a pre-incubation with unlabelled nick DNA was needed to remove background activity without any cofactor. Similar to DV-Lig5, DV-1-1-Lig can ligate nicked DNA with the addition of ATP, ADP and GTP nucleotide cofactors. No additional ligation occurs, with NAD, above basal levels of ligation. Increasing the concentration of ADP or ATP above 0.5 mM in reactions, decreased the amount of ligated product. Ligation was not affected by increased concentration of GTP and ligation of nicked DNA was very inefficient with GTP as the nucleotide cofactor. These results make it difficult to confirm the role of GTP as a nucleotide cofactor for DV-1-1-Lig. As discussed in **Section 3.3** other ATP dependent DNA ligases, from Archaea, have been reported to utilize additional nucleotide cofactors, such as ADP and GTP (Kim et al., 2013; Seo et al., 2007; Sun et al., 2008). There has been some speculation around the use of ADP as a nucleotide cofactor for these Archaeal DNA ligases. In protein purifications of these DNA ligases, the *E. coli* adenylate kinase is a common contaminant, that catalyses the conversion of two molecules of ADP into AMP and ATP, the latter being a cofactor by DNA ligases (Chen et al., 2009) Crystallisation of ADP and GTP nucleotides bound to these DNA ligases, would elucidate if and how these cofactors bind to the active site of DNA ligases.

While members of the family of MBL-β-CASP nucleases have similar structures of their core catalytic domains, each has distinct functions and selectivities. SNM1A and SNM1B have been classified as exclusively 5’ to 3’ exonucleases, while SNM1C is an endonuclease, with minor 5’ to 3’ exonuclease activity (Malu et al., 2012; Niewolik et al., 2017). Human SNM1A participates in removal of DNA damage lesions and ICL repair. SNM1A and SNM1B prefer ssDNA substrates, with a requirement for free 5’-phosphate. In contrast, SNM1C preferentially cleaves hairpins and DNA junctions, although it is able to process ssDNA substrates. DV-1-1-Nuc, like SNM1C can cleave both Ds and Ss DNA substrates. It also exhibits nuclease activity on both 5’ and 3’ tail DNA substrates, with a preference for 5’-tail DNA substrates. This preference for 5’ DNA strand was also observed through activity on 3’-flapped and 5’-flapped DNA substrate, where DV-1-1-Nuc was most active on 5’-flapped DNA. Additional nuclease activity was also observed on splayed DNA substrates. DV-1-1-Nuc also shows activity on abasic and uracil mismatch DNA damage substrates. Repair of abasic sites is effected through the base excision repair pathway, while mismatch damages is usually repaired via the mismatch repair pathway. Both pathways require the use of a nuclease to remove the DNA damage before additional proteins can repair the DNA (Aydin, 2014; Niewolik et al., 2017). In summary, DV-1-1-Nuc is most active on single stranded DNA substrates, especially damaged or flapped DNA substrates, with a preference for activity in a 5’ to 3’ direction.

Some enzymes belonging to the MBL-β-CASP family can only utilize zinc ions for catalytic activity, such as CPSF-73 (Mandel et al., 2006). While others, such as SNM1C, SNM1A and SNM1B, can utilize magnesium, manganese and zinc as metal cofactors in the hydrolysis of DNA substrates (Chang et al., 2015; Sengerová et al., 2012). These studies also noted that nuclease activity, with the addition of zinc ions, only occurred at low concentrations of zinc (0.01 mm) and high concentrations of zinc inhibited nuclease activity. These results are comparable to what was observed from DV-1-1-Nuc, which could utilize magnesium, manganese, and low concentrations of zinc for nuclease activity on different DNA substrates. DSF thermal melts revealed that DV-1-1-Nuc was more

stable with the addition of manganese over magnesium, suggesting that manganese invokes a structural change in the protein, that supports stability.

DV-1-1-Nuc shows clear nuclease activity from 1 ºC up to 50 ºC. There is some basal level of degradation at lower temperatures (-40 and -20 ºC), however obvious degraded product is only clear from 1 ºC. No cleaved product is observed at temperature above 50 ºC, and these resulted are supported by findings from DSF thermal melts, where DV-1-1-Nuc had a T_m of 45 ºC, indicating protein unfolding at higher temperatures. Activity at low temperatures is common for psychrophilic bacteria, that need to be able to repair damaged DNA at these lower temperatures.

Design of the DV-1-1-Nuc mutant was based around disrupting interactions involved in metal binding. Residues Asp-36 and His-37 have been implicated in metal binding coordination in other MBL-β-CASP nucleases such as SNM1A, SNM1B and SNM1C. These identified residues belong to the highly conserved sequence motif II (HxHxDH) of the MBL domain. Mutation of these conserved residues has been shown to reduce or abolish nuclease activity on DNA substrates (Yosaatmadja et al., 2021). The DV-1-1-Nuc mutant still retained nuclease activity, although this activity was slightly reduced compared to wild-type. A second metal binding site is present in the structures of many MBL-β-CASP nucleases and supports binding of an additional metal ion. Residues His-32, His-34, His-87 and Asp-108 from Motif I, II and IV, of DV-1-1-Nuc have been implicated in the coordination of a metal ion, from structural investigations in some MBL-β-CASP nucleases (Yosaatmadja et al., 2021). It is likely that the nuclease activity of DV-1-1-Nuc mutant was only reduced as it was still able to bind metal ions at this additional binding site. Future mutational design will involve mutation of residues from this additional metal binding site, as well as conserved residues implicated in DNA binding from other MBL-β-CASP nucleases. Inactive mutants of DV-1-1-Nuc are required to confirm that the observed activity on DNA substrates is from DV-1-1-Nuc and not due to background activity by *E. coli* contaminants.

Characterisation studies on DV-1-1-Lig-Nuc showed this protein exhibited both ligase and nuclease activity in a similar manner to the separate domains. DV-1-1-Lig-Nuc was able to successfully ligate nicked DNA as well as several non-canonical DNA substrates; UB_duplex 14, UB_duplex 13 and UBSB_duplex 14. These results are comparable to ligation seen by DV-1-1-Lig, although DV-1-1-Lig-Nuc was more successful at ligating UBSB_duplex 13 over DV-1-1-Lig. Specific nuclease activity by DV-1-1-Lig-Nuc was observed on abasic, flapped 3', flapped 5' and 5'-tail DNA substrate, this specific nuclease activity was also observed with DV-1-1-Nuc on the same DNA substrate. There is also non-specific degradation observed on Ds, Ss and splayed DNA substrates, which was also seen by DV-1-1-Nuc domain. Nuclease activity by DV-1-1-Lig-Nuc was observed in reactions with the addition of magnesium, manganese, and zinc. There was some background degradation seen in reactions with no metal cofactor, however the addition of metals ions increased the degradation of substrate further. Nuclease activity appears to be inhibited by an excess of zinc ions in the reaction which was a similar observation with DV-1-1-Nuc. DV-1-1-Lig-Nuc has the ability to degrade and ligate from 5 ºC, however only ligation was observed in reactions above 30 ºC. The ligase domain appears to be more stable at higher temperatures than the nuclease domain, these results are also seen when comparing activity of the separate domains, at higher temperatures.

Structural characterisation of DV-1-1-Lig-Nuc identified that the nuclease domain forms several interactions with the ligase domain, as discussed earlier. Analysis of the electrostatic surface potential of DV-1-1-Lig-Nuc reveals electronegative regions, in both domains, where DNA may bind. Overall, the combined structural and biochemical findings from the separate and fused domains, of DV-1-1-Lig-Nuc, suggests a role in repair of damaged DNA. Owing to the additional DNA repair proteins, in the same gene cluster as DV-1-1-Lig-Nuc, including two DNA polymerases, it is likely that these proteins coordinate DNA repair together. The nuclease domain of DV-1-1-Lig-Nuc was particularly active on flapped DNA substrates, which could mimic the flaps formed after a polymerase, with strand displacement activity has replicated past a damage. The nuclease domain then removes this flap and in turn stimulates the additional DNA

repair neighbouring proteins to the site, ending in ligation of the nick by the DNA ligase.

Currently there is limited information in the literature for a proposed role of ligase nuclease fusion proteins in DNA repair. In addition, the gene arrangement for DV-1-1-Lig-Nuc along with its neighbouring genes, has currently not been discussed in the literature. What we do know, comes from studies of proteins like DNA ligase D (Lig D) and separate ligase and nuclease proteins that interact together in the same DNA repair pathway. DNA ligase D (Lig D) is essential in the repair of DNA double stranded breaks (DSBs) through the process of non-homologous end joining (NHEJ) in bacteria.

Lig D is made up of three distinct domains. The N-terminal polymerase domain, a central domain with 3'-phoshpoesterase and nuclease activity and a C-terminal ligase domain. While the nuclease and ligase domains of Lig D, share little structural and sequence homology to DV-1-1-Lig-Nuc protein, the role of Lig D in DNA repair, might suggest a potential role for DV-1-1-Lig-Nuc, especially considering the genome arrangement of DV-1-1-Lig-Nuc, in a gene cluster with genes encoding for: RecA and two DNA polymerases. Together the translated protein products, might coordinate the repair of Ds DNA breaks, through the process of NHEJ (Zhu & Shuman, 2006).

Additional studies, also support the role of DV-1-1-Lig-Nuc in NHEJ. Human DNA ligase IV and SNM1C are central components of the NHEJ pathway. Recently it was discovered that the C-terminal region of SNM1C directly interacts with the DB domain of LigIV. Together these proteins form a complex with additional NHEJ factors, to a site of a double stranded break (DSB) (De Ioannes et al., 2012). As previously discussed, the separate ligase and nuclease proteins identified from the four gene cluster, within *P. putida*, display structural and sequence similarity to the ligase and nuclease domains, that make up DV-1-1-Lig-Nuc. Ejaz and Shuman speculate that one or more enzymes from this gene cluster can contribute to NHEJ, either in an alternative pathway to Ku-Lig D or in the Ku-Lig D pathway as backup components, such as the DNA ligase (Ejaz & Shuman, 2018). While further characterisation of DV-1-1-Lig-Nuc and its

adjacent DNA modifying proteins is required, it is tempting to speculate that DV-1-1-Lig-Nuc could also participate in NHEJ or similar repair pathways.

5 Chapter 5

DV-Nuc3 (NucS) protein

5.1 Introduction

Nuclease enzymes are found in almost all known DNA repair pathways, with endonucleases playing an important role in excision repair pathways, that remove damaged or mis-match nucleotides within DNA duplexes (Wozniak & Simmons, 2022). A recently discovered family of nucleases, belonging to a subset of the Endonuclease V superfamily of proteins, are the NucS nucleases (Ren et al., 2007; Ren et al., 2009). This family represents a group of highly conserved enzymes found in various species of Euryarchaea, Crenarchaea, and Bacteria (Zhang et al., 2020). It is hypothesised, that NucS plays a role in the mis-match repair pathway, in place of the canonical MutS/MutL proteins, which are absent from some species of hyperthermophilic Archaea and Actinobacteria (Castañeda-García et al., 2017; Zhang et al., 2020).

NucS proteins were first discovered in the archaeal species; *Pyrococcus abyssi*, as a member of a new family of novel structure-specific DNA endonucleases in Archaea and were named NucS (for nuclease specific for ssDNA) (Ren et al., 2009). In *P. abyssi*, the gene sequence for NucS is encoded in the open reading frame PAB2263. Initial investigations of this protein predicted it to belong to the RecB family of nucleases. PAB2263 and its orthologues are part of a superfamily that, in addition to RecB nucleases, included several restriction endonucleases, Holliday junction resolvase and XPF/Rad1/Mus81 nuclease (Ren et al., 2007). NucS from *P. abyssi* was identified to have specificity toward branched and splayed DNA substrates (Ren et al., 2009).

Interestingly, a study revealed that the NucS protein from another archaeal species; *Thermococcus kodakarensis* (renamed EndoMS (Endonuclease Mismatch-Specific)) is a mismatch-specific endonuclease, acting on double-stranded DNA (dsDNA) substrates containing mismatched bases. Here the archaeal NucS protein binds and cleaves both strands at dsDNA mismatched substrates, with the mis-paired bases in the middle position, leaving 5-nucleotide

long 5'-cohesive ends (Ishino et al., 2016). This finding suggests that these DSBs may initiate the repair of mismatches, acting in a novel mismatch repair process (Ishino et al., 2016). Another study by (Nakae et al., 2016) revealed the structure of the *T. kodakarensis* NucS in complex with mismatched dsDNA, hence supporting the theory that EndoMs/NucS acts in non-canonical mismatch repair (MMR) pathways.

NucS proteins have since been discovered in Bacteria, specifically Actinobacteria. Castañeda-García and collegues identifed a non-canonical mismatch repair pathway in prokaryotes, specifically looking into *Mycobacterial* species (Castañeda-García et al., 2017). Here they report that *M. smegmatis* contains a putative endonuclease NucS/EndoMS, with no structural homology to known MMR factors. This NucS type homolog is required for mutation avoidance and anti-recombination, which is a signature of the canonical MMR pathway. They also undertook a phylogenetic analysis of NucS proteins, and these findings indicated a complex evolutionary process leading to a wide distribution pattern in prokaryotes. This discovery indicates that distinct pathways for MMR have evolved at least twice in nature (Castañeda-García et al., 2017).

Another study looking into EndoMS/NucS proteins in Bacteria, identified this protein within *Corynebacterium glutamicum* (Takemoto et al., 2018). Through biochemical and genetic analysis, they demonstrated that *C. glutamicum* EndoMS/NucS is a mismatch-specific endonuclease that functions cooperatively with a sliding clamp of the replisome to correct replication errors. They also discovered that unlike MutS-dependent systems, the EndoMS-pathway is highly specific to G/T, G/G and T/T mismatches.

Overall this NucS family of novel endonucleases, has been shown to act on branched, mismatch and deaminated DNA, implying that this endonuclease is a multifunctional enzyme involved in NER, MMR and deaminated base repair in a non-canonical function (Ishino et al., 2016; Ren et al., 2009; Zhang et al., 2020).

There is a lack of readily detectable sequence similarity outside of these proteins' active sites, which has complicated their characterisation and functional

annotation. The structures of EndoMS/NucS proteins from *T. kodakarensis* and *P. abyssi* were solved in the last few years (Ren et al., 2009) (Nakae et al., 2016). These structures reveal that these proteins share structural similarity, possessing 13 β-sheet and 5 α-helical structures. However, the location and length of these β-sheet and 5 α-helical structures differ between the two endonucleases.

Analysis of the EndoMS/NucS structure (5GKE), from *T. kodakarensis*, revealed that the enzyme takes on a dimeric conformation with the C-terminal domain forming the interface and mobile N-terminal domains in an open conformation without DNA and closing around DNA when it binds (**Figure 5.1**, **A**) (Nakae et al., 2016). The C-terminal domains contain catalytic sites like those in RecB. The structure of EndoMS dimer bound to dsDNA revealed that the mismatched bases were turned out into the binding sites and the overall structure mimicked that of restriction enzymes (Nakae et al., 2016).

In the structure of NucS, from *P. abyssi* (2VLD), each subunit of this protein clearly displays a dumbbell-like two-domain structure (**Figure 5.1**, **B**) The N- and C-terminal domains are separated from each other by a stretched polypeptide linker. The N-terminal domain has a unique half-closed β-barrel structure (Ren et al., 2009). This study (Ren et al., 2009) found that the N-terminal domain, of NucS, has some structural homology to the Sm-fold of the eukaryotic RNA-binding domain of the small nuclear ribonucleoproteins (Sm proteins) and the OB-fold of ssDNA-binding proteins (Kambach et al., 1999; Theobald et al., 2003) The C-terminal domain contains an α/β structure with a five-stranded central β-sheet and four flanking α-helices, representing a minimal-endonuclease fold (Pingoud et al., 2005). Within the C-terminal domain is the active site, with sequence motifs conserved in the family of RecB-like nucleases (Aravind et al., 2000). The dumbbell-like subunit structure of NucS has a large hydrophobic patch exposed on the six-stranded N-terminal β-sheet. By domain swapping, the subunits assemble to form a dimer that constitutes the asymmetric unit (**Figure 5.1**, **B**, **II**). This dimer formation is thought to be critical for the folding and stabilization of the NucS structure (Ren et al., 2009).

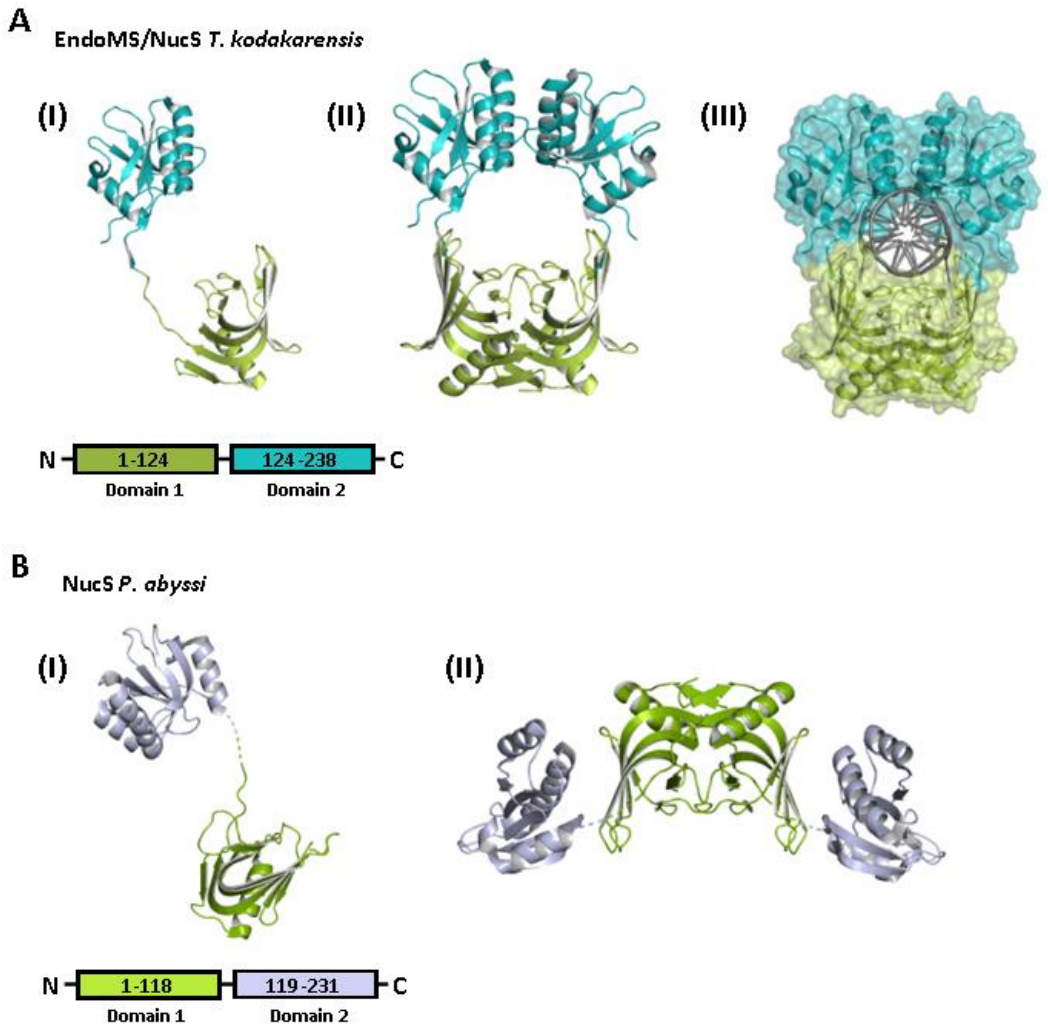


Figure 5.1. Structural arrangement of NucS proteins from *T. kodakarensis* (EndoMS/NucS), and *P. abyssi* (NucS). **A**) (I) Crystal structure of EndoMS/NucS, as a monomer, from *T. kodakarensis* (5GKE) (Nakae et al., 2016). Domains are coloured accordingly: domain 1 (green), domain 2 (blue) (II) EndoMS/NucS structure presented as a dimer. (III) Structure of EndoMS/NucS in complex with Ds DNA. Protein structure is presented as a surface, with 20 % transparency. **B**) (I) Crystal structure of NucS from *P. abyssi*, as a monomer. (2VLD) (Ren et al., 2009). Domains are coloured accordingly: domain 1 (green), domain 2 (purple). (II) NucS structure presented as a dimer. Structures were visualized and presented using PyMOL (Schrödinger, 2020).

The following results section describes the preliminary *in silico* identification of NucS nuclease proteins from the Dry Valley metagenomes and the *in silico* and biochemical characterisation of one of these NucS nucleases denotated DV-Nuc3.

5.2 Results

5.2.1 *In silico* characterisation and homology modelling of DV-Nuc3 NucS protein

As previously discussed in **Section 1.11** many sequences with matches to the NucS profile were discovered in the Dry Valley metagenomes. Using sequence similarity networks, two major Dry Valley metagenome containing clusters were resolved at 28 % edge, identifying each with 761 and 115 metagenomic sequences respectively. No additional known domains were identified by hmmscan for sequences in either cluster and for both, sequence homology searches revealed that proteins from these clusters, showed significant homology to the Endonuclease NucS family (Rzoska-Smith et al., 2023).

Cluster (1) contains representative nodes for sequences with >50 % identity in UniRef. This cluster represents the majority of Bacterial NucS sequences including *M. tuberculosis* and *C. glutamicum*. NucS proteins from *C. glutamicum* (Ishino et al., 2018), and *M. tuberculosis* (Cebrián-Sastre et al., 2021) have previously been biologically characterised.

Cluster (2) contains only Dry Valley metagenome sequences, with some being more than 500 amino acids in length. Most characterised homologs are much shorter than this. Alignments of 25 full length sequences from this cluster against *C. glutamicum*, *M. tuberculosis* and *T. kodakarensis*, show that the Dry Valley sequences have N-terminal sequence extensions in comparison to their NucS homologs. The 100-residue extension is ubiquitous among the Dry valley NucS homologs and shows no sequence identity to the characterised sequences. The C-terminal domain residues show the best alignment similarity and correspond to the RecB-like endonuclease domain. Among these Dry Valley NucS proteins belongs DV-Nuc3 (Ga0136640_100017415), which was originally annotated as hypothetical protein (Rzoska-Smith et al., 2023) (**Figure 5.2**).

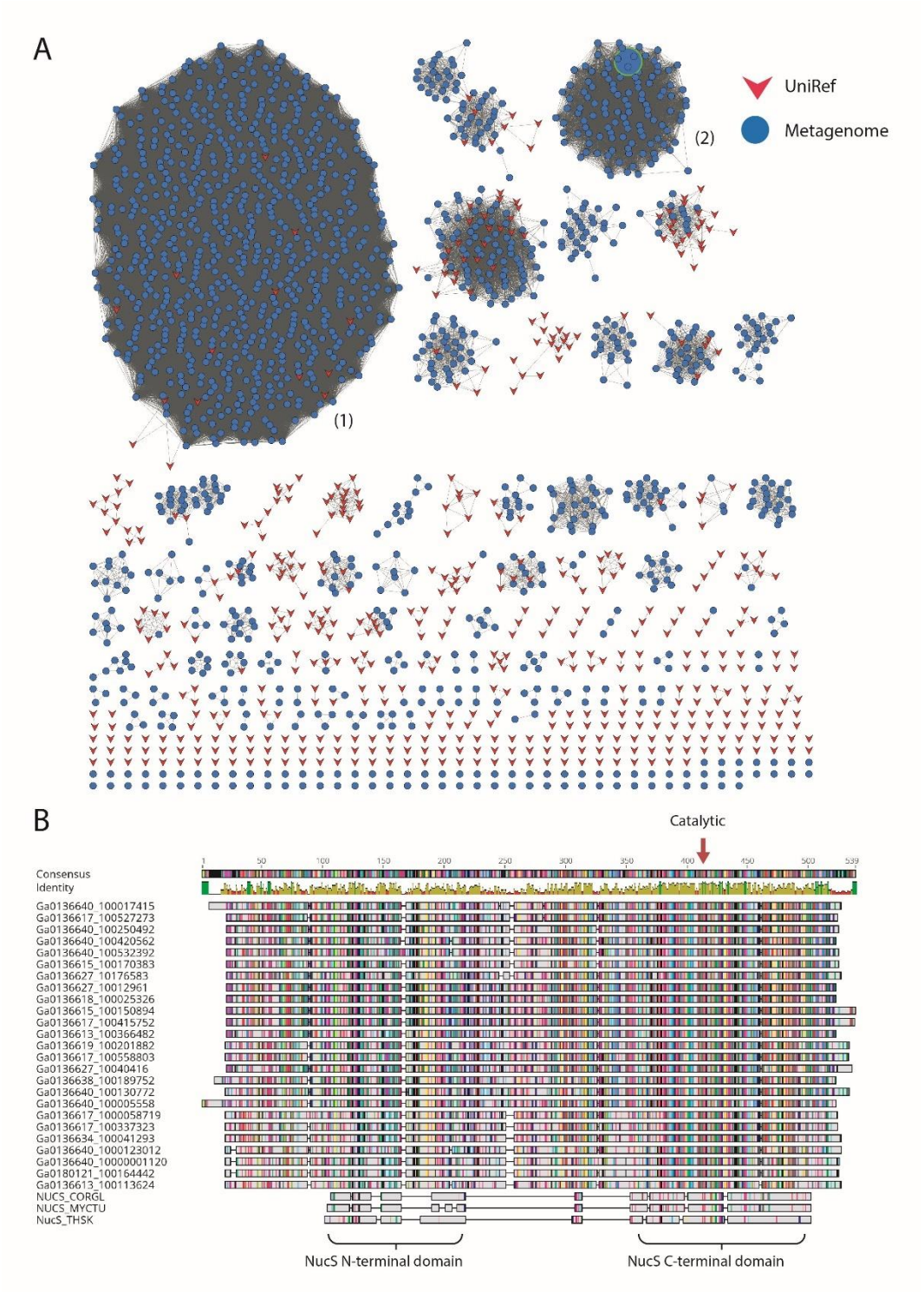


Figure 5.2. Sequence Similarity Network (SSN) for the NucS-type proteins at 28 % identify threshold. Sequence similarity networks were constructed for each set of sequences identified by hmmsearch using the EFI-EST server. **A**) Dry-Valley metagenome nodes are coloured blue, UniRef50 nodes are indicated in red. The node corresponding to the recombinantly produced homolog (DV-Nuc3) from Cluster (2) is indicated with a large, green-boarded symbol. **B**) alignment of full-length sequences from Cluster (2) with characterised EndoMS/NucS from *C. glutamicum* (NUCS_CORGL), *M. tuberculosis* (NUCS_MYCTU) and *T. kodakarensis* (NUCS_THSK). Domains identified in the crystal structure of *T. kodakarensis* EndoMS/NucS are indicated below the alignment and the position of the conserved catalytic aspartic acid residue (D) is noted by a red arrow (Rzoska-Smith et al., 2023).

DV-Nuc3 NucS protein belongs to the Acidobacteria bacterial lineage, of *Pyrinomonas methylaliphatogenes* based on analysis of the contig in JGI. Currently no known NucS proteins, have been identified from this bacterial lineage. The gene encoding DV-Nuc3 NucS is in a contig, alongside genes that encode for Lysophospholipase L1, a UV damage protection protein, a zinc carboxypeptidase, and an Acyl-coenzyme A thioesterase (**Figure 5.3**). Analysis of other Acidobacterial genomes, from the DV-metagenomes, shows there is also gene conservation of the Lysophospholipase L1 gene, always found in close proximately to the NucS gene. Interestingly the DV-Nuc3 NucS protein is classified as a hypothetical protein, within the Dry Valley genome and among other homologs of Acidobacteria. SSN revealed that this protein belonged to the NucS endonuclease family, as discussed above and in **Section 1.11**.

A search through the genome of *P. methylaliphatogenes* K22, using the Integrated Microbial Genomes & Microbiomes system (IMG) from the Joint Genome Institute (JGI) (Chen et al., 2023; Markowitz et al., 2014), identified a homolog of DV-Nuc3 (38 % identity) but also revealed that MutS and MutL genes are present in the genome. Due to the presence of these genes, in the same bacterial lineage of DV-Nuc3 NucS, it is possible that DV-Nuc3 NucS plays a different role within Acidobacterial species.

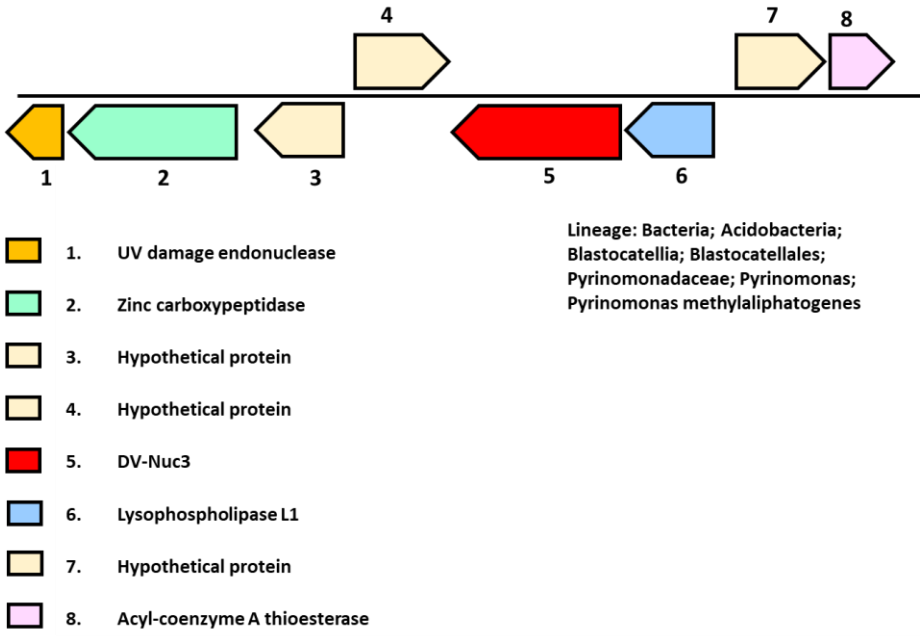


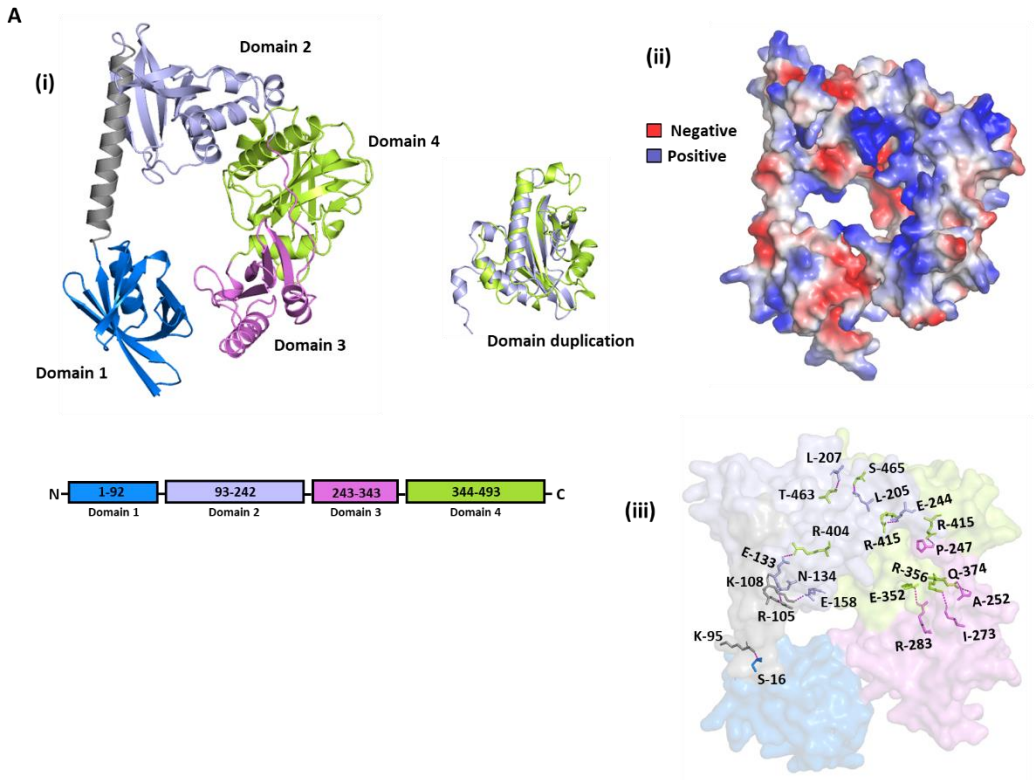
Figure 5.3. Location of DV-Nuc3 NucS gene (Ga0136640_100017) and neighbouring genes from the DV gene contig, with a predicted lineage from Acidobacteria.

The canonical dimeric NucS protein structures from Archaea and Actinobacteria appear to have a similar conformation, unlike DV-Nuc3 NucS which appears to be unique to Acidobacterial species. Modelling other NucS-like proteins from Acidobacteria using AlphaFold indicates they almost all are predicted to have the same structural arrangement as DV-Nuc3, which is consistent with their sequence similarity (data not shown). Unlike the Archaeal and Actinobacterial NucS, which are dimeric, the predicted structure of DV-Nuc3 appears to be monomeric with the N-terminal portion forming two additional domains contributing to an equivalent binding surface as the second monomer in the Archaeal/Actinobacterial NucS. Interestingly, there appears to be a domain duplication between domain two (purple) and domain four (green) of DV-Nuc3 NucS, which when super imposed onto each other, shows a very similar arrangement of α-helices and β-sheets (RMSD 3.238) (**Figure 5.4**, **A, I**). Generation of protein contact potential for DV-Nuc3 protein shows pockets of positive contact sites on the surface of the protein, particularly on the surface of domain two and four. This may indicate a potential area for DNA binding on the surface of domains, two, three and four (**Figure 5.4**, **A, II**). There are numerous interactions between the domains of DV-Nuc3, with a number of these including polar contacts and salt bridges, potentially increasing protein stability (**Figure 5.4**, **A, III**).

In PyMOL the domains of DV-Nuc3 were superimposed separately, onto the dimeric structures of NucS from *T. kodakarensis* and *P. abyssi*. Here it is observed that the N-terminal domain of DV-Nuc3 (blue) overlayed onto the N-terminal domains (pink) of *T. kodakarensis* (RMSD 10.707) and *P. abyssi* (RMSD 7.680) NucS proteins. While the generated RMSD values are relatively high for each superimposition, the overall β-barrel shape and direction of secondary structural elements is retained. Domains two (purple) and four (green) of DV-Nuc3 overlayed onto the C-terminal domains (orange) from T. *kodakarensis* and *P. abyssi* NucS proteins (**Figure 5.4**, **B**). The overlayed domains revealed a very similar arrangement of α-helices and β-sheets, with only minor differences between them. The C-terminal domains from *T. kodakarensis* and *P. abyssi* NucS proteins, contain the catalytic residues, which are shared by DV-Nuc3 NucS protein, hence why these domains have high structural homology

between them. Domains two and four, from DV-Nuc3 are superimposable, but not symmetrically positioned. Hence why when superimposed onto the dimeric forms of *T. kodakarensis* and *P. abyssi* NucS proteins, they positioned onto separate C-terminals of each monomer. Overall domain four gave lower RMSD values, when superimposed onto the C-terminal domains of both *T. kodakarensis* (domain 4 1.770, domain 2 5.002) and *P. abyssi* (domain 4 2.182, domain 2 4.837) NucS proteins.

Sequence alignments of DV-Nuc3 NucS protein, against NucS proteins of *P. abyssi*, *T. kodakarensis* and *Thermococcus gammatolerans*, show very low similarity at the N-terminal region of the protein sequence (**Figure 5.4**, **C**) There are also many gaps in sequence between DV-Nuc3 NucS and the other NucS proteins. With this in mind, there is high conservation of residues within the C-terminal region of the protein sequence, with the same conserved RecB-type motifs, except for the glycine residue in motif II, which is an aspartic acid in DV-Nuc NucS (Aravind et al., 2000). Of interest, all the conserved RecB catalytic motifs are found within domain 4 (green) of DV-Nuc3.



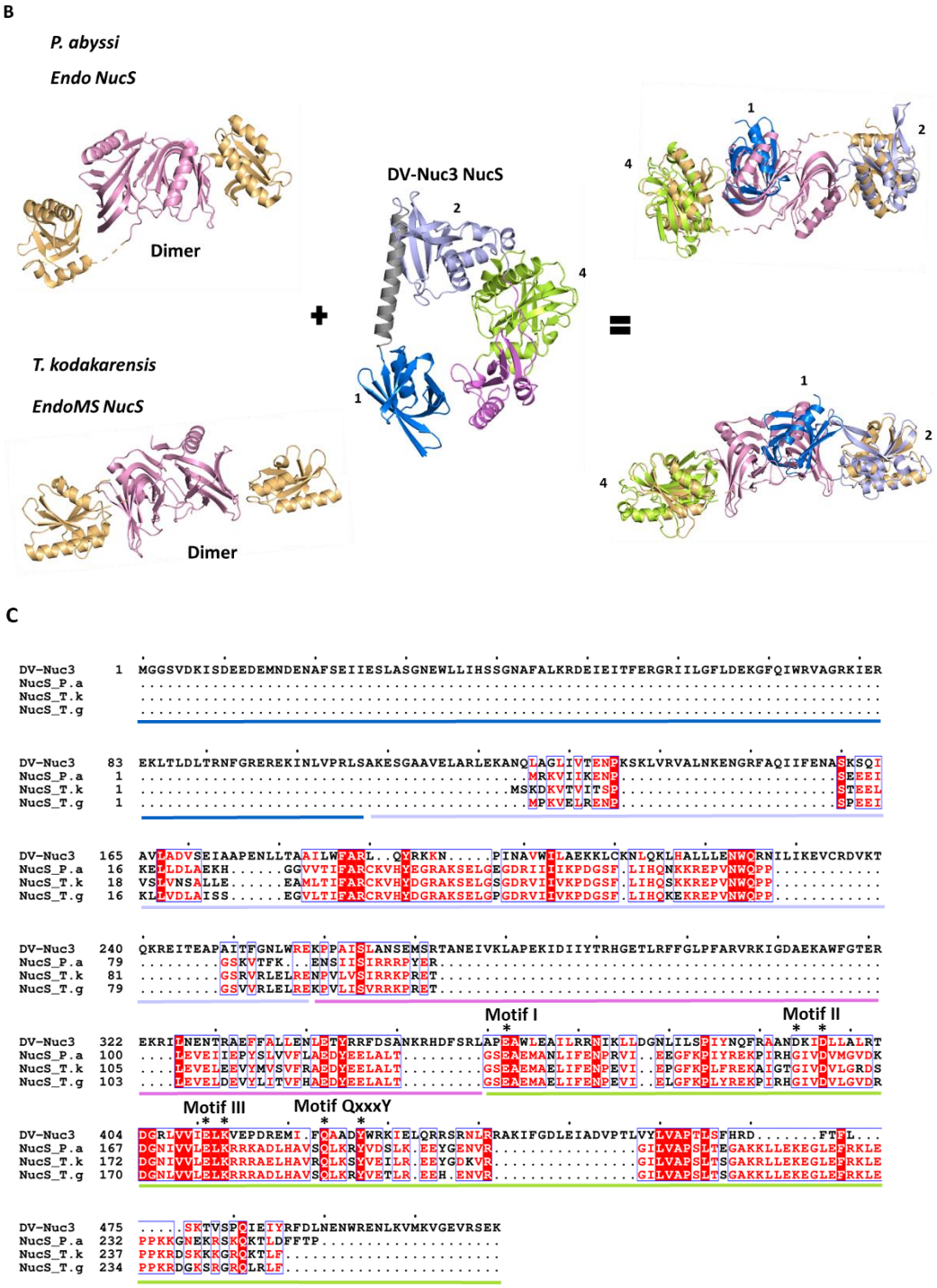


Figure 5.4. Structural arrangement of DV-Nuc3 NucS protein in comparison with other NucS proteins. **A**) (i) Alpha fold prediction of DV-Nuc3 NucS protein, with the four different domains coloured accordingly: domain 1 (blue), domain 2 (purple), domain 3 (pink), domain 4 (green). (ii) Protein contact potential, with positive potential in blue and negative in red. Domain 2 and 4 are domain duplications and are super imposable (iii) Polar contacts between domains of DV-Nuc3 protein. **B**) Domains 1, 2 and 4 of DV-Nuc3 super imposed onto dimeric crystal structures of EndoMS/NucS from *T. kodakarensis* (5GKE) (Nakae et al., 2016) and from *P. abyssi* (2VLD) (Ren et al., 2009). Domains are coloured accordingly: domain 1 (pink), domain 2 (orange). DV-Nuc3 NucS protein structure was generated by AlphaFold2, from Google Colab, version v2.3.1 (Jumper, Evans et al. 2021). The model was presented using PyMOL (Schrödinger, 2020). **C**) Structure guided amino-acid sequence alignment of NucS protein from a Dry-Valley metagenome (metaG UQ864, (DV-Nuc3 NucS)), *P. abyssi, T. kodakarensis* and *T. gammatolerans*. Highly conserved residues are highlighted in red, with a white text and less conserved residues are colored red, with a white background. The active sites residues in the conserved sequence motif of RecB-like nuclease family are indicated by an asterisk (*). Domain arrangement for DV-Nuc3 is represented by coloured lines above the sequences; domain 1 (purple),

domain 2 (orange), domain 3 (green) and domain 4 (blue). Sequence alignment was created using Clustal Omega version 1.2.4. Alignment was created using ESPript 3 (Robert & Gouet, 2014).

5.2.2 DV-Nuc3 protein expression and purification

5.2.2.1 Small scale protein expression testing

The gene sequence for DV-Nuc3 was cloned into pDEST17 (His-tagged) and pHMGWA (MBP-tagged) expression plasmids and transformed into BL21 pLysS *E. coli* expression strains, as described in **Section 2.2.1**. Small scale expression trials were performed, at 15 and 20 ºC and results of these trials were run on SDS PAGEs, as described in **Section 2.4.7**. The best protein expression was seen in pHMGWA expression plasmid, at both 15 and 20 ºC, with results shown below. No protein expression was seen in pDEST17 plasmid (data not shown). Soluble protein expression was observed in pHMGWA plasmid, with soluble protein bound to Ni beads, at both 15 and 20 ºC (**Figure 5.5**).

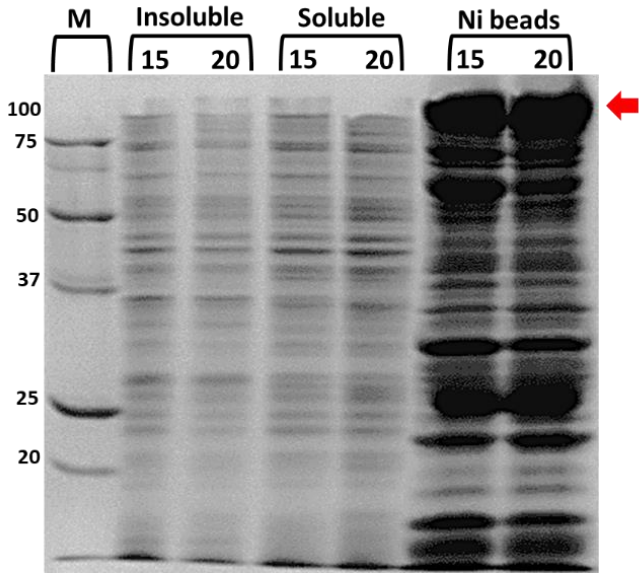


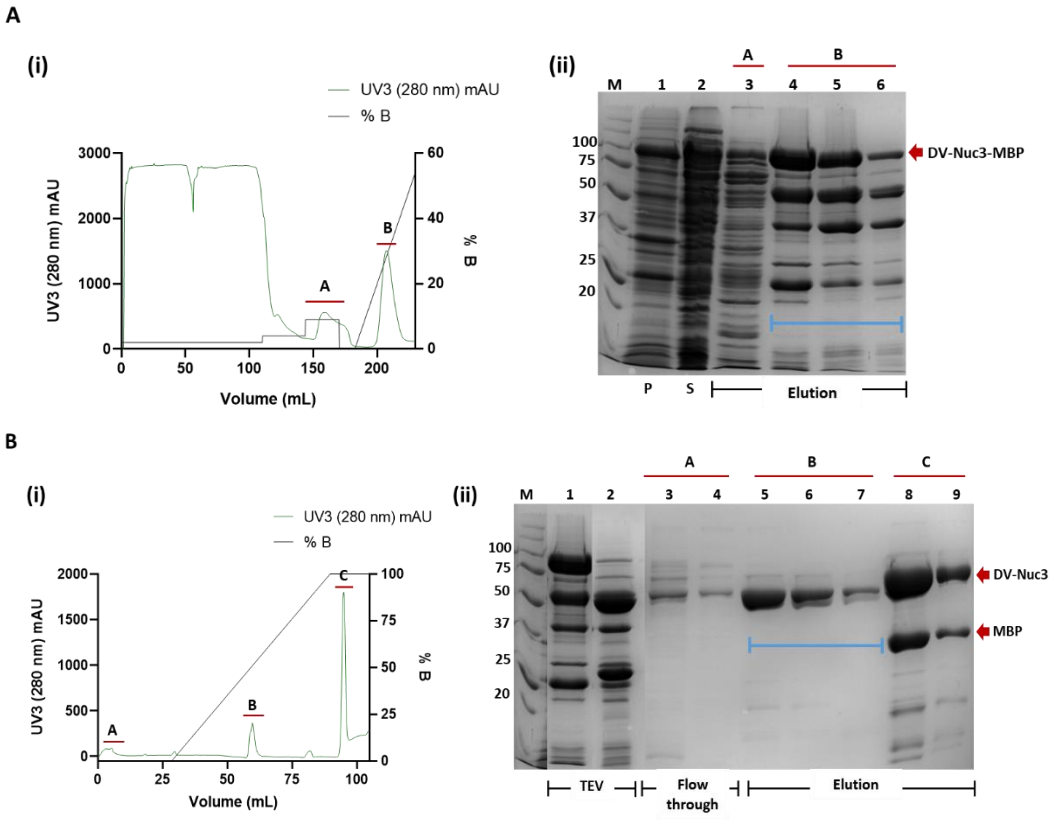
Figure 5.5. SDS PAGE of small scale protein expression results for DV-Nuc3 in pHMGWA plasmid, expressed in BL21 pLysS *E. coli.* Protein expression was tested at 15 and 20 ^{o}C. Results of expression are shown on the gel as insoluble protein, soluble protein lysate and soluble protein bound to Ni beads. Red arrow indicates expression of DV-Nuc3 protein, at the expected size for MBP-tagged (103.4 kDa). A precison plus protein ladder was used as a molecular weight marker (M).

5.2.2.2 Large scale protein purification

Following on from results of soluble protein expression of MBP-tagged (pHMGWA) DV-Nuc3 protein, in small scale screens (**Figure 5.5**), protein

expression cultures were scaled up following methods from **Section 2.3.3**. A three-step purification *via* IMAC and an overnight TEV digest and reverse IMAC, followed by gel filtration chromatography (**Section 2.4**), produced soluble, active protein, suitable for characterisation experiments.

DV-Nuc3$_{MBP}$ eluted off the IMAC column with the addition of 30 mM imidazole (**Figure 5.6**, **A**). An overnight incubation with TEV protease, resulted in 80 % cleavage efficiency of the MBP tag and a reverse IMAC purification resulted in cleaved DV-Nuc3 eluting off the IMAC column, with the addition of 10 mM imidazole (**Figure 5.6**, **B**). Fractions containing DV-Nuc3 protein were further purified by gel filtration chromatography and DV-Nuc3 eluted off the column in two main peaks. Fractions from peak A (**Figure 5.6**, **C**) were up concentrated and stored for further use in characterisation experiments.



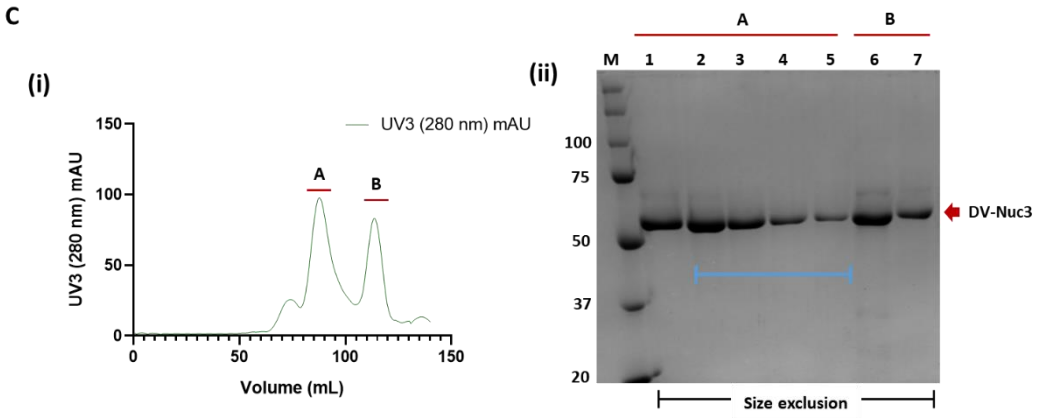


Figure 5.6. IMAC and gel filtration chromatograms (i) and SDS PAGE gels for production of DV-Nuc3 protein, recombinantly expressed from *E. coli* BL21 (DE3) pLysS (ii). **A**) IMAC purification of DV-Nuc3$_{MBP}$. (i) Peak A represents proteins that fell off IMAC column during the 8 % imidazole wash step, peak B represents where the protein of interest (103.5 kDa) eluted during the elution step of the IMAC purification, with 30 mM imidazole. Lanes 1-2 are; insoluble (P), soluble (S) lane 3 represents proteins that eluted during a 8 % imidazole wash step, lane 4-6 are fractions that eluted off the IMAC column during the imidazole gradient. The blue bar indicates fractions that were pooled and incubated overnight with Tev protease. **B**) Reverse IMAC purification of DV-Nuc3. (i) Peak A represents fractions of protein that come through the flowthrough. Peak B represents where the protein of interest (58.8 kDa) eluted off the IMAC column, with 12 % imidazole. Peak C represent protein fraction that eluted off the IMAC column, at a high concentration of imidazole (95 %). (ii) Lanes 1 (before TEV) and 2 (after TEV) show TEV cleave reaction results. Lanes 3-4 represent flow through protein fractions, lanes 5-7 represent de-tagged DV-Nuc3 protein fractions, lanes 8-9 represent protein fractions that eluted off the IMAC column with 95 % imidazole. The blue bar indicates fractions that were pooled and further purified by gel filtration chromatography. **C**) Gel filtration purification of DV-Nuc3. (i) represents protein peaks that eluted off the gel filtration column at different volumes of buffer C. (ii) Lanes 1-5 represent protein fractions from peak A and lanes 6-7 represent protein fractions from peak B. DV-Nuc3 protein is present in both peaks at 58.8 kDa. The blue bar indicates fractions that were up concentrated and stored at -80 ^{o}C. A precison plus protein ladder was used as a molecular weight marker (M). Chromatogram graphs were designed in GraphPad Prism, version 9.0.0.

5.2.3 DV-Nuc3 mutant cloning, expression, and purification

5.2.3.1 Design of DV-Nuc3 mutant

To ensure any nuclease activity observed in activity assays was coming from DV-Nuc3 protein and not contaminants, a mutant protein would need to be used as a control in the assays. The protein sequence of DV-Nuc3 was aligned against other NucS proteins and had the same catalytic aspartic acid (D) in the C-terminus region of the protein, belonging to the conserved motif II (**Section 5.2.1**). A point mutation was made to change the aspartic acid residue, to an alanine, making a null mutation (D397A) (**Figure 5.7**). DV-Nuc3 mutant construct was ordered and cloned into expression plasmids, followed by expression in *E. coli*, as described in **Section 2.2.1**.

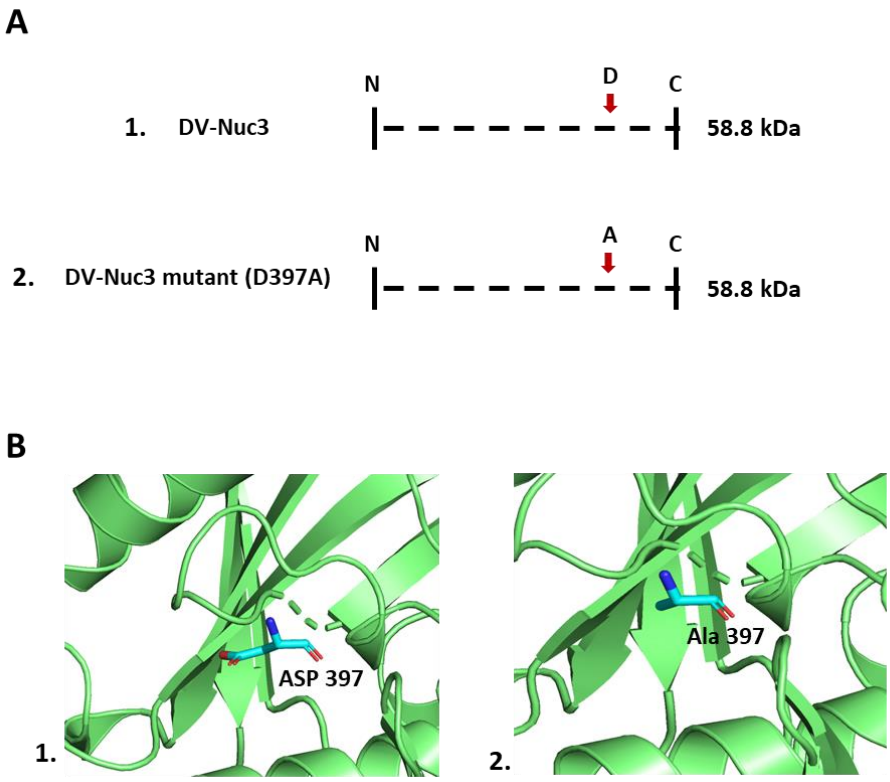


Figure 5.7. Design of DV-Nuc3 mutant (D397A). **A**) schematic of amino acid change for DV-Nuc3 mutant. **B**) Pymol images (Schrödinger, 2020) of AlphaFold predicted (John Jumper, 2021) DV-Nuc3 (1.) with Asp-397 and (2.) mutant with Ala-397.

5.2.3.2 Large scale purification of DV-Nuc3 mutant

Following on from successful small scale protein expression of DV-Nuc3 mutant, in (MBP-tagged) pHMGWA plasmid, expressed in *E. coli* BL21 pLysS (data not shown) protein expression cultures were scaled up following methods from **Section 2.3**. A three-step purification *via* IMAC, reverse IMAC and gel filtration chromatography (**Section 2.4**), produced soluble, active protein, suitable for characterisation experiments. The chromatograms and corresponding SDS-PAGE gels in **Appendix C.4.5** depict the purification, column load and flow through fractions.

5.2.4 Protein folding and stability of DV-Nuc3

The folded structure of DV-Nuc3 protein was investigated using circular dichroism. Secondary structure predictions from both CD spectra and PDBsum analysis of the DV-Nuc3 AlphaFold model were similar with only slight

differences between strand and other percentage contributions. These results provide additional confidence in the accuracy of the AlphaFold predicted structure and confirm that the purified protein is folded (**Figure 5.8**).

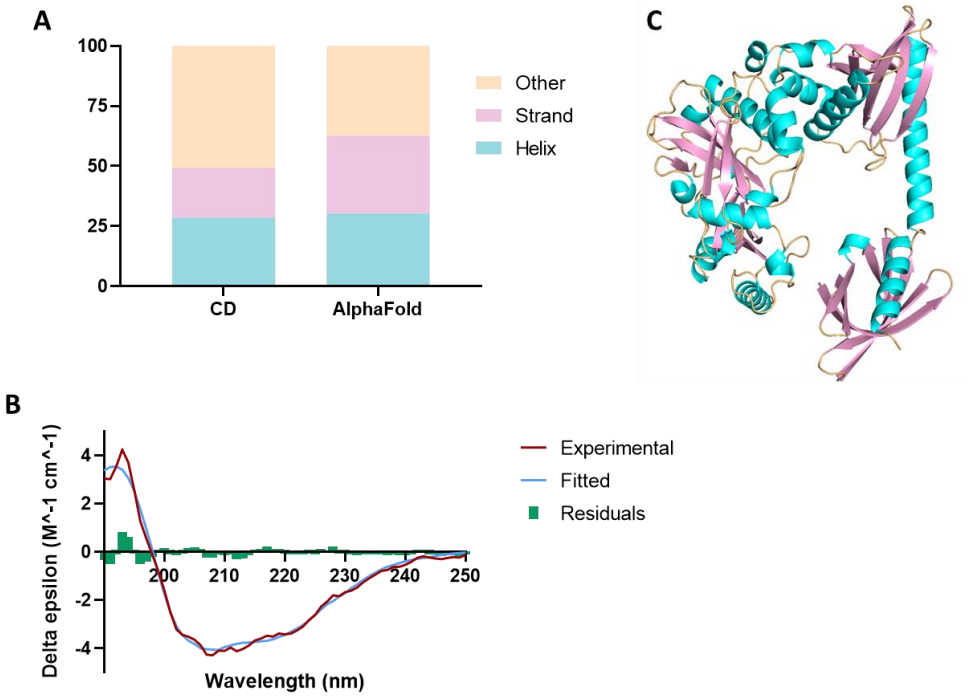


Figure 5.8. Circular dichroism (CD) and AlphaFold secondary structural composition of DV-Nuc3 protein. **A**) A graph showing comparison of secondary structural predictions from CD and AlphaFold prediction model. **B**) Single spectrum analysis of CD spectra, using BeStSel database (Micsonai et al., 2018). **C**) AlphaFold 3D structural prediction of DV-Nuc3, coloured based on secondary structure (Helix in blue, strand in pink and other orange. (John Jumper, 2021). Graphs were produced using Prism version 8 (GraphPadSoftware). Wavelength range (190-250 nm) and scale factor (1). RMSD value (0.2153). NRMSD value (0.02527).

CD thermal melts as described in **Section 2.7** were used to compare the thermal stabilities of DV-Nuc3 wild-type and DV-Nuc3 mutant. DSF as described in **Section 2.8**, was used to generate protein melt curves at different protein concentrations, different pH values and different metal groups. The combined findings of the thermal melt data are presented in **Figure 5.9**.

The CD thermal melt curves revealed that DV-Nuc3 wild-type gave an average T_m of 45 ºC, whereas the DV-Nuc3 mutant gave a slightly lower average T_m of 43 ºC. In the DSF thermal melts, with different protein concentrations, DV-Nuc3 wild-type exhibited the same average T_m as observed in the CD thermal melts. Furthermore, the DSF thermal melts conducted with different pH values indicated that a pH range between 7.5 and 8 yielded the highest T_m values. This

finding suggests an optimal pH range (7.5-8) for protein stability. When various metal ions to DSF experiments, minimal differences in T_m values were observed compared to melts without the addition of metal ions, see **Appendix C.6**.

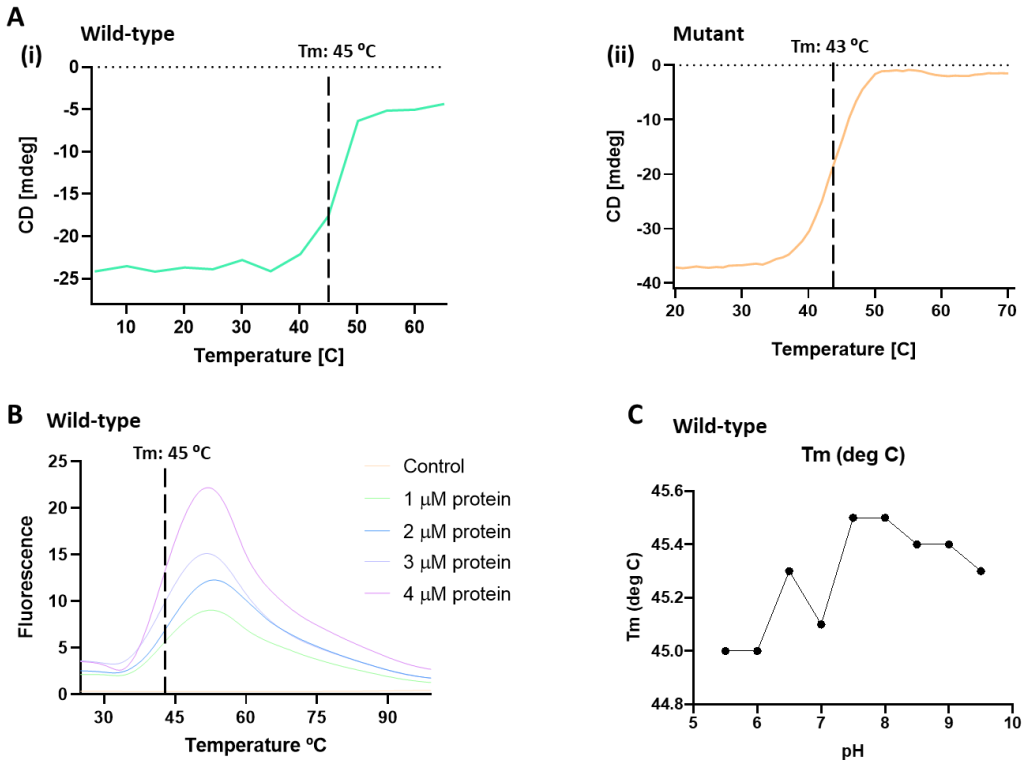


Figure 5.9. Results from thermal melts of DV-Nuc3 and DV-Nuc3 mutant, using CD and DSF. **A**) CD thermal melt data of DV-Nuc3 wild-type (i) and DV-Nuc3 mutant (ii) at 222 nm. T_m values were determined from the midpoint in the unfolding equilibrium and are indicated on the graph, by a dotted line. **B**) DSF, with SYPRO orange, showing melt curve of DV-Nuc3 wildtype at four different protein concentrations (1, 2, 3 & 4 µM). Reactions were carried out in triplicate. T_m values were determined from the midpoint in the unfolding transition and are indicated on the graph, by a dotted line. **C**) First derivative T_m plots, from a DSF, with SYPRO orange, with T_m values derived from the first derivative midpoint of each peak, at different pH values (5.5-9.5). Protein was at a final concentration of 2 µM. Graphs were generated using GraphPad Prism version 8 (GraphPadSoftware).

5.2.5 Biochemical activity characterisation of DV-Nuc3

Nuclease activity of DV-Nuc3 was tested on a range of different DNA substrates, to determine if the protein showed similar substrate specificity and activity, as other characterized NucS homologs. DV-Nuc3 mutant (D397A) was tested alongside DV-Nuc3 to confirm specific nuclease activity observed in reactions, was coming from DV-Nuc3 protein and not *E. coli* contaminants. Activity assays were also performed with different metal ions, varying salt concentrations and different reaction temperatures to determine the optimum conditions for nuclease activity by DV-Nuc3. Initial testing of DV-Nuc3 on

different DNA damage substrates, **Section 5.2.5.5**, indicated that the protein was particularly active on uracil match DNA and this substrate was used in subsequential activity assays to determine optimum conditions.

5.2.5.1 DNA binding by DV-Nuc3

The binding ability of DV-Nuc3 was tested with different DNA substrates (Double stranded (Ds) matched, single stranded (Ss), 3'-tail and 5'-tail, using EMSA, at 15 ºC and 25 ºC. DV-Nuc3 can bind to all DNA substrates, evident by bands at the top of the gel (bound substrate). There isn't a big difference in binding affinity between 15 ºC and 25 ºC. Qualitatively, it appears that DV-Nuc3 binds more effectively to substrates with single stranded segments as a larger amount of these is retained in the wells (**Figure 5.10**, **A**).

A second EMSA was used to identify the binding affinity of DV-Nuc3 on DNA damage/mis-match substrates (8-oxo-dG, Abasic (dSpacer), uracil match, uracil mismatch, A/C mismatch and T/G mismatch), at 15 ºC. Again DV-Nuc3 shows binding across all substrates, particularly with abasic, uracil mismatch and A/C mis-match DNA substrates (**Figure 5.10**, **B**).

A final EMSA was performed on DNA damage/mismatch substrates, using DV-Nuc3 wild-type and DV-Nuc3 mutant protein, to identify if the mutant was still able to bind to DNA substrates. Both DV-Nuc3 and its mutant are capable of binding tightly to all DNA damage substrates (**Figure 5.10**, **C**).

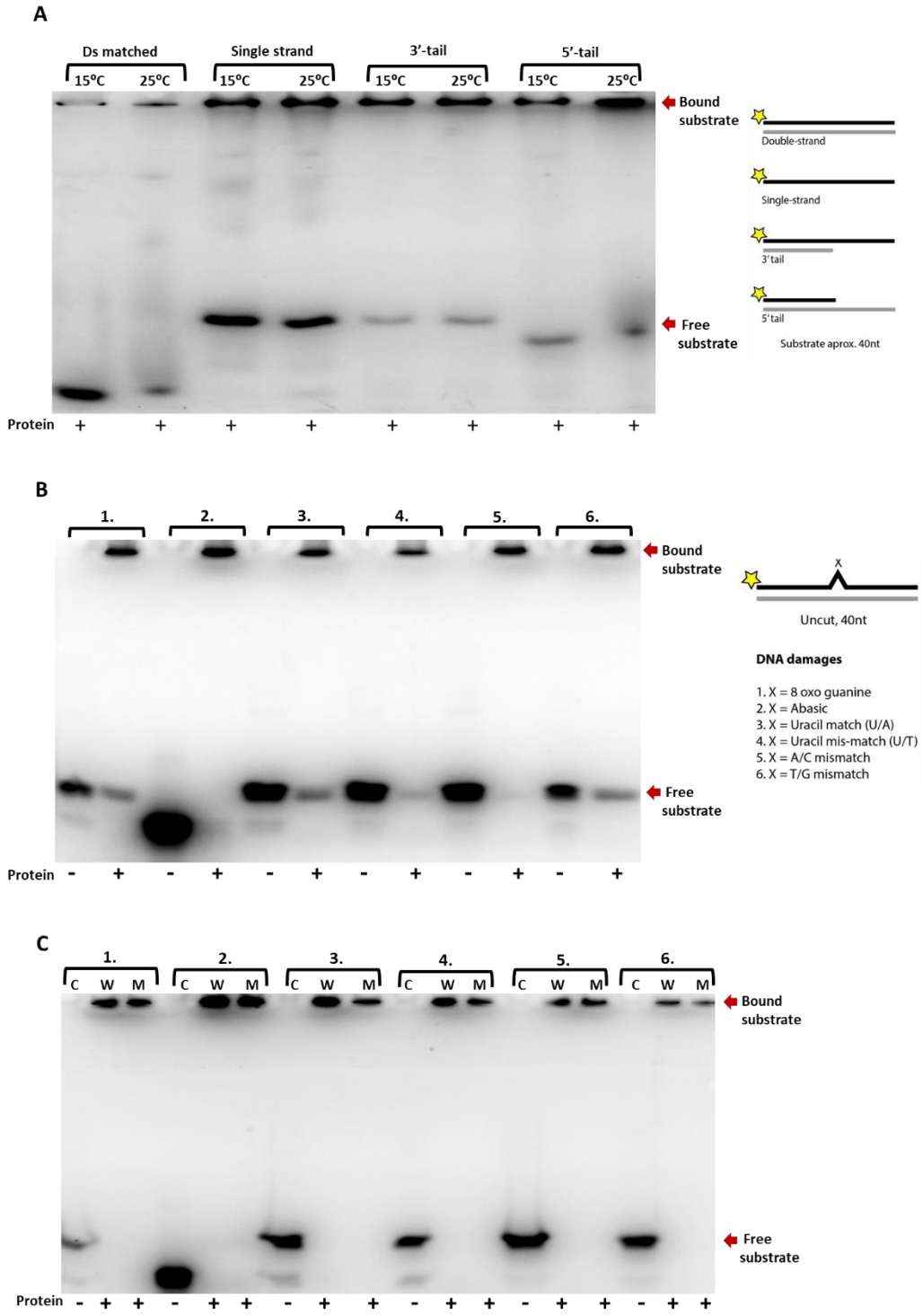


Figure 5.10. Electrophoretic mobility shift assay (EMSA) of DV-Nuc3 protein with DNA substrates, run on a native PAGEs. **A**) EMSA with DV-Nuc3 and un-modified DNA substrates (Ds, Ss, 5'-tail, & 3'-tail DNA). **B**) EMSA with DV-Nuc3 and damaged/mis-matched DNA substrates (8 oxo guanine, abasic, uracil match, uracil mismatch, A/C mis-match, T/G mis-match). **C**) EMSA with DV-Nuc3 wild-type and DV-Nuc3 mutant on damaged and mis-matched DNA substrates. Controls don't contain any protein (-). Free and bound substrate are indicated by red arrows. DNA substrates are indicated to the right of gels. Reactions were incubated at 20 °C (B, C) or 15 °C and 25 °C (A) for 1 hour, with 2.2 µM DV-Nuc3 wild-type and 2.4 µM DV-Nuc3 mutant. Results of EMSAs were visualized using iBright™ CL750 Imaging System, Invitrogen™.

5.2.5.2 Activity of DV-Nuc3 on modified and un-modified DNA substrates

To determine the type of nuclease activity DV-Nuc3 possessed, a range of different DNA substrates were used in a series of activity assays. DV-Nuc3 mutant protein was included in these assays as a control, to ensure nuclease activity observed in DV-Nuc3 samples, was coming from the protein itself and not contaminating nucleases.

The following **Figure 5.11** represents urea PAGEs showing the results of activity assays on double stranded (Ds), single stranded (Ss), 3'-tail and 5'-tail DNA substrate, with DV-Nuc3 wild-type and its mutant. Here a low level of activity was observed on all un-modified substrates, with DV-Nuc3 being most active on Ss DNA. Activity by the mutant was minimal comparatively, except on the Ss DNA substrate.

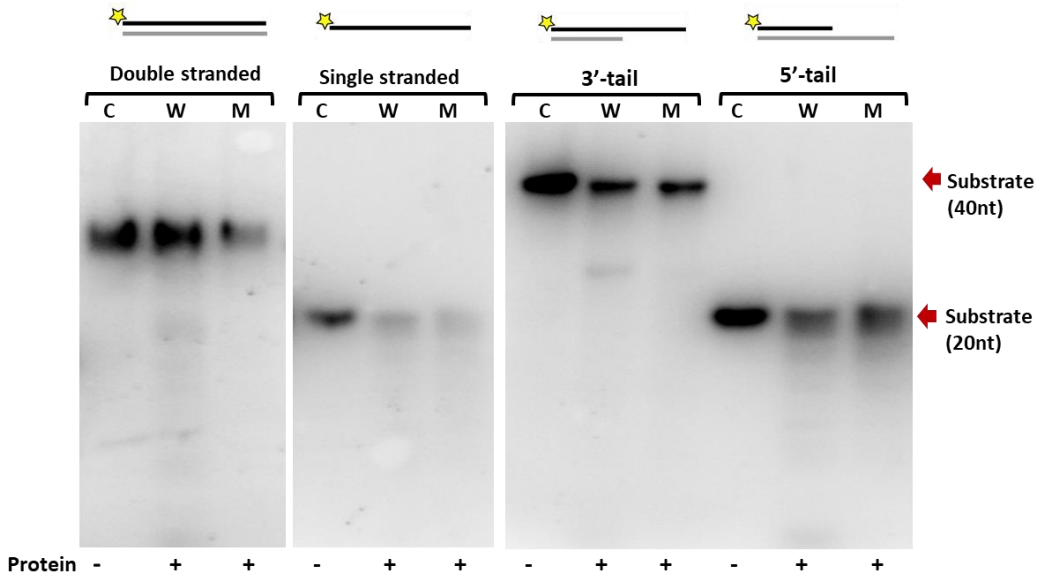


Figure 5.11. Urea PAGE gels of nuclease activity assays on DNA substrates by DV-Nuc3 and DV-Nuc3 (D397A) mutant protein. Samples containing protein are annotated (+) and samples with no protein are annotated (-). For each of the DNA substrates (Double stranded, single stranded, 3'-tail, 5'-tail) there is a control (C) reaction, with no protein, a reaction containing DV-Nuc3 protein (W) and a reaction containing the mutant protein (M). Substrates sizes vary and are indicated by red arrows. Reactions were carried out for 3 hours, at 25 ^{o}C, with 2 µM final concentration for both DV-Nuc3 and mutant and 10 mM final concentration of magnesium. Results of activity assays were visualized using iBright™ CL750 Imaging System, Invitrogen™.

A similar activity assay, from above, was carried out on DNA damage/mis-match substrates (8-oxo-dG, abasic, uracil match, uracil mismatch,

A/C mismatch and T/G mis-match) again with DV-Nuc3 mutant included as a control (Refer to **Section 4.2.12.3** for schematic of DNA substrates). Here nuclease activity is observed with abasic (dSpacer), uracil match, uracil mismatch A/C mismatch and a small amount with T/G mis-match. The best nuclease activity, by DV-Nuc3 wild-type, is observed on abasic DNA substrate. DV-Nuc3 mutant protein only shows a small amount of nuclease activity on abasic DNA substrate (**Figure 5.12**).

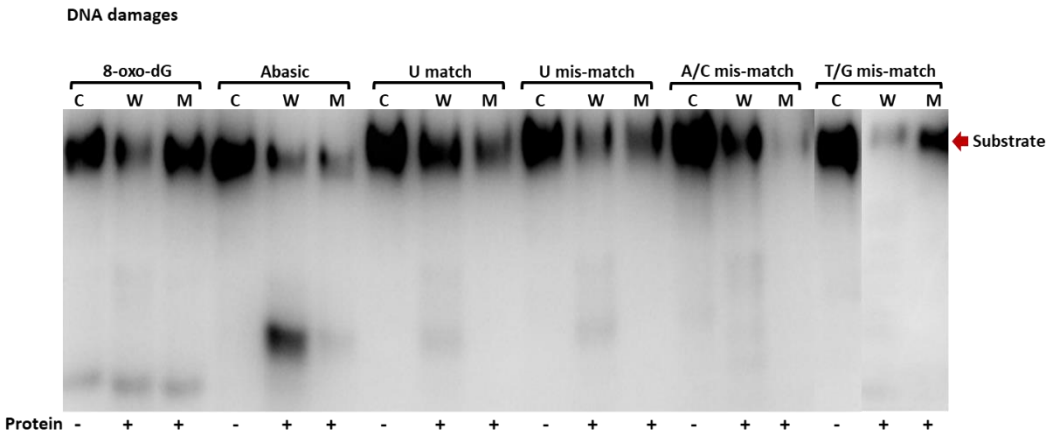


Figure 5.12. Urea PAGE gels of nuclease activity assays on DNA damage substrates by DV-Nuc3 and DV-Nuc3 (D397A) mutant protein. Samples containing protein are annotated (+) and samples with no protein are annotated (-). For each of the DNA substrates (uracil match, uracil mismatch A/C mismatch and T/G mismatch) there is a control (C) reaction, with no protein, a reaction containing DV-Nuc3 protein (W) and a reaction containing the mutant protein (M). Substrates are indicated by a red arrow and cut bands are observed below these substrates. Reactions were carried out for 3 hours, at 25 ^{o}C, with 2 µM final concentration for both DV-Nuc3 and mutant and 10 mM final concentration of magnesium. Results of activity assays were visualized using iBright™ CL750 Imaging System, Invitrogen™.

A last set of DNA substrates (flapped 3', flapped 5' and splayed) were tested for nuclease activity with DV-Nuc3 protein (**Figure 5.13**). Here no specific cut sites are observed, instead there is smearing of the DNA substrates, indicating non-specific activity.

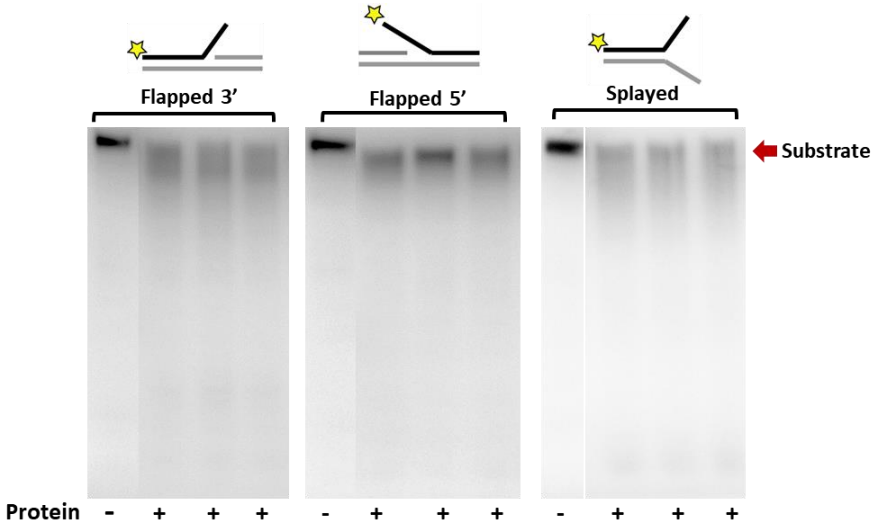


Figure 5.13. Urea PAGE gels showing nuclease activity assays on flapped 3', flapped 5' and splayed DNA substrates by DV-Nuc3 protein. Samples containing protein are annotated (+) and samples with no protein are annotated (-). Reactions were run in replicates of 3. Substrates are indicated by a red arrow and cut bands are observed below these substrates. Reactions were carried out for 10 hours, at 25 ^{o}C, with 2 µM final concentration of DV-Nuc3 protein and 10 mM final concentration of magnesium. Results of activity assays were visualized using iBright™ CL750 Imaging System, Invitrogen™.

5.2.5.3 Metal ion preference of DV-Nuc3

NucS proteins are predicted to require a metal ion for nuclease activity on DNA substrates. DV-Nuc3 wild-type protein, along with DV-Nuc3 mutant protein, were used in a series of activity assays on uracil match DNA substrate, with different metal ions (zinc, manganese, and magnesium).

These results show that nuclease activity is observed without the addition of a metal ion and this activity is also seen with the addition of EDTA. Based off these results, activity seen upon the addition of zinc, at a low concentration (1 mM), appears to be caused by either a contaminating *E. coli* nuclease or by a contaminant metal ion already bound to the active site of DV-Nuc3. Increasing the concentration of zinc in the reactions appears to inhibit nuclease activity on the DNA substrate. Addition of manganese to reactions results in almost complete degradation of the DNA substrate, with this degradation increasing with higher concentrations of manganese. Reactions containing magnesium show a slight increase in product band, compared to reactions with no metal ion. Increasing the concentration of magnesium in the reactions results in a decrease in degradation of substrate. With the addition of 20 mM magnesium, there is a smaller size band under the substrate, and it is unclear if this band is resulting from nuclease activity

or degradation of substrate, due to a high concentration of magnesium (**Figure 5.14**).

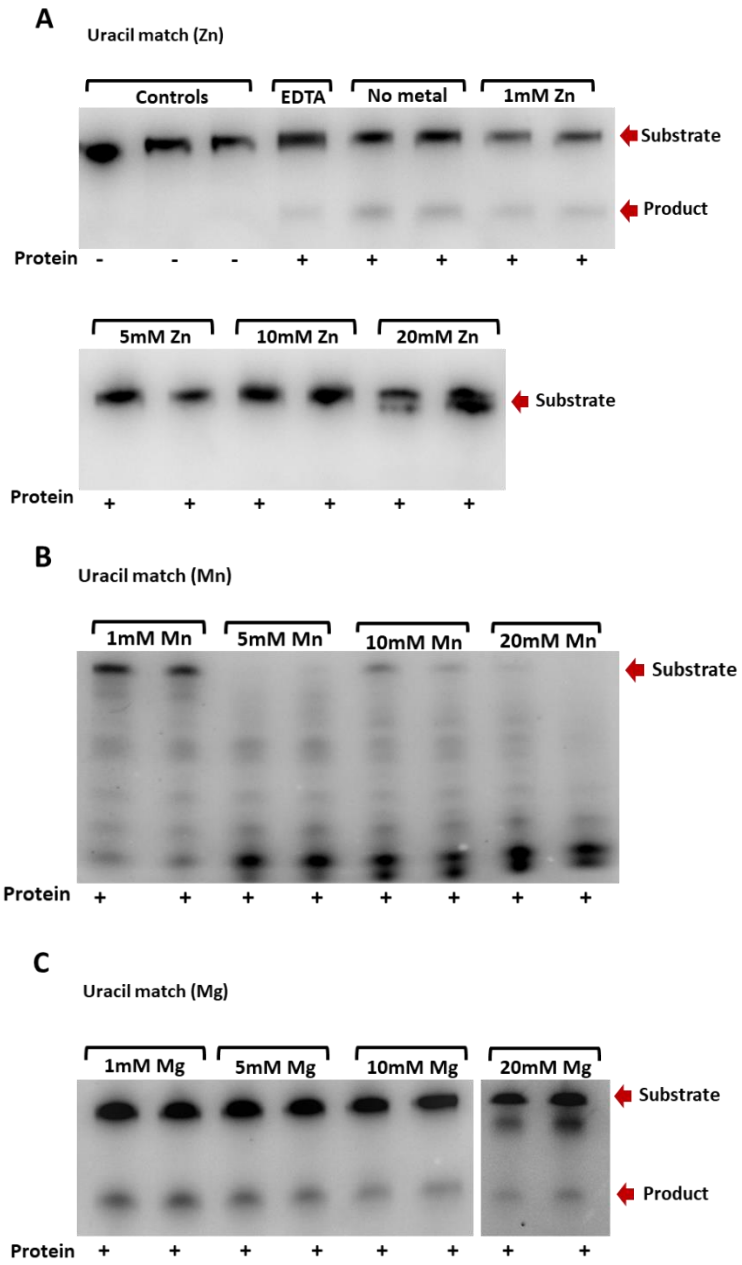


Figure 5.14. Urea PAGE gels showing nuclease activity assays on uracil match DNA substrate by DV-Nuc3 protein, with different metal ions. **A**) represents activity assays results on uracil match DNA with the addition of different zinc (Zn) ion concentrations. **B**) represents activity assays results on uracil match DNA with the addition of different manganese (Mn) ion concentrations. C) represents activity assays results on uracil match DNA with the addition of different magnesium (Mg) ion concentrations. Controls don't contain any protein (-). The reaction with EDTA contains 10 mM magnesium and protein (+). No metal reactions, contain protein (+) but no metal ion. Reactions were run in replicates of 2. Substrates are indicated by a red arrow and product bands are observed below these substrates, with some product being indicated by a red arrow. Reactions were carried out for 8 hours, at 25 ^{o}C, with 2 µM final concentration of DV-Nuc3 protein and varying concentrations of metal ions (1, 5, 10 and 20 mM). Results of activity assays were visualized using iBright™ CL750 Imaging System, Invitrogen™.

These activity assays on different metal cofactors, were repeated with the addition of DV-Nuc3 mutant, to ensure degradation was coming from DV-Nuc3 protein and not contaminating nucleases. As there was no obvious activity with zinc as a metal cofactor, only magnesium and manganese were used for the following activity assays.

Here nuclease activity was seen only in reactions with DV-Nuc3 wild-type protein and not in the reaction with DV-Nuc3 mutant. Although there was some nuclease activity observed in the reactions without the addition of a metal ion, this was less compared to reactions with the addition of magnesium and particularly manganese. Reactions containing magnesium showed improved nuclease activity on uracil match DNA over time, with potentially longer incubation periods required for sufficient activity on substrate. Reactions with the addition of manganese showed nuclease activity, on uracil match DNA occurring just after 30 minutes of incubation with protein. DV-Nuc3 activity on uracil match DNA, with the addition of manganese appears to be non-specific, with the generation of multiple product bands. While activity, with the addition of magnesium, is more specific, with less product bands present on the gel (**Figure 5.15**).

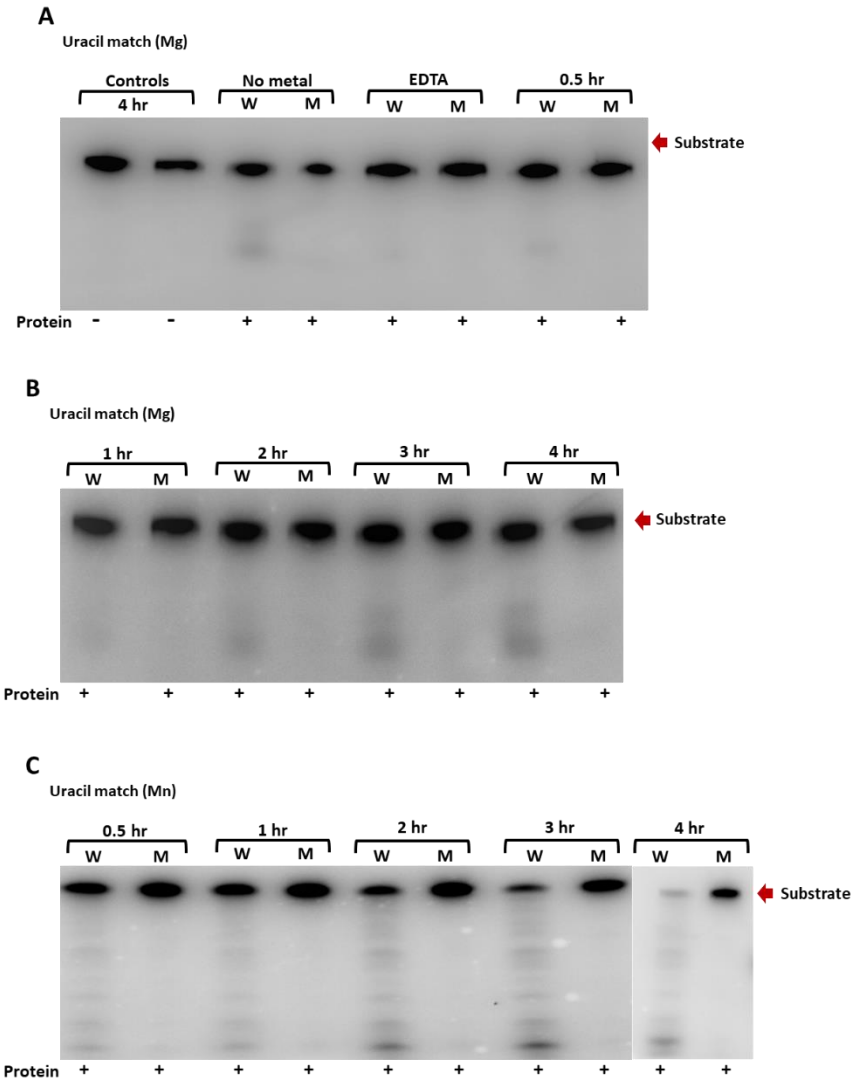


Figure 5.15. Urea PAGE gels showing nuclease activity assays on uracil match DNA substrate by DV-Nuc3 wild-type and DV-Nuc3 mutant, with different metal ions. Reactions with protein were incubated for 30 minutes and 1-4 hours. **A**) represents activity assays results on uracil match DNA with controls and addition of magnesium (Mg). **B**) represents activity assays results on uracil match DNA with the addition of magnesium (Mg) ion at different incubation periods (1-4 hours). **C**) represents activity assays results on uracil match DNA with the addition of manganese (Mn) ion at different incubation periods (30 mins, 1-4 hours). Controls don't contain any protein (-). The reaction with EDTA contains 10 mM magnesium and protein (+). No metal reactions, contain protein (+) but no metal ion. Substrates are indicated by a red arrow. Reactions were carried out at different incubation periods, at 20 ^{o}C, with 1.2 µM final concentration of DV-Nuc3 wild-type and 1.4 µM final concentration of DV-Nuc3 mutant and 10 mM final concentration of metal ions. Results of activity assays were visualized using iBright™ CL750 Imaging System, Invitrogen™.

Further activity assays were performed with DV-Nuc3 wild-type and DV-Nuc3 mutant on uracil match DNA substrate, with the addition of magnesium or manganese metal ions. Activity assays were carried out for four or sixteen hours, to ensure sufficient activity by proteins on DNA substrate. Results from these activity assays were visualized on native PAGEs so that the duplex was not denatured, to distinguish between double and single stranded (**Figure 5.16**). Here

results show that with the addition of manganese to reaction with DV-Nuc3 wild-type, both strands of substrate are degraded through nuclease activity, compared to one strand with the addition of magnesium. There is some nuclease activity in DV-Nuc3 mutant reactions with the addition of manganese, but this activity is low comparatively to reactions with DV-Nuc3 wild-type.

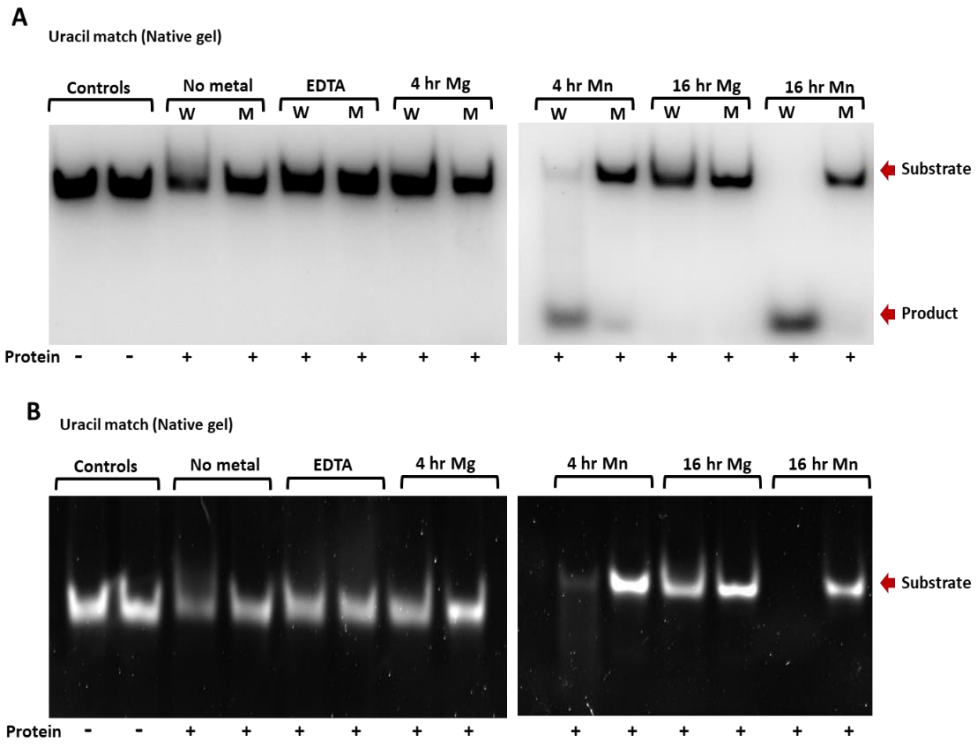


Figure 5.16. Native PAGE gels showing results of nuclease activity on uracil match DNA, by DV-Nuc3 wild-type and mutant. Reactions were incubated for 4 or 16 hours, with the addition of magnesium or manganese. **A**) Results of nuclease activity was visualized using the fluorescent blot setting on the iBright™ CL750 Imaging System, Invitrogen™. **B**) Results of nuclease activity visualized using SYBR Gold nucleic acid stain followed by imagining on the iBright™ CL750 Imaging System, Invitrogen™. Controls don't contain any protein (-). The reaction with EDTA contains 10 mM magnesium and protein (+). No metal reactions, contain protein (+) but no metal ion. Substrate and product are indicated by red arrows. Reactions were carried out at different incubation periods, at 20 ^{o}C, with 1.2 µM final concentration of DV-Nuc3 wild-type and 1.4 µM final concentration of DV-Nuc3 mutant and 10 mM final concentration of metal ions. Results of activity assays were visualized using iBright™ CL750 Imaging System, Invitrogen™.

Cumulative results of nuclease activity on uracil match DNA, visualized on denaturing gels (**Figure 5.14 and 5.15**) and native gels (**Figure 5.16**), suggest that nuclease activity by DV-Nuc3, with the addition of manganese shows non-specific degradation of substrate, rather than a specific cut at the uracil lesion, that is observed with the addition of magnesium.

5.2.5.4 Activity by DV-Nuc3 at varying salt concentrations

Nuclease activity on uracil match DNA substrate, by DV-Nuc3 protein, with different concentrations of NaCl, were evaluated by activity assays. Here there is little change in nuclease activity over the increasing concentrations of NaCl, although best activity is observed in reactions containing a lower NaCl concentration. (**Figure 5.17**).

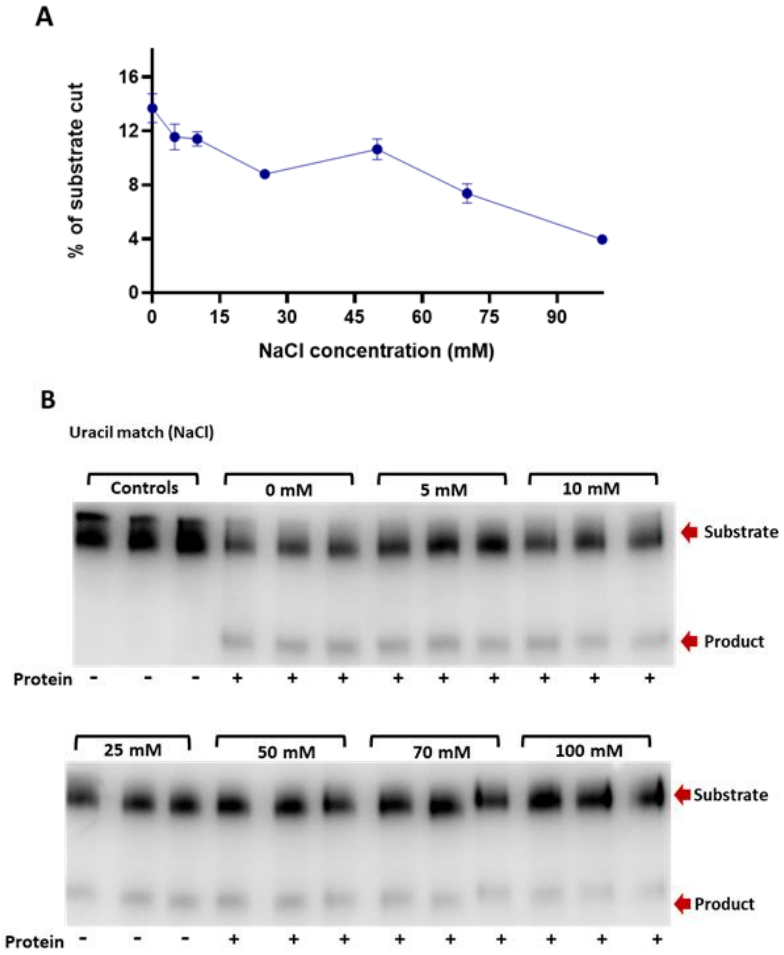


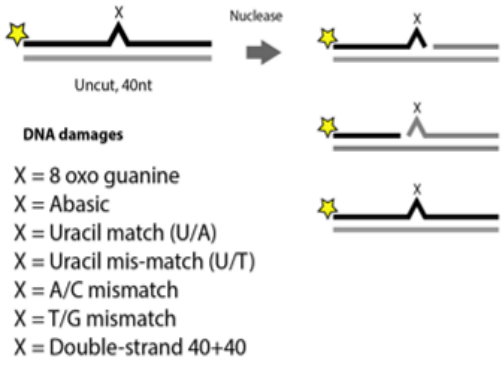
Figure 5.17. Urea PAGE gels showing nuclease activity assays on uracil match DNA substrate by DV-Nuc3, with increasing concentrations of NaCl salt (0-100 mM). **A**) Results of activity assay visualized on a quantitative graph. Points on the graph represent percentage average of product formation in reactions. Standard deviation error bars are included. **B**) Results of nuclease activity by DV-Nuc3 on uracil match DNA, with varying concentrations of NaCl salt. Controls don't contain any protein (-) and contain 100 mM NaCl salt buffer. Substrate and product are indicated by red arrows. Reactions were carried out for 8 hours, at 20 ^{o}C, with 2.1 µM final concentration of DV-Nuc3 and 10 mM final concentration of magnesium ion. Results of activity assays were visualized using iBright™ CL750 Imaging System, Invitrogen™.

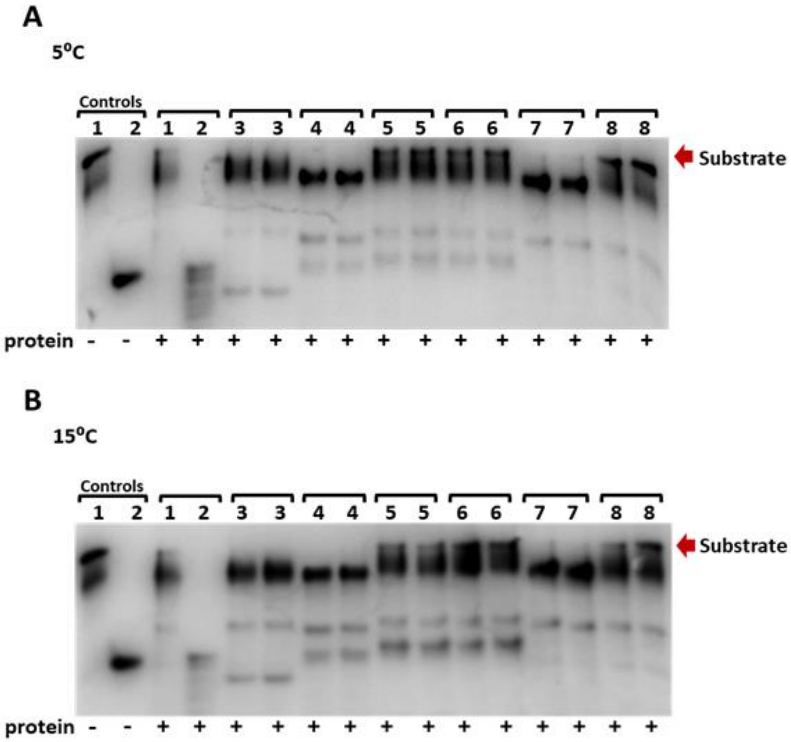
5.2.5.5 Temperature dependence of DV-Nuc3 activity

The temperature dependence of nuclease activity by DV-Nuc3 protein was tested on different DNA damage and mismatch substrates, from 5 to 40 ºC, with magnesium as the metal ion cofactor.

Here nuclease activity is observed across all temperature ranges, with best activity seen at 30 ºC. At 40 ºC, nuclease degradation is reduced compared to at 30 ºC. Degradation activity observed here shows specific cutting at DNA damage and mismatch lesion sites on the substrates. DV-Nuc3 is particularly active on uracil match and uracil mis-match DNA substrates, with degradation of substrates being most prominent in the 30 ºC activity assays (**Figure 5.18**).

DNA substrates	
1.	Double stranded matched
2.	20+20 single stranded
3.	8-oxo-dG linear
4.	Abasic (dspacer) linear
5.	Uracil match
6.	Uracil mis-match
7.	Regular mis-match A/C
8.	Regular mis-match T/G





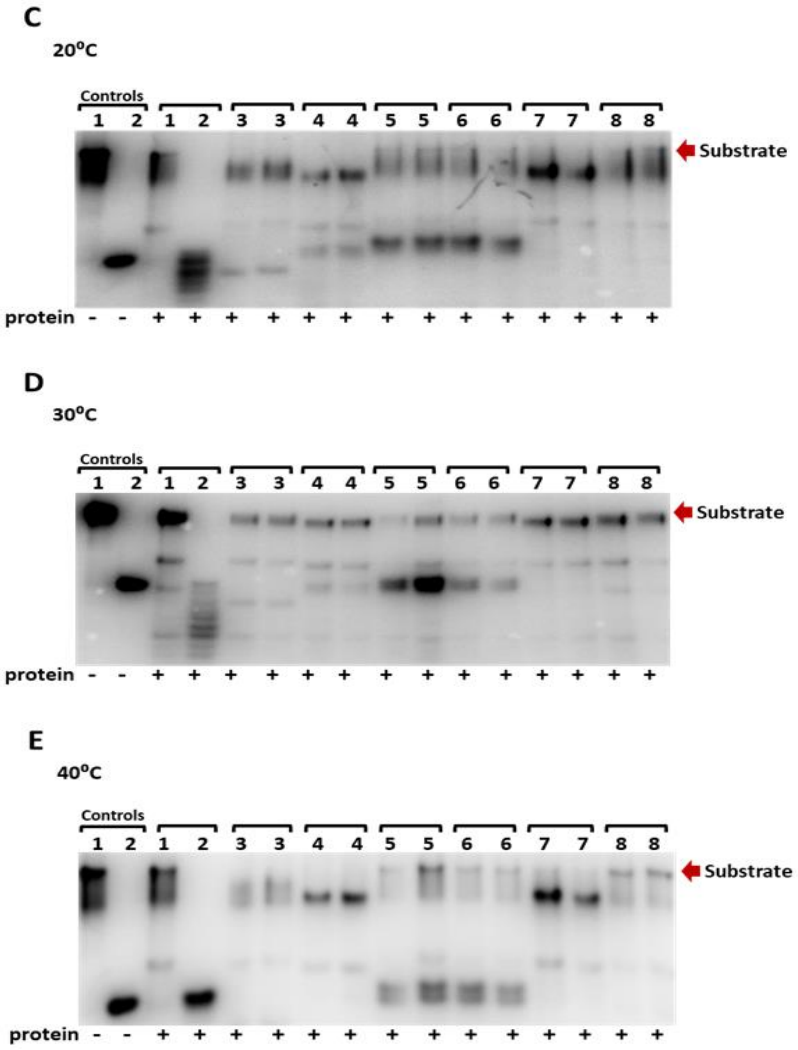


Figure 5.18. Urea PAGE gels of nuclease activity assays on DNA substrates by DV-Nuc3 protein at different incubation temperatures (5, 15, 20, 30 & 40 °C). Samples containing protein are annotated (+) and samples with no protein are annotated (-). Lanes labelled 1 contain double stranded matched DNA, lanes labelled 2 contain 20+20 single stranded DNA, lanes labelled 3 contain 8-oxo-dG linear, lanes labelled 4 contain abasic DNA, lanes labelled 5 contain uracil match DNA, lanes labelled 6 contain uracil mismatch DNA, lanes labelled 7 contain regular mismatch A/C and lanes labelled 8 contain regular mismatch T/G) DNA. DNA substrates are indicated by a red arrow, except for the 20+20 single stranded DNA substrates, which runs as a smaller substrate (Lanes 2). Reactions were carried out for 16 hours, at the annotated temperature, with 2 µM final DV-Nuc3 protein concentration and 10 mM final concentration of magnesium. Results of activity assays were visualized using iBright™ CL750 Imaging System, Invitrogen™.

Additional temperature gradient activity assays were performed, with uracil match DNA substrate, ranging from 5 to 80 ºC. Unfortunately, in reactions above 50 ºC, the integrity of the substrate was affected, potentially by the high temperature with the addition of magnesium and a similar banding pattern under the substrate was observed across all reactions. No product band, at the expected size was seen in these reactions, which indicates no nuclease activity by the

protein was occurring. However, due to controls not being incubated at these higher temperature ranges, these results are inconclusive (**Appendix C.5**).

5.3 DV-Nuc3 N-terminal truncation

5.3.1 Construct design

A construct was designed for DV-Nuc3 to remove the N-terminal domain from the protein. It has been hypothesised, **Section 5.2.1**, that the N-terminal domain is important for DNA binding. Here 137 amino acids were removed from the N-terminal sequence of DV-Nuc3, to completely remove the first domain and the helical linker from the protein (**Figure 5.19**). The DV-Nuc3 N-terminal truncation construct was ordered and cloned into expression plasmids, followed by expression in *E. coli*, as described in **Section 2.2.1**.

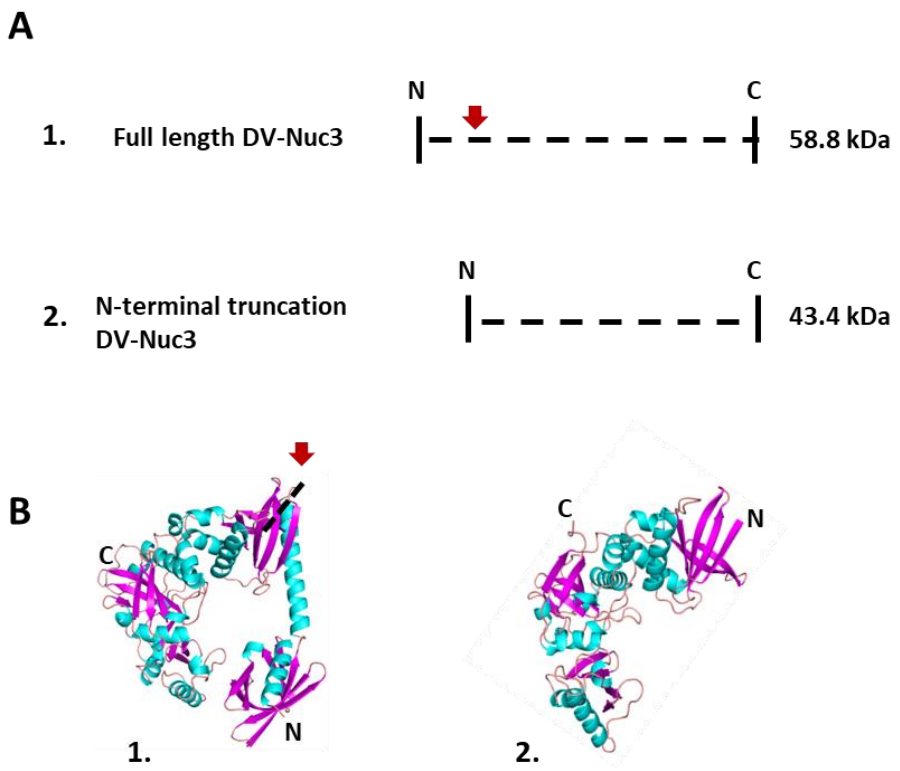


Figure 5.19. Design of DV-Nuc3 N-terminal truncation. **A**) schematic of location of new start site for DV-Nuc3 truncation protein and the size differences between original and new construct. **B**) Pymol images (Schrödinger, 2020) of AlphaFold predicted (John Jumper, 2021) DV-Nuc3 (1.) with N-terminal domain and (2.) removal of N-terminal domain from protein.

5.3.2 Small scale expression trials

The gene for DV-Nuc3 N-terminal truncation was cloned into pDEST17 (His-tagged) and pHMGWA (MBP-tagged) expression plasmids and transformed into BL21 pLysS, BL21 (DE3) and arctic express *E. coli* expression strains, as described in **Section 2.2.1**. Small scale expression trials were performed on all strains, with the greatest soluble protein expression seen in BL21 pLysS cells (**Figure 5.20**). Here soluble protein expression was observed for both His-tagged and MBP-tagged DV-Nuc3 truncation protein, with best expression seen for cultures grown at 15 ºC.

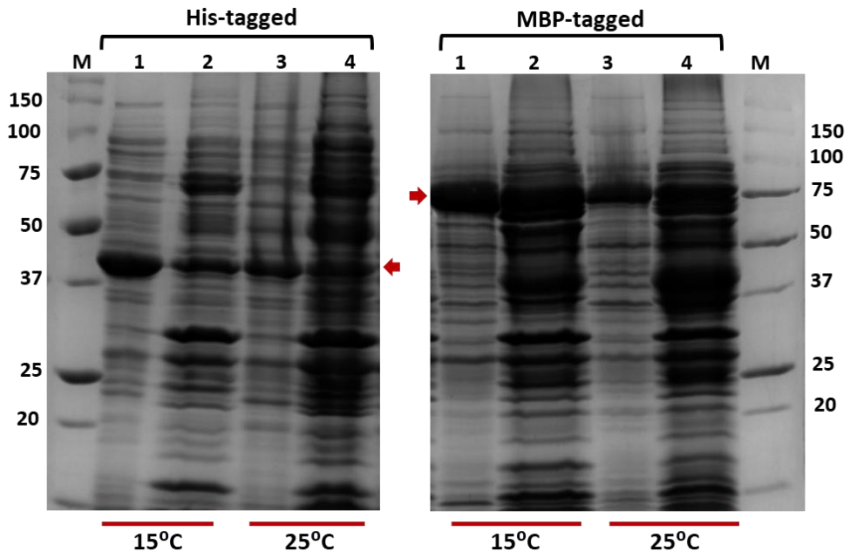


Figure 5.20. SDS PAGE gels of small-scale protein expression results for DV-Nuc3 N-terminal truncation in pDEST17 (His-tagged) and pHMGWA (MBP-tagged) plasmids, expressed in BL21 pLysS *E. coli*. Protein expression was tested at 15 and 25 °C. Results of expression are shown on the gel as insoluble protein (lanes 1 & 3) and soluble protein bound to Ni beads (lanes 2 & 4). Red arrows indicates expression of DV-Nuc3 truncation protein, at the expected size for His-tagged protein (46.9 kDa) and MBP-tagged protein (87.3 kDa). A precison plus protein ladder was used as a molecular weight marker (M).

5.3.3 Large scale purification

Following on from results of soluble protein expression of DV-Nuc3 truncation , in small scale screen (**Figure 5.20**), protein expression cultures were scaled up, following methods from **Section 2.3.3**. His-tagged protein is easier to purify, as the his-tag can be left on the protein.

An IMAC purification of DV-Nuc3 truncation showed expressed protein, however the protein was mostly insoluble and remained in inclusion bodies

(**Figure 5.21**). Due to the low expression level of soluble protein, no further purifications were attempted.

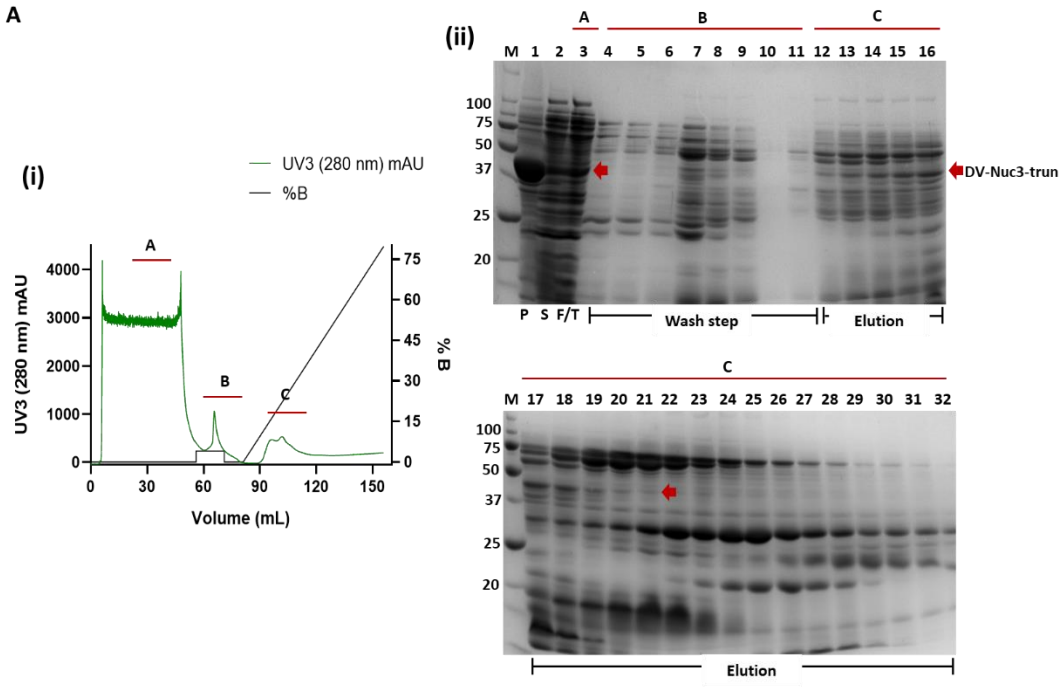


Figure 5.21. IMAC chromatogram (i) and SDS PAGE gel for production of DV-Nuc3 truncation protein, recombinantly expressed from *E. coli* BL21 (DE3) pLysS (ii). Peak A represents protein fractions from the flow through, Peak B represents protein fractions, from a 4 % imidazole wash step, peak C represents protein fraction that eluted from the IMAC column, with 12 % imidazole. Lanes 1-3 are; insoluble (P), soluble (S) and flowthrough (F/T). Lanes 4-11 represent protein fraction that eluted from the column during a 4 % imidazole wash step. Lanes 12-32 represent protein fractions that eluted from the IMAC column, during the imidazole gradient, starting from 12 %. Red arrows indicate the presence of DV-Nuc3 truncation protein (43.4 kDa). Chromatogram graph was designed in GraphPad Prism, version 9.0.0.

Due to time restraints on this project, no further work was carried out with the truncated DV-Nuc3 protein. However, future plans for this protein will involve purifying the MBP-tagged DV-Nuc3 truncation protein, to determine if this construct will produce soluble expressed protein.

5.4 Discussion

DV-Nuc3 NucS was identified from the DV-metagenomes and originates from the bacterial lineage of Acidobacteria, with 36.6 % identify to *Pyrinomonas methylaliphatogenes*. DV-Nuc3 was initially annotated as a hypothetical protein from the IMG annotation pipeline and no functional attributes were assigned to the protein. Bioinformatic investigation using tools such as hmmsearch and SSN analysis, revealed DV-Nuc3 is a homolog of NucS type endonucleases (PF01939).

Alignments with DV-Nuc3 against known NucS homologs indicates that the N-terminal sequence of DV-Nuc3 is extended relative to the characterized NucS homologs. The best aligned portion are the final 200 C-terminal residues, which correspond to the RecB-like endonuclease domain of structurally-characterized NucS/EndoMS of *T. kodakarensis* (Rzoska-Smith et al., 2023; Zhang et al., 2020). There is little conservation of gene synteny for identified NucS proteins, across different species. Within the *Thermococcus* species, there was conservation of a RadA gene, found either just up or down stream of the NucS gene, along with other conserved genes; a DUF473 domain containing protein, a proteasome assembly chaperon family protein, and an S-methyl-5'-thioadenosine phosphorylase. Investigations into the genomes of *Pyrococcus* species, containing a NucS protein, revealed that the RadA gene was missing, but the same DUF473 domain containing protein, proteasome assembly chaperon family protein and S-methyl-5'-thioadenosine phosphorylase genes were present. There was also conservation of a DUF63 family protein gene, across these genomes, from *Pyrococcus* species (Nakae et al., 2016). In the gene contig of DV-Nuc3 and other gene contigs from Acidobacteria, belonging to DV-metagenomes, there was always conservation of a gene encoding a Lysophospholipase, found in close proximity to the NucS gene.

The modelled structure of DV-Nuc3 has four discrete folded domains, where domains one, two and four align with previously characterized NucS proteins. Further, there appears to be a domain duplication, where domain two has pseudo-symmetry with domain four. The domains two-four make a cup-shaped structure which potentially may clasp the DNA duplex, while a more mobile N-terminal domain (domain one) likely moves to allow binding. DV-Nuc3 appears to function as a monomer, with only one active site present, in contrast to previously studied homologs which are structural and functional dimers, with two active sites. Overlay of DV-Nuc3 predicted model with *P. abyssi* and *T. kodakarensis* NucS dimers, reveals that the RecB-like domains occupy equivalent positions to the duplicated domains two and four from DV-Nuc3. The duplicated domains of DV-Nuc3 superimpose well with each other and domains two and four superimpose well with RecB domains of *P. abyssi* and *T. kodakarensis* NucS dimers, including the active -site aspartic acid residue in domain four. Domain

one from DV-Nuc3 also superimposes onto the N-terminals domains *P. abyssi* and *T. kodakarensis,* but give high RMSD values, indicating differences between these domains. The polypeptide sequence of DV-Nuc3 aligned poorly against other NucS proteins, in the N-terminal region, but there was high conservation in the C-terminal portion, particularly in domains two and four, that have homology to the C-terminal domains of *P. abyssi* and *T. kodakarensis* NucS proteins. Domain four contains four highly conserved RecB-type-motifs, comprising mostly negatively charged amino acid residues, found across other NucS proteins from Archaea and bacteria. However, the expected glycine residue in motif I, is an aspartic acid in DV-Nuc3 NucS.

In DNA binding experiments DV-Nuc3 was able to bind to both double and single stranded DNA substrates, with a preference for single stranded DNA. Binding was also observed with 3'-tail and 5'-tail DNA substrates, with no obvious preference between the two. DV-Nuc3 also showed the ability to bind to a wide range of DNA damage substrates, with the best binding observed on abasic, uracil mismatch (U/A) and A/C mismatch DNA substrates. Binding of DNA damage substrates was also observed by DV-Nuc3 mutant (D397A), indicating that mutation of the aspartic acid residue to an alanine had no effect on the binding ability of the protein to DNA. In activity assays DV-Nuc3 was active on all un-modified DNA substrates (Ds, Ss, 5'-tail and 3'-tail). DV-Nuc3 was more active on Ss DNA, compared to Ds, with most of the substrate been degraded. A specific cutting pattern was observed on 3'- tail DNA, was non-specific degradation was observed on 5'-tail DNA. Apparent endonuclease activity was observed on several DNA damage substrates, with the best activity observed on abasic, uracil match, uracil mismatch and A/C mismatch DNA substrates. The DV-Nuc3 mutant had either minimal or no activity on these DNA substrates, indicating that the nuclease activity observed from DV-Nuc3 reactions were a result of the proteins activity and not likely from contaminants. DV-Nuc3 was also active on flapped (3', or 5') and splayed DNA substrates, showing non-specific degradation activity.

NucS homologs have been found across Archaea and some bacterial species. NucS homologs from *T. kodakarensis, P. furiosus, P. abyssi* , and

C. glutamicum have been biochemically characterized (Creze et al., 2011; Ishino et al., 2018; Zhang et al., 2020). Overall, these described NucS proteins have been shown to act on branched, mismatched and deaminated DNA, suggesting that this protein is a multifunctional enzyme involved in several DNA repair pathways. Biochemical characterizations of the NucS homolog from *T. kodakarensis* clearly showed that NucS specifically cleaves both strands of double-stranded DNA into 5'-protruding forms, with the mismatched base pair in the central position.

The mismatch endonuclease activity of *T. kodakarensis* can use both magnesium and manganese metal ions for catalysis, similar to DV-Nuc3, that can utilize both metal ions (Ishino et al., 2018). DV-Nuc3 appears to have specific nuclease activity on Ss DNA substrates, with the addition of magnesium and cuts at the site of the DNA lesion, while non-specific activity is observed when manganese is added. Magnesium is the most abundant divalent cation inside cells, while other metal ions, like manganese are found at lower concentrations. Magnesium is larger than manganese and has more ridged geometry preferences, while geometry preference of manganese can be more distorted. Because of the stringent coordination geometry and charge requirements of magnesium, nuclease activity observed on DNA substrates, with magnesium bound, is often highly selective and specific. While on the other hand, activity observed with manganese bound appears less specific towards substrates (Yang, 2011). This difference has been observed in several metalloenzymes, possibly attributed to manganese being a transitional element with less stringent coordinate requirement compared to magnesium (Bertini & Turano, 2007; Lee et al., 2015).

NucS from *P. furiosus* cleaves G/T, G/G, T/T, T/C and A/G mismatches, with a more preference for G/T, G/G and T/T, but has very little or no effect on C/C, A/C and A/A mismatches (Ishino et al., 2016). Many of the characterized NucS homologs are found in thermophilic organisms, that are active at high temperatures (Ishino et al., 2018), whereas DV-Nuc3 is from a low temperature environment, and accordingly is more active on DNA substrates at a lower temperature range. DV-Nuc3 shows nuclease activity from 5-40 ºC, with best activity seen at 35 ºC. DSF and CD thermal melts, indicated that DV-Nuc3 has a T_m of 45 ºC and it is likely that no nuclease activity would be observed around

45 ºC or over. DSF thermal melts also indicate that DV-Nuc3 is most stable at a moderate pH of 7.5, which has also been observed in some NucS homologs, such as NucS from *C. glutamicum*, which has a very narrow pH optimum around 6.4. Other homologs, such as NucS from *T. kodakarensis* can function in a broad pH range, from 6.0 to 11.0 (Ishino et al., 2018).

In some Archaea and Actinobacteria there is no evidence of genes that encode the canonical MMR proteins (Ishino et al., 2018), which play an important role in in MMR. It has been suggested that these NucS proteins, which can cleave mismatched and branched DNA, are involved in non-canonical MMR and NER pathways in organisms, that are missing proteins involved in these canonical repair pathways (Zhang et al., 2020). P. *methylaliphatogenes* comes from the lineage Acidobacteria and possess canonical mismatch repair systems. The genome of DV-Nuc3 is predicted to come from this lineage and is likely to also possess this canonical mismatch repair system. This suggests that, in contrast to other characterised NucS proteins, this is not the function of DV-Nuc3 proteins from Acidobacteria.

6 Chapter 6

Conclusion and future recommendations

5.5 Research motivation

The McMurdo Dry Valleys (DVs) of Antarctica are dominated by extreme dryness and cold temperatures, making them among one of the harshest environments on Earth (Raggio et al., 2016). There is minimal water availability, they are subject to high levels of UV light and they experience multiple cycles of freezing and thawing daily (Tamppari et al., 2012). These conditions are extremely damaging to DNA, and consequently organisms inhabiting the DVs must possess highly efficient DNA repair systems to survive. The conditions experienced here are comparable to those found on the surface of Mars (Salvatore & Levy, 2021).

While previous studies have investigated the taxonomic composition of microbial communities, from the McMurdo DVs (Wei et al., 2016) and compared the *in silico* annotations of these metagenomes to other extreme environments, there is a sparse number of studies that have experimentally investigated the enzyme functions of microbes from the DVs.

Most of what is known about bacterial DNA repair enzymes comes from mesophiles, occupying habitats of moderate temperatures. There is less known about bacteria inhabiting extreme environments, particularly psychrophiles, and this partly comes down to the fact that many are unculturable. By studying these enzymes, we can provide insight into the mechanisms that may enable resident microbes to survive these threats to their genomic integrity and provide explanations for the possible exitance of life on Mars and even other frozen planets. DNA repair proteins identified here have the potential to be used as molecular biological tools. For example, cold adapted DNA ligases can be used in experiments that require a low temperature and they can easily be heat inactivated when they are no longer required. Another important motivation for this research

was to identify and chacterise novel DNA repair enzymes from bacteria inhabiting the DVs. These novel DNA repair enzymes may play a role in alternative DNA repair pathways, and by characterising the activity of these enzymes in DNA repair, we might be able to elucidate new DNA repair pathways used by microorganisms from the DVs.

The research covered by this project provides detailed systematic exploration of DNA repair enzymes from DV metagenomes and one of the first attempts at recombinant protein production from these samples.

5.6 Summary of key findings and implications

Owing to the difficulty of cultivating these organisms in a laboratory environment, soil samples were collected from thirty sample sites within these DVs, and these metagenomes were sequenced to identify novel or highly divergent DNA-processing enzymes that enable effective DNA repair. Before I joined this project the protein coding sequences from these sequenced metagenomes were used to construct comprehensive sequence similarity networks (SSNs). The SSNs grouped several of the DV metagenome protein sequences in separate clusters from sequences already available in NCBI. The SSNs revealed genes implicated in specialized repair processes, encompassing novel nucleases and a diverse array of ATP-dependent DNA ligases that participate in DNA repair pathways during stationary-phase (Rzoska-Smith et al., 2023). In summary, several genes associated with repair functions were identified from bioinformatic analysis of the DV metagenomes. Several of these genes appear to be novel to this environment or have very few representatives in the current database.

From the list of ATP-dependent DNA ligases identified from bioinformatic search of the DV-metagenomes, three candidate DNA ligases were selected for biochemical and structural investigations. Based on structural and sequence analysis, DV-Lig2 and DV-Lig5 were identified to belong to the bacterial family of LigB type ATP-dependent DNA ligases. These ligases structurally comprise of three domains common to the LigB family of ligases: a DNA binding (DB) domain, an adenylation (AD) domain and an oligonucleotide

binding (OB) domain. The gene encoding DV-Lig2 belongs to a four gene cluster of DNA repair genes that have been discovered in other bacterial species. The Lhr helicase and MPE gene products from this cluster have been structurally and biochemically characterized in *P. putida* (Ejaz & Shuman, 2018). The ligase and nuclease components from this cluster have not been biochemically or structurally defined. The third candidate DNA ligase was revealed to be a ligase nuclease fusion protein, which has been designated DV-1-1-Lig-Nuc. The ligase domain (DV-1-1-Lig) of DV-1-1-Lig-Nuc also belongs to the LigB family of DNA ligases, sharing the same conserved sequence motifs and is made up of the same three sub domains. The nuclease domain (DV-1-1-Nuc) shares structural and sequence homology to nucleases of the metallo-β-lactamase (MBL) structural super family and reassembles nucleases of the β-CASP nucleic acid processing subfamily (Fernandez et al., 2011). DV-1-1-Nuc possess a characteristic α/β/β/α MBL core fold and contains a β-CASP domain inserted within the MBL domain. Sequence alignments against members from the β-CASP subfamily show that DV-1-1-Nuc contains conserved MBL and β-CASP motifs.

Along with these DNA ligases, a protein characterized as a NucS type nuclease was also discovered from bioinformatic analysis of the DV-metagenomes, designated DV-Nuc3. Protein sequence alignments of DV-Nuc3 against known NucS homologs indicates that the N-terminal sequence of DV-Nuc3 is extended N-terminally relative to the characterized NucS homologs. While the C-terminus has sequence homology to the RecB-like endonuclease domain of the structurally characterized NucS from *T. kodakarensis* (Zhang et al., 2020). Structurally, DV-Nuc3 contains four discrete folded domains, where domain two has pseudo-symmetry with domain four. Domains one, two and four share structural homology to domains of previously characterized NucS proteins. The overall structure of DV-Nuc3 suggests that this protein functions as a monomer, in contrast to previously characterised homologs which are structural and functional dimers.

All three ligases, DV-Lig2, DV-Lig5 and DV-1-1-Lig can utilize either magnesium or manganese metal ions, use ATP and ADP as nucleotide cofactors for ligation activity, and they all possess the ability to ligate nicked and A/C

mismatched DNA substrates. Only DV-Lig5 and DV-1-1-Lig can ligate cohesive DNA substrates, this is likely owing to their extended DB domains. DV-Lig5 was also particularly active in ligating non-canonical DNA substrates. DV-1-1-Lig is more stable at higher temperatures than DV-Lig2 and DV-Lig5, possessing ligation activity up to 80 ºC, while DV-Lig2 and DV-Lig5 were only active up to 55 ºC.

DV-1-1-Nuc can utilize magnesium, manganese, and zinc metal ions for degradation of DNA substrates. It can cleave both Ds and Ss DNA substrates, with a preference for activity in a 5' to 3' direction. DV-1-1-Nuc also shows specific activity on abasic and uracil mismatch DNA damages, making cuts at the location of the lesions. Nuclease activity is observed from 1 ºC up to 50 ºC.

DV-Nuc3, can utilize both magnesium and manganese metal ions, which is also observed with NucS from *T. kodakarensis* (Zhang et al., 2020). DV-Nuc3 appears to have specific nuclease activity with the addition of magnesium and cuts at the site of the DNA lesion. With the addition of manganese, the observed nuclease activity on DNA substrates is non-specific. DV-Nuc3 protein displays nuclease activity on an array of DNA damages, with a preference for abasic and uracil match DNA substrates. Optimum nuclease activity is observed at a lower temperature range from 5-40 ºC, compared to archaeal NucS homologs (Ishino et al., 2018; Zhang et al., 2020).

5.7 Project challenges and solutions

5.7.1 Recombinant protein expression

All target proteins were initially recombinantly expressed in *E. coli* (DE3) BL21 and plysS cells for production. However, not all the target proteins expressed well in these *E. coli* strains. The mesophilic organism *E. coli* is a suitable host for expression of a number of heterologous proteins at standard cultivation temperatures. The target proteins all come from microorganisms inhabiting the McMurdo DVs, and therefore expression of these proteins was performed at lower than standard cultivation temperatures for *E. coli*. Chaperonins

expressed by *E. coli*, lose protein folding activity at reduced temperatures, and can result in unfolded recombinant protein expression.

The ArcticExpress (DE3) *E. coli* expression strain co-express cold-adapted chaperonins Cpn10 and Cpn60, that show high protein refolding activities at temperatures 4-12 ºC. ArcticExpress (DE3) strains were trialed for expression of the target proteins, as research shows that the chaperonins from ArcticExpress can improve protein processing at lower temperatures and can increase the yield of active, soluble recombinant protein (Ferrer et al., 2003).

In addition to the use of ArcticExpress (DE3) *E. coli* as a new expression strain, Origami 2 (DE3) was also trialed as another potential expression strain. Origami 2 strains have mutations in glutathione reductase (gor) and thioredoxin reductase (trxB), facilitating proper disulfide bond formation. While it was uncertain at the time if any of the proteins required disulfide bond formation for proper folding, these strains were used as they have been known to improve soluble expression of recombinant proteins (Brüsehaber et al., 2010).

Several of the target proteins, when expressed in either ArcticExpress (DE3) or Origami 2 (DE3) expression strains, showed an improvement in solubility as well as overall expression levels. Therefore ArcticExpress (DE3) and Origami 2 (DE3) *E. coli* expression strains were used for the expression of target proteins that exhibited poor expression in *E. coli* (DE3) BL21 and plysS expression cells.

For each target protein optimum expression conditions were trailed by using different expression temperatures (15, 20, 25 and 30 ºC) as well as different concentrations of IPTG (200mM and 500mM) and finally different $O.D^{600}$ induction growth points (0.3, 0.5 and 0.8 $O.D^{600}$).

5.7.2 Expression and purification of the separate domains from DV-1-1-Ligase

Before I joined this project small scale expression trial (50 mls) were conducted on DV-1-1-Ligase protein, using pDEST17 (His-tagged) and

pHMGWA (His-tagged and MBP tagged) expression vectors, with the BL21 (DE3) pLysS *E. coli* expression strain. Results from the small scale expression trials revealed that there was some soluble expression of DV-1-1-Ligase. However, large scale (1 L) purifications of DV-1-1-Ligase did not result in enough soluble protein to proceed to characterisation experiments. DV-1-1-Ligase contains an N-terminal nuclease domain, connected to a C-terminal ligase domain, by a polypeptide linker. Instead of expressing this protein as a whole, the new plan was to recombinantly express and purify the domains separately, to determine if this would improve protein solubility and expression levels.

DV-1-1-Lig domain expressed really well in the pDEST17 plasmid, from Origami 2 (DE3) expression cells, with a high concentration of protein being produced after purification. On the other hand, DV-1-1-Nuclease domain, was not soluble in either pDEST17 or pHMGWA, and exhibited very poor expression during small scale expression trials. Three new constructs were designed for DV-1-1-Nuclease domain, where residues at the start of the sequence and or end of the sequence were removed. The N-terminal of the nuclease domain appeared to have additional residues that did not align with homologous protein sequences and was observed as a region of high disorder from the AlphaFold2 structural predictions. At the C-terminal end of the nuclease domain, was the beginning of the polypeptide linker and when the original nuclease construct was designed, a small region of this linker was left on the sequence. The three new nuclease constructs showed improved solubility and expression in the pHMGWA plasmid with either ArcticExpress (DE3) or Origami 2 (DE3) in small scale expression trials. From here the construct that was truncated at both the C and N terminus was used in large scale purifications, followed by biological characterisation experiments.

Following on from the success of generating truncations to the N-terminus of the nuclease domain, the following truncation was also applied to the full-length enzyme product DV-1-1-Ligase. This new construct of DV-1-1-Ligase showed improved expression now that the high disorder residues had been removed from the N-terminus. This new construct was tested in biological

activities assay and displayed both ligation and nuclease activity similar to that observed in the separate ligase and nuclease domains.

5.7.3 Structural characterisation of proteins

The ligase domain (DV-1-1-Lig) from DV-1-1-Ligase was able to crystallise using the hanging drop crystallization method, with Natrix conditions (**Section 2.6**). However, none of these crystals diffracted during X-ray crystallography. No crystals were obtained from any of the other target proteins. From here, an *in-silico* approach was used to structurally predict the fold of the candidate proteins. 3D protein structures were predicted using AlphaFold prediction software accessible through Google ColabFold-v2.3.1. For each predicted model, a predicted aligned error (PAE) plot and a predicted Local Distance Difference Test (pLDDT) plot was generated. Most of the proteins showed low prediction error, except for regions of high flexibility. These predicted models of the proteins were helpful to better understand how these proteins might fold up and whether they were structurally similar to their protein homologs. Using the AlphaFold predicted model of DV-1-1-Nuclease domain, I was able to identify both the MBL and β-CASP domains and show that the arrangement of these domains matched several other MBL β-CASP proteins, such as those belonging to the SNM1 family of nucleases (Baddock et al., 2021; Schmiester & Demuth, 2017).

5.8 Future directions

5.8.1 Structural determination

Attempts to obtain X-ray crystallography data for the candidate proteins were unsuccessful due to various issues, including challenges in producing soluble and purified proteins, as well as obtaining crystals of sufficient quality for diffraction. Instead, AlphaFold2 was used to predict structural models for all the candidate proteins. While the models predicted by AlphaFold2 were useful to predict how these proteins might fold up, they are not as reliable as structural data obtained from *in vivo* experiments. To support the predicted protein structures from AlphaFold2, there are alternative techniques that can be used to determine

the structural properties of these proteins, using the protein in its native state, without the requirements of crystallisation.

While X-ray crystallography is the most common method for protein structural determination, other techniques such as nuclear magnetic resonance (NMR) spectroscopy, small angle X-ray diffraction (SAX) and electron cryo-microscopy (CryoEM) have increased in popularity, for structural investigations into proteins that are difficult to crystallise.

Nuclear magnetic resonance (NMR) spectroscopy uses strong local magnetic fields to analyse the alignment of nuclei in an atom and the data collected can be used for the determination of three-dimensional protein structures at atomic resolution (Billeter et al., 2008; Wüthrich, 2001). NMR spectroscopy offers an advantage in that it does not require protein crystallisation; instead, it only necessitates a small volume of concentrated protein solution. However, there is an upper limit to the size of protein whose structure can be determined, with structure determination being particularly challenging for protein larger than 50 kDa (Gauto et al., 2019). Due to this upper size limit, this technique would only be useful for some of my smaller proteins, such as DV-Nuc3 and DV-1-1-Nuc, that are both smaller than 60 kDa. DV-1-1-Nuc would also require further optimisation, during purification, to remove any contaminating *E. coli* proteins, such as chaperones that often purify alongside DV-1-1-Nuc.

In the past, X-ray crystallography and NMR have stood as the primary techniques for high-resolution structural analysis of macromolecules. However, X-ray crystallography's reliance on sample crystallisation and NMR's upper molecular mass limitation of approximately 50 kDa imposes significant constraints. In recent years, CryoEM has emerged as an extremely powerful tool for achieving high-resolution structural analysis and has now established its position as a main technique for the structural analysis of macromolecules (Namba & Makino, 2022). CryoEM can achieve near atomic resolution or sometimes true atomic resolution in structural analysis of macromolecules from a very small amount of solution, with no requirement of a crystallisation and virtually no upper limit in the molecular mass of proteins. CryoEM is particularly

useful for complexes that are either too large or too heterogeneous to be investigated by conventional X-ray crystallography or nuclear magnetic resonance (NMR) (Jonic & Vénien-Bryan, 2009). Here, CryoEM would be ideal to validate the interactions between domains of DV-1-1-Lig-Nuc and analyze binding interactions of the complex with different DNA substrates, e.g. Ss DNA, flapped DNA, mismatches, etc.

Small angle X-ray scattering (SAXS) is a technique for the low-resolution structural characterisation of biological macromolecules in solution. The method of SAXS involves dissolved macromolecules that are exposed to a X-ray beam and the scattered intensity is recorded as a function of the scattering angle. It is a useful technique to assess the oligomeric state of proteins and protein complexes. While SAXS currently doesn't provide atomic resolution of protein structures it is useful for structure validation and quantitative analysis of flexible systems. This technique would be useful for several of my proteins, to validate the structural predictions obtained from AlphaFold2. However, as sample homogeneity is important to achieve reliable data, the proteins that co-purify with several *E. coli* chaperones would require further optimisation before using this technique.

5.8.2 New protein expression systems

Expression and purification of several proteins from the DV-metagenomes was extremely difficult, with a number of candidate proteins being terminated due to the difficulty in obtaining soluble protein. The candidate proteins were all recombinantly expressed in *E. coli* expression strains. The use of different *E. coli* strains, such as ArcticExpress (DE3) and Origami 2 (DE3), saw an improvement in overall soluble protein expression compared to those expressed in BL21 (DE3) and BL21 (DE3) pLysS. However, there were still continuous issues with protein precipitation, especially after the removal of the solubility maltose binding protein (MBP) tag. Biochemical characterisation experiments with DV-Lig2 and DV-1-1-Nuc were mostly performed with the MBP tag left on. While there was enough protein expression from all the candidate proteins to perform biological activity assays, several of the proteins could not be used in structural determination

experiments such as X-ray crystallography or NMR, due to insufficient protein production.

The use of *E. coli* to express high levels of heterologous proteins can often result in the production of incorrectly folded protein, which results in inactive proteins known as inclusion bodies. Obtaining active protein from inclusion bodies typically requires several *in vitro* re-folding steps and doesn't ensure that the resulting purified protein will be biologically active (Belval et al., 2015). For future work with these DV-proteins, it would be useful to investigate the use of an alternative expression system that has the potential to overcome these problems. For example, the potential use of yeast or mammalian expression systems for producing larger proteins. Large proteins are usually expressed in a eukaryotic expression system while smaller proteins are expressed in prokaryotic systems. *E. coli* was initially chosen as the expression host for the candidate proteins, as they are all bacterial proteins and protein expression in *E. coli* is inexpensive, easy, and quick. However, expression of large proteins in *E. coli* often results in the formation of insoluble target protein, due to misfolding (Demain & Vaishnav, 2009).

Yeast expression systems are often used to produce recombinant proteins that are not expressed well in *E. coli* due to problems with proteins not folding correctly or the need for glycosylation. The advantages of using yeast as expression hosts includes post translational modifications, fast growth, simple genetic manipulation, scalable fermentation and high biomass concentrations (Baghban et al., 2019).

5.8.3 Mutant design for DV-1-1-Nuclease domain

Proteins belonging to the family of MBL β-CASP nucleases contain conserved residue motifs, motifs 1-4 are typical for the whole MBL superfamily, while motifs A-C are only found in the β-CASP containing family (Fernandez et al., 2011). Motifs 1-4 are responsible for metal ion coordination in both RNA and DNA processing MBL enzymes (de Villartay et al., 2009).

DV-1-1-Nuc domain protein belongs to the family of MBL β-CASP nucleases and contains the same conserved motifs, as previously described (**Section 4.2.2**). To validate the specific nuclease activity of DV-1-1-Nuc on DNA substrates, a mutant lacking nuclease activity must serve as a control. This control is essential to confirm that any observed activity can be attributed solely to DV-1-1-Nuc, and not to contaminating *E. coli* nucleases.

A mutant was designed for DV-1-1-Nuc that targeted two out of the four conserved residues in motif II (D36A-H37A). The activity of the mutant was tested on DNA substrates, alongside the wild-type DV-1-1-Nuc, using biological activity assays. The mutant was still active on DNA substrates, but the activity was reduced when compared to that of the wild-type DV-1-1-Nuc.

The current design of DV-1-1-Nuc mutant involves mutation of key residues important in metal ion coordination in the active site. Structural investigations of the predicted model for DV-1-1-Nuc shows that there is an additional metal binding site, involving several key residues. Future mutational design should focus on these additional key residues involved in metal ion coordination at this second binding site. Additional targets for mutation could involve conserved motif residues found within the MBL and β-CASP domains, especially those important in DNA binding.

References

Abraham, W. P., Raghunandanan, S., Gopinath, V., Suryaletha, K., & Thomas, S. (2020). Deciphering the cold adaptive mechanisms in Pseudomonas psychrophila MTCC12324 isolated from the Arctic at 79 N. *Current Microbiology*, *77*, 2345-2355.

Adul Rahman, R., Jongsareejit, B., Fujiwara, S., & Imanaka, T. (1997). Characterization of recombinant glutamine synthetase from the hyperthermophilic archaeon Pyrococcus sp. strain KOD1. *Applied and environmental microbiology*, *63*(6), 2472-2476.

Aislabie, J., Jordan, S., & Barker, G. (2008). Relation between soil classification and bacterial diversity in soils of the Ross Sea region, Antarctica. *Geoderma*, *144*(1-2), 9-20.

Amare, B., Mo, A., Khan, N., Sowa, D. J., Warner, M. M., Tetenych, A., & Andres, S. N. (2021). LigD: a structural guide to the multi-tool of bacterial non-homologous end joining. *Frontiers in Molecular Biosciences*, *8*, 787709.

Aravind, L., & Koonin, E. V. (2001). Prokaryotic homologs of the eukaryotic DNA-end-binding protein Ku, novel domains in the Ku protein and prediction of a prokaryotic double-strand break repair system. *Genome research*, *11*(8), 1365-1374.

Aravind, L., Makarova, K. S., & Koonin, E. V. (2000). SURVEY AND SUMMARY: holliday junction resolvases and related nucleases: identification of new families, phyletic distribution and evolutionary trajectories. *Nucleic acids research*, *28*(18), 3417-3432.

Atkinson, H. J., Morris, J. H., Ferrin, T. E., & Babbitt, P. C. (2009). Using sequence similarity networks for visualization of relationships across diverse protein superfamilies. *PloS one*, *4*(2), e4345.

Au, N., Kuester-Schoeck, E., Mandava, V., Bothwell, L. E., Canny, S. P., Chachu, K., Colavito, S. A., Fuller, S. N., Groban, E. S., & Hensley, L. A. (2005). Genetic composition of the Bacillus subtilis SOS system. *Journal of bacteriology*, *187*(22), 7655-7666.

Aydin, Ö. (2014). Chromatin remodeling in the UV-induced DNA damage response.

Baddock, H. T., Newman, J. A., Yosaatmadja, Y., Bielinski, M., Schofield, C. J., Gileadi, O., & McHugh, P. J. (2021). A phosphate binding pocket is a key determinant of exo-versus endo-nucleolytic activity in the SNM1 nuclease family. *Nucleic acids research*, *49*(16), 9294-9309.

Baddock, H. T., Yosaatmadja, Y., Newman, J. A., Schofield, C. J., Gileadi, O., & McHugh, P. J. (2020). The SNM1A DNA repair nuclease. *DNA repair*, *95*, 102941.

Baghban, R., Farajnia, S., Rajabibazl, M., Ghasemi, Y., Mafi, A., Hoseinpoor, R., Rahbarnia, L., & Aria, M. (2019). Yeast expression systems: overview and recent advances. *Molecular biotechnology*, *61*, 365-384.

Bagwell, C. E., Bhat, S., Hawkins, G. M., Smith, B. W., Biswas, T., Hoover, T. R., Saunders, E., Han, C. S., Tsodikov, O. V., & Shimkets, L. J. (2008). Survival in nuclear waste, extreme resistance, and potential applications gleaned from the genome sequence of Kineococcus radiotolerans SRS30216. *PloS one*, *3*(12), e3878.

Baharoglu, Z., & Mazel, D. (2014). SOS, the formidable strategy of bacteria against aggressions. *FEMS microbiology reviews*, *38*(6), 1126-1145.

Bargagli, R., Skotnicki, M., Marri, L., Pepi, M., Mackenzie, A., & Agnorelli, C. (2004). New record of moss and thermophilic bacteria species and physico-chemical properties of geothermal soils on the northwest slope of Mt. Melbourne (Antarctica). *Polar Biology*, *27*, 423-431.

Bauer, R. J., Zhelkovsky, A., Bilotti, K., Crowell, L. E., Evans Jr, T. C., McReynolds, L. A., & Lohman, G. J. (2017). Comparative analysis of the end-joining activity of several DNA ligases. *PloS one*, *12*(12), e0190062.

Becherel, O. J., & Fuchs, R. P. (2001). Mechanism of DNA polymerase II-mediated frameshift mutagenesis. *Proceedings of the National Academy of Sciences*, *98*(15), 8566-8571.

Belval, L., Marquette, A., Mestre, P., Piron, M.-C., Demangeat, G., Merdinoglu, D., & Chich, J.-F. (2015). A fast and simple method to eliminate Cpn60 from functional recombinant proteins produced by E. coli Arctic Express. *Protein expression and purification*, *109*, 29-34.

Benner, S. A., Karalkar, N. B., Hoshika, S., Laos, R., Shaw, R. W., Matsuura, M., Fajardo, D., & Moussatche, P. (2016). Alternative Watson–Crick synthetic genetic systems. *Cold Spring Harbor perspectives in biology*, *8*(11), a023770.

Berg, J. M., Tymoczko, J. L., & Stryer, L. (2012). Biochemistry/Jeremy M. Berg, John L. Tymoczko, Lubert Stryer; with Gregory J. Gatto, Jr. In: New York: WH Freeman.

Berg, K., Leiros, I., & Williamson, A. (2019). Temperature adaptation of DNA ligases from psychrophilic organisms. *Extremophiles*, *23*, 305-317.

Bertini, I., & Turano, P. (2007). Metal ions and Proteins: Binding, stability, and folding. *Biological Inorganic Chemistry Structure and Reactivity*, 31-41.

Betz, K., Malyshev, D. A., Lavergne, T., Welte, W., Diederichs, K., Dwyer, T. J., Ordoukhanian, P., Romesberg, F. E., & Marx, A. (2012). KlenTaq polymerase replicates unnatural base pairs by inducing a Watson-Crick geometry. *Nature chemical biology*, *8*(7), 612-614.

Billeter, M., Wagner, G., & Wüthrich, K. (2008). Solution NMR structure determination of proteins revisited. *Journal of biomolecular NMR*, *42*, 155-158.

Billi, D., Friedmann, E. I., Hofer, K. G., Caiola, M. G., & Ocampo-Friedmann, R. (2000). Ionizing-radiation resistance in the desiccation-tolerant cyanobacterium Chroococcidiopsis. *Applied and environmental microbiology*, *66*(4), 1489-1492.

Bilotti, K., Potapov, V., Pryor, J. M., Duckworth, A. T., Keck, J. L., & Lohman, G. J. (2022). Mismatch discrimination and sequence bias during end-joining by DNA ligases. *Nucleic acids research*, *50*(8), 4647-4658.

Borges, N., Ramos, A., Raven, N. D., Sharp, R. J., & Santos, H. (2002). Comparative study of the thermostabilizing properties of mannosylglycerate and other compatible solutes on model enzymes. *Extremophiles*, *6*(3), 209-216.

Bottos, E. M., Laughlin, D. C., Herbold, C. W., Lee, C. K., McDonald, I. R., & Cary, S. C. (2020). Abiotic factors influence patterns of bacterial diversity and community composition in the Dry Valleys of Antarctica. *FEMS microbiology ecology*, *96*(5), fiaa042.

Bradley, N. P., Washburn, L. A., Christov, P. P., Watanabe, C. M., & Eichman, B. F. (2020). Escherichia coli YcaQ is a DNA glycosylase that unhooks DNA interstrand crosslinks. *Nucleic acids research*, *48*(13), 7005-7017.

Bridges, B. A. (2005). Error-prone DNA repair and translesion DNA synthesis: II: The inducible SOS hypothesis. *DNA repair*, *4*(6), 725-739.

Brüggemann, H., & Chen, C. (2006). Comparative genomics of Thermus thermophilus: plasticity of the megaplasmid and its contribution to a thermophilic lifestyle. *Journal of biotechnology*, *124*(4), 654-661.

Brüsehaber, E., Schwiebs, A., Schmidt, M., Böttcher, D., & Bornscheuer, U. T. (2010). Production of pig liver esterase in batch fermentation of E. coli Origami. *Applied microbiology and biotechnology*, *86*, 1337-1344.

Burby, P. E., & Simmons, L. A. (2019). A bacterial DNA repair pathway specific to a natural antibiotic. *Molecular microbiology*, *111*(2), 338-353.

Cadet, J., Grand, A., & Douki, T. (2015). Solar UV radiation-induced DNA bipyrimidine photoproducts: formation and mechanistic insights. *Photoinduced phenomena in nucleic acids II: DNA fragments and phenomenological aspects*, 249-275.

Callebaut, I., Moshous, D., Mornon, J. P., & de Villartay, J. P. (2002). Metallo‐β‐lactamase fold within nucleic acids processing enzymes: the β‐CASP family. *Nucleic acids research*, *30*(16), 3592-3601.

Cannan, W. J., & Pederson, D. S. (2016). Mechanisms and consequences of double‐strand DNA break formation in chromatin. *Journal of cellular physiology*, *231*(1), 3-14.

Caron, P. R., Kushner, S. R., & Grossman, L. (1985). Involvement of helicase II (uvrD gene product) and DNA polymerase I in excision mediated by the uvrABC protein complex. *Proceedings of the National Academy of Sciences*, *82*(15), 4925-4929.

Cary, S. C., McDonald, I. R., Barrett, J. E., & Cowan, D. A. (2010). On the rocks: the microbiology of Antarctic Dry Valley soils. *Nature Reviews Microbiology*, *8*(2), 129.

Casanueva, A., Tuffin, M., Cary, C., & Cowan, D. A. (2010). Molecular adaptations to psychrophily: the impact of 'omic'technologies. *Trends in microbiology*, *18*(8), 374-381.

Castañeda-García, A., Prieto, A., Rodriguez-Beltran, J., Alonso, N., Cantillon, D., Costas, C., Pérez-Lago, L., Zegeye, E., Herranz, M., & Plociński, P. (2017). A non-canonical mismatch repair pathway in prokaryotes. *Nature communications*, *8*(1), 14246.

Cebrián-Sastre, E., Martín-Blecua, I., Gullón, S., Blázquez, J., & Castañeda-García, A. (2021). Control of genome stability by EndoMS/NucS-mediated non-canonical mismatch repair. *Cells*, *10*(6), 1314.

Chang, H. H., Watanabe, G., & Lieber, M. R. (2015). Unifying the DNA end-processing roles of the artemis nuclease: Ku-dependent artemis resection at blunt DNA ends. *Journal of Biological Chemistry*, *290*(40), 24036-24050.

Chattopadhyay, M. (2006). Mechanism of bacterial adaptation to low temperature. *Journal of biosciences*, *31*(1), 157-165.

Chauleau, M., & Shuman, S. (2016). Kinetic mechanism and fidelity of nick sealing by Escherichia coli NAD+-dependent DNA ligase (LigA). *Nucleic acids research*, *44*(5), 2298-2309.

Chen, B., Sysoeva, T. A., Chowdhury, S., Guo, L., & Nixon, B. T. (2009). ADPase activity of recombinantly expressed thermotolerant ATPases may be caused by copurification of adenylate kinase of Escherichia coli. *The FEBS journal*, *276*(3), 807-815.

Chen, I.-M. A., Chu, K., Palaniappan, K., Ratner, A., Huang, J., Huntemann, M., Hajek, P., Ritter, S. J., Webb, C., & Wu, D. (2023). The IMG/M data management and analysis system v. 7: content updates and new features. *Nucleic acids research*, *51*(D1), D723-D732.

Chen, R., Li, C., Pei, X., Wang, Q., Yin, X., & Xie, T. (2014). Isolation an aldehyde dehydrogenase gene from metagenomics based on semi-nest touch-down PCR. *Indian journal of microbiology*, *54*, 74-79.

Cheng, C., & Shuman, S. (1997). Characterization of an ATP-dependent DNA ligase encoded by Haemophilus influenzae. *Nucleic acids research*, *25*(7), 1369-1374.

Cherepanov, A. V., & De Vries, S. (2003). Kinetics and thermodynamics of nick sealing by T4 DNA ligase. *European Journal of Biochemistry*, *270*(21), 4315-4325.

Chuzel, L., Ganatra, M. B., Rapp, E., Henrissat, B., & Taron, C. H. (2018). Functional metagenomics identifies an exosialidase with an inverting catalytic mechanism that defines a new glycoside hydrolase family (GH156). *Journal of Biological Chemistry*, *293*(47), 18138-18150.

Cole, R. S. (1973). Repair of DNA containing interstrand crosslinks in Escherichia coli: sequential excision and recombination. *Proceedings of the National Academy of Sciences*, *70*(4), 1064-1068.

Coleine, C., Pombubpa, N., Zucconi, L., Onofri, S., Stajich, J. E., & Selbmann, L. (2020). Endolithic fungal species markers for harshest conditions in the McMurdo Dry valleys, Antarctica. *Life*, *10*(2), 13.

Cooper, G. M., & Hausman, R. (2000). A molecular approach. *The Cell. 2nd ed. Sunderland, MA: Sinauer Associates*.

Cox, M. M., & Battista, J. R. (2005). Deinococcus radiodurans—the consummate survivor. *Nature Reviews Microbiology*, *3*(11), 882-892.

Creze, C., Lestini, R., Kühn, J., Ligabue, A., Becker, H. F., Czjzek, M., Flament, D., & Myllykallio, H. (2011). Structure and function of a novel endonuclease acting on branched DNA substrates. In: Portland Press Ltd.

Dalhus, B., Laerdahl, J. K., Backe, P. H., & Bjørås, M. (2009). DNA base repair–recognition and initiation of catalysis. *FEMS microbiology reviews*, *33*(6), 1044-1078.

Dallo, S. F., & Weitao, T. (2010). Bacteria under SOS evolve anticancer phenotypes. *Infectious Agents and Cancer*, *5*(1), 1-5.

Dalmaso, G. Z. L., Ferreira, D., & Vermelho, A. B. (2015). Marine extremophiles: a source of hydrolases for biotechnological applications. *Marine drugs*, *13*(4), 1925-1965.

Daly, M. J. (2012). Death by protein damage in irradiated cells. *DNA repair*, *11*(1), 12-21.

De Ioannes, P., Malu, S., Cortes, P., & Aggarwal, A. K. (2012). Structural basis of DNA ligase IV-Artemis interaction in nonhomologous end-joining. *Cell reports*, *2*(6), 1505-1512.

De los Ríos, A., Cary, C., & Cowan, D. (2014). The spatial structures of hypolithic communities in the Dry Valleys of East Antarctica. *Polar Biology*, *37*(12), 1823-1833.

De Los Ríos, A., Wierzchos, J., & Ascaso, C. (2014). The lithic microbial ecosystems of Antarctica's McMurdo Dry Valleys. *Antarctic Science*, *26*(5), 459-477.

de Villartay, J.-P., Shimazaki, N., Charbonnier, J.-B., Fischer, A., Mornon, J.-P., Lieber, M. R., & Callebaut, I. (2009). A histidine in the β-CASP domain of Artemis is critical for its full in vitro and in vivo functions. *DNA repair*, *8*(2), 202-208.

Decottignies, A. (2013). Alternative end-joining mechanisms: a historical perspective. *Frontiers in genetics*, *4*, 48.

Della, M., Palmbos, P. L., Tseng, H.-M., Tonkin, L. M., Daley, J. M., Topper, L. M., Pitcher, R. S., Tomkinson, A. E., Wilson, T. E., & Doherty, A. J. (2004). Mycobacterial Ku and ligase proteins constitute a two-component NHEJ repair machine. *Science*, *306*(5696), 683-685.

Demain, A. L., & Vaishnav, P. (2009). Production of recombinant proteins by microbes and higher organisms. *Biotechnology advances*, *27*(3), 297-306.

Đermić, D. (2015). Double-strand break repair mechanisms in Escherichia coli: recent insights.

Doetsch, P. W., & Cunningham, R. P. (1990). The enzymology of apurinic/apyrimidinic endonucleases. *Mutation Research/DNA Repair*, *236*(2-3), 173-201.

Doherty, A. J., Ashford, S. R., Subramanya, H. S., & Wigley, D. B. (1996). Bacteriophage T7 DNA Ligase: Overexpression, purification, crystallization, and characterization (∗). *Journal of Biological Chemistry*, *271*(19), 11083-11089.

Doherty, A. J., Jackson, S. P., & Weller, G. R. (2001). Identification of bacterial homologues of the Ku DNA repair proteins. *FEBS letters*, *500*(3), 186-188.

Doherty, A. J., & Suh, S. W. (2000). Structural and mechanistic conservation in DNA ligases. *Nucleic acids research*, *28*(21), 4051-4058.

Dominski, Z. (2007). Nucleases of the metallo-β-lactamase family and their role in DNA and RNA metabolism. *Critical reviews in biochemistry and molecular biology*, *42*(2), 67-93.

Donald, J. E., Kulp, D. W., & DeGrado, W. F. (2011). Salt bridges: Geometrically specific, designable interactions. *Proteins: Structure, Function, and Bioinformatics*, *79*(3), 898-915.

Doran, P. T., McKay, C. P., Clow, G. D., Dana, G. L., Fountain, A. G., Nylen, T., & Lyons, W. B. (2002). Valley floor climate observations from the McMurdo Dry Valleys, Antarctica, 1986–2000. *Journal of Geophysical Research: Atmospheres*, *107*(D24), ACL 13-11-ACL 13-12.

Douki, T., von Koschembahr, A., & Cadet, J. (2017). Insight in DNA repair of UV ‐ induced pyrimidine dimers by chromatographic methods. *Photochemistry and photobiology*, *93*(1), 207-215.

Duffy, K., Arangundy-Franklin, S., & Holliger, P. (2020). Modified nucleic acids: replication, evolution, and next-generation therapeutics. *Bmc Biology*, *18*(1), 1-14.

Dupuy, P., Howlader, M., & Glickman, M. S. (2020). A multilayered repair system protects the mycobacterial chromosome from endogenous and antibiotic-induced oxidative damage. *Proceedings of the National Academy of Sciences*, *117*(32), 19517-19527.

Dziewit, L., & Bartosik, D. (2014). Plasmids of psychrophilic and psychrotolerant bacteria and their role in adaptation to cold environments. *Frontiers in microbiology*, *5*, 596.

Ejaz, A., Goldgur, Y., & Shuman, S. (2019). Activity and structure of Pseudomonas putida MPE, a manganese-dependent single-strand DNA endonuclease encoded in a nucleic acid repair gene cluster. *Journal of Biological Chemistry*, *294*(19), 7931-7941.

Ejaz, A., & Shuman, S. (2018). Characterization of Lhr-Core DNA helicase and manganese-dependent DNA nuclease components of a bacterial gene cluster encoding nucleic acid repair enzymes. *Journal of Biological Chemistry*, *293*(45), 17491-17504.

Enderle, J., Dorn, A., & Puchta, H. (2019). DNA-and DNA-protein-crosslink repair in plants. *International journal of molecular sciences*, *20*(17), 4304.

Ericsson, U. B., Hallberg, B. M., DeTitta, G. T., Dekker, N., & Nordlund, P. (2006). Thermofluor-based high-throughput stability optimization of proteins for structural studies. *Analytical biochemistry*, *357*(2), 289-298.

Erill, I., Campoy, S., & Barbé, J. (2007). Aeons of distress: an evolutionary perspective on the bacterial SOS response. *FEMS microbiology reviews*, *31*(6), 637-656.

Feller, G. (2013). Psychrophilic enzymes: from folding to function and biotechnology. *Scientifica*, *2013*.

Fernandez, F. J., Lopez-Estepa, M., & Vega, M. C. (2011). Nucleases of metallo-beta-lactamase and protein phosphatase families in DNA repair. In *DNA Repair: On the Pathways to Fixing DNA Damage and Errors*. Intech.

Ferrer, M., Chernikova, T. N., Yakimov, M. M., Golyshin, P. N., & Timmis, K. N. (2003). Chaperonins govern growth of Escherichia coli at low temperatures. *Nature biotechnology*, *21*(11), 1266-1267.

Ferrer, M., Martínez-Abarca, F., & Golyshin, P. N. (2005). Mining genomes and 'metagenomes' for novel catalysts. *Current Opinion in Biotechnology*, *16*(6), 588-593.

Fleck, O., & Nielsen, O. (2004). DNA repair. *Journal of Cell Science*, *117*(4), 515-517. https://doi.org/10.1242/jcs.00952

Flores, M. R., Ordoñez, O. F., Maldonado, M. J., & Farías, M. E. (2009). Isolation of UV-B resistant bacteria from two high altitude Andean lakes (4,400 m) with saline and non saline conditions. *The Journal of General and Applied Microbiology*, *55*(6), 447-458.

Forterre, P. (2002). A hot story from comparative genomics: reverse gyrase is the only hyperthermophile-specific protein. *Trends in Genetics*, *18*(5), 236-237.

Fountain, A. G., Nylen, T. H., Monaghan, A., Basagic, H. J., & Bromwich, D. (2010). Snow in the McMurdo dry valleys, Antarctica. *International Journal of Climatology: A Journal of the Royal Meteorological Society*, *30*(5), 633-642.

Friedberg, E. C., Walker, G. C., Siede, W., & Wood, R. D. (2005). *DNA repair and mutagenesis*. American Society for Microbiology Press.

Friedmann, E. I. (1982). Endolithic microorganisms in the Antarctic cold desert. *Science*, *215*(4536), 1045-1053.

Friedmann, E. I., & Thistle, A. B. (1993). Antarctic microbiology.

Fuchs, R. P., & Fujii, S. (2013). Translesion DNA synthesis and mutagenesis in prokaryotes. *Cold Spring Harbor Perspectives in Biology*, *5*(12), a012682.

Fuentes‐León, F., Peres de Oliveira, A., Quintero‐Ruiz, N., Munford, V., Satoru Kajitani, G., Coimbra Brum, A., Schuch, A. P., Colepicolo, P., Sá

nchez‐Lamar, A., & Menck, C. F. M. (2020). DNA Damage Induced by Late Spring Sunlight in Antarctica. *Photochemistry and Photobiology*.

Fujii, S., & Fuchs, R. P. (2004). Defining the position of the switches between replicative and bypass DNA polymerases. *The EMBO journal*, *23*(21), 4342-4352.

Fujiwara, S., Lee, S., Haruki, M., Kanaya, S., Takagi, M., & Imanaka, T. (1996). Unusual enzyme characteristics of aspartyl-tRNA synthetase from hyperthermophilic archaeon Pyrococcus sp. KOD1. *FEBS letters*, *394*(1), 66-70.

Gauto, D. F., Estrozi, L. F., Schwieters, C. D., Effantin, G., Macek, P., Sounier, R., Sivertsen, A. C., Schmidt, E., Kerfah, R., & Mas, G. (2019). Integrated NMR and cryo-EM atomic-resolution structure determination of a half-megadalton enzyme complex. *Nature communications*, *10*(1), 2697.

Gerlt, J. A., Bouvier, J. T., Davidson, D. B., Imker, H. J., Sadkhin, B., Slater, D. R., & Whalen, K. L. (2015). Enzyme function initiative-enzyme similarity tool (EFI-EST): a web tool for generating protein sequence similarity networks. *Biochimica Et Biophysica Acta (BBA)-Proteins and Proteomics*, *1854*(8), 1019-1037.

Ghosh, S., Ejaz, A., Repeta, L., & Shuman, S. (2021). Pseudomonas putida MPE, a manganese-dependent endonuclease of the binuclear metallophosphoesterase superfamily, incises single-strand DNA in two orientations to yield a mixture of 3′ -PO4 and 3′ -OH termini. *Nucleic acids research*, *49*(2), 1023-1032.

Gong, C., Martins, A., Bongiorno, P., Glickman, M., & Shuman, S. (2004). Biochemical and genetic analysis of the four DNA ligases of mycobacteria. *Journal of Biological Chemistry*, *279*(20), 20594-20606.

Gonzalez-Hunt, C. P., Wadhwa, M., & Sanders, L. H. (2018). DNA damage by oxidative stress: Measurement strategies for two genomes. *Current Opinion in Toxicology*, *7*, 87-94.

Gottesman, M. M., Hicks, M. L., & Gellert, M. (1973). Genetics and function of DNA ligase in Escherichia coli. *Journal of molecular biology*, *77*(4), 531-547.

GraphPadSoftware. *GraphPad Prism version 8 for Windows*. www.graphpad.com

Greenfield, S. R., Tighe, S. W., Bai, Y., Goerlitz, D. S., Von Turkovich, M., Taatjes, D. J., Dragon, J. A., & Johnson, S. S. (2020). Life and its traces in Antarctica's McMurdo Dry Valley paleolakes: a survey of preservation. *Micron*, *131*, 102818.

Grogan, D. W. (2000). The question of DNA repair in hyperthermophilic archaea. *Trends in microbiology*, *8*(4), 180-185.

Grossman, L., & Yeung, A. T. (1990). The UvrABC endonuclease system of Escherichia coli—a view from Baltimore. *Mutation Research/DNA Repair*, *236*(2-3), 213-221.

Gupta, G., Srivastava, S., Khare, S., & Prakash, V. (2014). Extremophiles: an overview of microorganism from extreme environment. *International Journal of Agriculture, Environment and Biotechnology*, *7*(2), 371-380.

Haldenby, S., White, M. F., & Allers, T. (2009). RecA family proteins in archaea: RadA and its cousins. In: Portland Press Ltd.

Handelsman, J. (2004). Metagenomics: application of genomics to uncultured microorganisms. *Microbiol. Mol. Biol. Rev.*, *68*(4), 669-685.

Handelsman, J. (2005). Metagenomics: application of genomics to uncultured microorganisms. *Microbiol. Mol. Biol. Rev.*, *69*(1), 195-195.

Hegde, M. L., Izumi, T., & Mitra, S. (2012). Oxidized base damage and single-strand break repair in mammalian genomes: role of disordered regions and posttranslational modifications in early enzymes. *Progress in molecular biology and translational science*, *110*, 123-153.

Ho, C. K., Van Etten, J. L., & Shuman, S. (1997). Characterization of an ATP-dependent DNA ligase encoded by Chlorella virus PBCV-1. *Journal of virology*, *71*(3), 1931-1937.

Hofreiter, M., Jaenicke, V., Serre, D., Haeseler, A. v., & Pääbo, S. (2001). DNA sequences from multiple amplifications reveal artifacts induced by cytosine deamination in ancient DNA. *Nucleic acids research*, *29*(23), 4793-4799.

Horowitz, N., Cameron, R. E., & Hubbard, J. S. (1972). Microbiology of the dry valleys of Antarctica. *Science*, *176*(4032), 242-245.

Hossain, M. A., Lin, Y., & Yan, S. (2018). Single-strand break end resection in genome integrity: mechanism and regulation by APE2. *International journal of molecular sciences*, *19*(8), 2389.

Hoyoux, A., Jennes, I., Dubois, P., Genicot, S., Dubail, F., François, J.-M., Baise, E., Feller, G., & Gerday, C. (2001). Cold-adapted β-galactosidase from the Antarctic psychrophile Pseudoalteromonas haloplanktis. *Applied and environmental microbiology*, *67*(4), 1529-1535.

Huang, J., Liu, X., Sun, Y., Li, Z., Lin, M.-H., Hamilton, K., Mandel, C. R., Sandmeir, F., Conti, E., & Oyala, P. H. (2023). An examination of the metal ion content in the active sites of human endonucleases CPSF73 and INTS11. *Journal of Biological Chemistry*, *299*(4).

Huang, Y., & Li, L. (2013). DNA crosslinking damage and cancer-a tale of friend and foe. *Translational cancer research*, *2*(3), 144.

Humann, J. L., & Kahn, M. L. (2015). Genes involved in desiccation resistance of rhizobia and other bacteria. *Biological nitrogen fixation*, 397-404.

Husain, I., Van Houten, B., Thomas, D. C., Abdel-Monem, M., & Sancar, A. (1985). Effect of DNA polymerase I and DNA helicase II on the turnover rate of UvrABC excision nuclease. *Proceedings of the National Academy of Sciences*, *82*(20), 6774-6778.

Ionescu, D., Oren, A., Hindiyeh, M. Y., & Malkawi, H. I. (2007). The thermophilic cyanobacteria of the Zerka Ma'in thermal springs in Jordan. *Algae and cyanobacteria in extreme environments*, 411-424.

Ishino, S., Nishi, Y., Oda, S., Uemori, T., Sagara, T., Takatsu, N., Yamagami, T., Shirai, T., & Ishino, Y. (2016). Identification of a mismatch-specific endonuclease in hyperthermophilic Archaea. *Nucleic acids research*, *44*(7), 2977-2986.

Ishino, S., Skouloubris, S., Kudo, H., l'Hermitte-Stead, C., Es-Sadik, A., Lambry, J.-C., Ishino, Y., & Myllykallio, H. (2018). Activation of the mismatch-specific endonuclease EndoMS/NucS by the replication clamp is required for high fidelity DNA replication. *Nucleic acids research*, *46*(12), 6206-6217.

Jin, M., Gai, Y., Guo, X., Hou, Y., & Zeng, R. (2019). Properties and applications of extremozymes from deep-sea extremophilic microorganisms: a mini review. *Marine drugs*, *17*(12), 656.

Jindal, S. (2020). Microbes in soil and their metagenomics. In *Microbial Diversity, Interventions and Scope* (pp. 85-96). Springer.

John Jumper, R. E., Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek,

Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A. A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, …Demis Hassabis. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, *596*, 583–589. https://doi.org/https://doi.org/10.1038/s41586-021-03819-2

Jones, P., Binns, D., Chang, H.-Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A., & Nuka, G. (2014). InterProScan 5: genome-scale protein function classification. *Bioinformatics*, *30*(9), 1236-1240.

Jonic, S., & Vénien-Bryan, C. (2009). Protein structure determination by electron cryo-microscopy. *Current opinion in pharmacology*, *9*(5), 636-642.

José, R., Goebel, B. M., Friedmann, E. I., & Pace, N. R. (2003). Microbial diversity of cryptoendolithic communities from the McMurdo Dry Valleys, Antarctica. *Applied and environmental microbiology*, *69*(7), 3858-3867.

Jun, S. H., Kim, T. G., & Ban, C. (2006). DNA mismatch repair system. *The FEBS journal*, *273*(8), 1609-1619.

Kaczmarek, F. S., Zaniewski, R. P., Gootz, T. D., Danley, D. E., Mansour, M. N., Griffor, M., Kamath, A. V., Cronan, M., Mueller, J., & Sun, D. (2001). Cloning and functional characterization of an NAD+-dependent DNA ligase from Staphylococcus aureus. *Journal of bacteriology*, *183*(10), 3016-3024.

Kambach, C., Walke, S., Young, R., Avis, J. M., De La Fortelle, E., Raker, V. A., Lührmann, R., Li, J., & Nagai, K. (1999). Crystal structures of two Sm protein complexes and their implications for the assembly of the spliceosomal snRNPs. *Cell*, *96*(3), 375-387.

Kawahara, H. (2002). The structures and functions of ice crystal-controlling proteins from bacteria. *Journal of bioscience and bioengineering*, *94*(6), 492-496.

Kelley, L. A., Mezulis, S., Yates, C. M., Wass, M. N., & Sternberg, M. J. (2015). The Phyre2 web portal for protein modeling, prediction and analysis. *Nature protocols*, *10*(6), 845-858.

Kemp, M. G., & Sancar, A. (2012). DNA excision repair: where do all the dimers go? *Cell Cycle*, *11*(16), 2997-3002.

Khan, B. (2014). Biotech Khan. https://biotechkhan.wordpress.com/2014/10/14/dna-damage

Khan, N., Tuffin, M., Stafford, W., Cary, C., Lacap, D. C., Pointing, S. B., & Cowan, D. (2011). Hypolithic microbial communities of quartz rocks from Miers Valley, McMurdo Dry Valleys, Antarctica. *Polar Biology*, *34*(11), 1657.

Kim, J.-H., Lee, K.-K., Sun, Y., Seo, G.-J., Cho, S. S., Kwon, S. H., & Kwon, S.-T. (2013). Broad nucleotide cofactor specificity of DNA ligase from the hyperthermophilic crenarchaeon Hyperthermus butylicus and its evolutionary significance. *Extremophiles*, *17*, 515-522.

Kisker, C., Kuper, J., & Van Houten, B. (2013). Prokaryotic nucleotide excision repair. *Cold Spring Harbor Perspectives in Biology*, *5*(3), a012591.

Kleibl, K. (2002). Molecular mechanisms of adaptive response to alkylating agents in Escherichia coli and some remarks on O6-methylguanine DNA-methyltransferase in other organisms. *Mutation Research/Reviews in Mutation Research*, *512*(1), 67-84.

Kountz, D. J., & Balskus, E. P. (2021). Leveraging microbial genomes and genomic context for chemical discovery. *Accounts of Chemical Research*, *54*(13), 2788-2797.

Kovačič, L., Paulič, N., Leonardi, A., Hodnik, V., Anderluh, G., Podlesek, Z., Žgur-Bertok, D., Križaj, I., & Butala, M. (2013). Structural insight into LexA–RecA* interaction. *Nucleic acids research*, *41*(21), 9901-9910.

Krokan, H. E., & Bjørås, M. (2013). Base excision repair. *Cold Spring Harbor perspectives in biology*, *5*(4), a012583.

Krwawicz, J., Arczewska, K. D., Speina, E., Maciejewska, A., & Grzesiuk, E. (2007). Bacterial DNA repair genes and their eukaryotic homologues: 1. Mutations in genes involved in base excision repair (BER) and DNA-end processors and their implication in mutagenesis and human disease. *Acta Biochimica Polonica*, *54*(3), 413-434.

Kumar, R., Acharya, V., Mukhia, S., Singh, D., & Kumar, S. (2019). Complete genome sequence of Pseudomonas frederiksbergensis ERDD5: 01 revealed genetic bases for survivability at high altitude ecosystem and bioprospection potential. *Genomics*, *111*(3), 492-499.

Kurthkoti, K., Kumar, P., Sang, P. B., & Varshney, U. (2020). Base excision repair pathways of bacteria: New promise for an old problem. *Future medicinal chemistry*, *12*(04), 339-355.

Lao-Sirieix, S.-h., Pellegrini, L., & Bell, S. D. (2005). The promiscuous primase. *Trends in Genetics*, *21*(10), 568-572.

Laskowski, R. A. (2022). PDBsum 1: A standalone program for generating PDBsum analyses. *Protein Science*, *31*(12), e4473.

Lawley, P., & Phillips, D. (1996). DNA adducts from chemotherapeutic agents. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, *355*(1-2), 13-40.

Le Chatelier, E., Bécherel, O. J., d'Alençon, E., Canceill, D., Ehrlich, S. D., Fuchs, R. P., & Janniere, L. (2004). Involvement of DnaE, the second replicative DNA polymerase from Bacillus subtilis, in DNA mutagenesis. *Journal of Biological Chemistry*, *279*(3), 1757-1767.

Lee, C. K., Laughlin, D. C., Bottos, E. M., Caruso, T., Joy, K., Barrett, J. E., Brabyn, L., Nielsen, U. N., Adams, B. J., & Wall, D. H. (2019). Biotic interactions are an unexpected yet critical control on the complexity of an abiotically driven polar ecosystem. *Communications biology*, *2*(1), 62.

Lee, K.-Y., Lee, K.-Y., Kim, J.-H., Lee, I.-G., Lee, S.-H., Sim, D.-W., Won, H.-S., & Lee, B.-J. (2015). Structure-based functional identification of Helicobacter pylori HP0268 as a nuclease with both DNA nicking and RNase activities. *Nucleic acids research*, *43*(10), 5194-5207.

Lehnman, I. R. (1974). DNA ligase: structure, mechanism, and function. *Science*, *186*(4166), 790-797.

Leiros, I., Moe, E., Lanes, O., Smalås, A. O., & Willassen, N. P. (2003). The structure of uracil-DNA glycosylase from Atlantic cod (Gadus morhua) reveals cold-adaptation features. *Acta Crystallographica Section D: Biological Crystallography*, *59*(8), 1357-1365.

Lieber, M. R. (2010). The mechanism of double-strand DNA break repair by the nonhomologous DNA end-joining pathway. *Annual review of biochemistry*, *79*, 181-211.

Lindahl, T. (1993). Instability and decay of the primary structure of DNA. *Nature*, *362*(6422), 709-715.

Liu, F., Li, N., & Zhang, Y. (2023). The radioresistant and survival mechanisms of Deinococcus radiodurans. *Radiation Medicine and Protection*.

Liu, P., Burdzy, A., & Sowers, L. C. (2004). DNA ligases ensure fidelity by interrogating minor groove contacts. *Nucleic acids research*, *32*(15), 4503-4511.

Lohman, G. J., Bauer, R. J., Nichols, N. M., Mazzola, L., Bybee, J., Rivizzigno, D., Cantin, E., & Evans Jr, T. C. (2016). A high-throughput assay for the comprehensive profiling of DNA ligase fidelity. *Nucleic acids research*, *44*(2), e14-e14.

Luo, J., & Barany, F. (1996). Identification of essential residues in Thermus thermophilus DNA ligase. *Nucleic acids research*, *24*(15), 3079-3085.

Lusetti, S. L., & Cox, M. M. (2002). The bacterial RecA protein and the recombinational DNA repair of stalled replication forks. *Annual review of biochemistry*, *71*(1), 71-100.

Madronich, S. (1994). Increases in biologically damaging UV-B radiation due to stratospheric ozone reductions: a brief review. *Arch. Hydrobiol., 43*, 17-30.

Magana-Schwencke, N., Henriques, J., Chanet, R., & Moustacchi, E. (1982). The fate of 8-methoxypsoralen photoinduced crosslinks in nuclear and mitochondrial yeast DNA: comparison of wild-type and repair-deficient strains. *Proceedings of the National Academy of Sciences*, *79*(6), 1722-1726.

Makarova, K. S., Aravind, L., Grishin, N. V., Rogozin, I. B., & Koonin, E. V. (2002). A DNA repair system specific for thermophilic Archaea and bacteria predicted by genomic context analysis. *Nucleic acids research*, *30*(2), 482-496.

Malu, S., De Ioannes, P., Kozlov, M., Greene, M., Francis, D., Hanna, M., Pena, J., Escalante, C. R., Kurosawa, A., & Erdjument-Bromage, H. (2012). Artemis C-terminal region facilitates V (D) J recombination through its interactions with DNA Ligase IV and DNA-PKcs. *Journal of Experimental Medicine*, *209*(5), 955-963.

Malyarchuk, S., Wright, D., Castore, R., Klepper, E., Weiss, B., Doherty, A. J., & Harrison, L. (2007). Expression of Mycobacterium tuberculosis Ku and Ligase D in Escherichia coli results in RecA and RecB-independent DNA end-joining at regions of microhomology. *DNA repair*, *6*(10), 1413-1424.

Mandel, C. R., Kaneko, S., Zhang, H., Gebauer, D., Vethantham, V., Manley, J. L., & Tong, L. (2006). Polyadenylation factor CPSF-73 is the pre-mRNA 3'-end-processing endonuclease. *Nature*, *444*(7121), 953-956.

Marco, D. (2010). Metagenomics. Theory, Methods and Applications. In.

Markowitz, V. M., Chen, I.-M. A., Palaniappan, K., Chu, K., Szeto, E., Pillay, M., Ratner, A., Huang, J., Woyke, T., & Huntemann, M. (2014). IMG 4 version of the integrated microbial genomes comparative analysis system. *Nucleic acids research*, *42*(D1), D560-D567.

Marti, T., & Fleck, O. (2004). DNA repair nucleases. *Cellular and Molecular Life Sciences CMLS*, *61*, 336-354.

Martin, I. V., & MacNeill, S. A. (2002). ATP-dependent DNA ligases. *Genome biology*, *3*(4), reviews3005. 3001.

Mattimore, V., & Battista, J. R. (1996). Radioresistance of Deinococcus radiodurans: functions necessary to survive ionizing radiation are also necessary to survive prolonged desiccation. *Journal of bacteriology*, *178*(3), 633-637.

McCullough, A. K., Dodson, M., & Lloyd, R. S. (1999). Initiation of base excision repair: glycosylase mechanisms and structures. *Annual review of biochemistry*, *68*(1), 255-285.

McHugh, P. J., Spanswick, V. J., & Hartley, J. A. (2001). Repair of DNA interstrand crosslinks: molecular mechanisms and clinical relevance. *The lancet oncology*, *2*(8), 483-490.

McIlwraith, M. J., Hall, D. R., Stasiak, A. Z., Stasiak, A., Wigley, D. B., & West, S. C. (2001). RadA protein from Archaeoglobus fulgidus forms rings, nucleoprotein filaments and catalyses homologous recombination. *Nucleic acids research*, *29*(22), 4509-4517.

McVey, M., & Lee, S. E. (2008). MMEJ repair of double-strand breaks (director's cut): deleted sequences and alternative endings. *Trends in Genetics*, *24*(11), 529-538.

Micsonai, A., Wien, F., Bulyáki, É., Kun, J., Moussong, É., Lee, Y.-H., Goto, Y., Réfrégiers, M., & Kardos, J. (2018). BeStSel: a web server for accurate protein secondary structure prediction and fold recognition from the circular dichroism spectra. *Nucleic acids research*, *46*(W1), W315-W322.

Misra, H., Rajpurohit, Y., & Kota, S. (2013). Physiological and molecular basis of extreme radioresistance in Deinococcus radiodurans. *Current Science(Bangalore)*, *104*(2), 194-205.

Monaghan, A. J., Bromwich, D. H., Powers, J. G., & Manning, K. W. (2005). The climate of the McMurdo, Antarctica, region as represented by one year of forecasts from the Antarctic Mesoscale Prediction System. *Journal of Climate*, *18*(8), 1174-1189.

Morita, R. Y. (1975). Psychrophilic bacteria. *Bacteriological reviews*, *39*(2), 144.

Moseley, B., & Mattingly, A. (1971). Repair of irradiated transforming deoxyribonucleic acid in wild type and a radiation-sensitive mutant of Micrococcus radiodurans. *Journal of bacteriology*, *105*(3), 976-983.

Mourad, R., Ginalski, K., Legube, G., & Cuvier, O. (2018). Predicting double-strand DNA breaks using epigenome marks or DNA at kilobase resolution. *Genome biology*, *19*(1), 1-14.

Mulder, N., & Apweiler, R. (2007). InterPro and InterProScan: tools for protein sequence classification and comparison. *Comparative genomics*, 59-70.

Nair, P. A., Nandakumar, J., Smith, P., Odell, M., Lima, C. D., & Shuman, S. (2007). Structural basis for nick recognition by a minimal pluripotent DNA ligase. *Nature structural & molecular biology*, *14*(8), 770-778.

Nakae, S., Hijikata, A., Tsuji, T., Yonezawa, K., Kouyama, K.-i., Mayanagi, K., Ishino, S., Ishino, Y., & Shirai, T. (2016). Structure of the EndoMS-DNA complex as mismatch restriction endonuclease. *Structure*, *24*(11), 1960-1971.

Namba, K., & Makino, F. (2022). Recent progress and future perspective of electron cryomicroscopy for structural life sciences. *Microscopy*, *71*(Supplement_1), i3-i14.

Niewolik, D., Peter, I., Butscher, C., & Schwarz, K. (2017). Autoinhibition of the nuclease ARTEMIS is mediated by a physical interaction between its catalytic and C-terminal domains. *Journal of Biological Chemistry*, *292*(8), 3351-3365.

Nikitaki, Z., Hellweg, C. E., Georgakilas, A. G., & Ravanat, J.-L. (2015). Stress-induced DNA damage biomarkers: applications and limitations. *Frontiers in chemistry*, *3*, 35.

Nishida, H., Kiyonari, S., Ishino, Y., & Morikawa, K. (2006). The closed structure of an archaeal DNA ligase from Pyrococcus furiosus. *Journal of molecular biology*, *360*(5), 956-967.

Nishino, T., & Morikawa, K. (2002). Structure and function of nucleases in DNA repair: shape, grip and blade of the DNA scissors. *Oncogene*, *21*(58), 9022-9032.

Noll, D. M., Mason, T. M., & Miller, P. S. (2006). Formation and repair of interstrand cross-links in DNA. *Chemical reviews*, *106*(2), 277-301.

Nunn, B. L., Slattery, K. V., Cameron, K. A., Timmins‐Schiffman, E., & Junge, K. (2015). Proteomics of C olwellia psychrerythraea at subzero temperatures–a life with limited movement, flexible membranes and vital DNA repair. *Environmental microbiology*, *17*(7), 2319-2335.

Oberg, N., Zallot, R., & Gerlt, J. A. (2023). EFI-EST, EFI-GNT, and EFI-CGFP: Enzyme Function Initiative (EFI) Web Resource for Genomic Enzymology Tools. *Journal of molecular biology*, 168018.

Odell, M., & Shuman, S. (1999). Footprinting of Chlorella virus DNA ligase bound at a nick in duplex DNA. *Journal of Biological Chemistry*, *274*(20), 14032-14039.

Ogura, T., & Wilkinson, A. J. (2001). AAA+ superfamily ATPases: common structure–diverse function. *Genes to Cells*, *6*(7), 575-597.

Ordonez, H., & Shuman, S. (2013). Mycobacterium smegmatis Lhr is a DNA-dependent ATPase and a 3′-to-5′ DNA translocase and helicase that prefers to unwind 3′-tailed RNA: DNA hybrids. *Journal of Biological Chemistry*, *288*(20), 14125-14134.

Ouaray, Z., Benner, S. A., Georgiadis, M. M., & Richards, N. G. (2020). Building better polymerases: Engineering the replication of expanded genetic alphabets. *Journal of Biological Chemistry*, *295*(50), 17046-17059.

Panda, A. K., Mishra, R., Miglani, R., Dewali, S., Kumar, A., Bora, S., & Bisht, S. S. (2022). Extremophilic Diversity and Climate Change. In *Biodiversity* (pp. 41-53). CRC Press.

Pascal, J. M., O'Brien, P. J., Tomkinson, A. E., & Ellenberger, T. (2004). Human DNA ligase I completely encircles and partially unwinds nicked DNA. *Nature*, *432*(7016), 473-478.

Pastwa, E., & Błasiak, J. (2003). Non-homologous DNA end joining. *Acta Biochimica Polonica*, *50*(4), 891-908.

Pergolizzi, G., Wagner, G. K., & Bowater, R. P. (2016). Biochemical and structural characterization of DNA ligases from bacteria and archaea. *Bioscience Reports*, *36*(5).

Phillips, R. W., Wiegel, J., Berry, C. J., Fliermans, C., Peacock, A. D., White, D. C., & Shimkets, L. J. (2002). Kineococcus radiotolerans sp. nov., a radiation-resistant, gram-positive bacterium. *International journal of systematic and evolutionary microbiology*, *52*(3), 933-938.

Pingoud, A., Fuxreiter, M., Pingoud, V., & Wende, W. (2005). Type II restriction endonucleases: structure and mechanism. *Cellular and molecular life sciences*, *62*, 685-707.

Poetsch, A. R. (2020). The genomics of oxidative DNA damage, repair, and resulting mutagenesis. *Computational and structural biotechnology journal*, *18*, 207-219.

Pointing, S. B., Bollard-Breen, B., & Gillman, L. N. (2014). Diverse cryptic refuges for life during glaciation. *Proceedings of the National Academy of Sciences*, *111*(15), 5452-5453.

Pointing, S. B., Chan, Y., Lacap, D. C., Lau, M. C., Jurgens, J. A., & Farrell, R. L. (2009). Highly specialized microbial diversity in hyper-arid polar desert. *Proceedings of the National Academy of Sciences*, *106*(47), 19964-19969.

Popovic, A., Hai, T., Tchigvintsev, A., Hajighasemi, M., Nocek, B., Khusnutdinova, A. N., Brown, G., Glinos, J., Flick, R., & Skarina, T. (2017). Activity screening of environmental metagenomic libraries reveals novel carboxylesterase families. *Scientific reports*, *7*, 44103.

Potter, S. C., Luciani, A., Eddy, S. R., Park, Y., Lopez, R., & Finn, R. D. (2018). HMMER web server: 2018 update. *Nucleic acids research*, *46*(W1), W200-W204.

Raggio, J., Green, T., & Sancho, L. (2016). In situ monitoring of microclimate and metabolic activity in lichens from Antarctic extremes: a comparison between South Shetland Islands and the McMurdo Dry Valleys. *Polar Biology*, *39*, 113-122.

Rangarajan, E. S., & Shankar, V. (2001). Sugar non-specific endonucleases. *FEMS microbiology reviews*, *25*(5), 583-613.

Rasband, W. S. (2011). National Institutes of Health, Bethesda, Maryland, USA. *http://imagej. nih. gov/ij/*.

Reardon, J. T., & Sancar, A. (2006). Purification and characterization of Escherichia coli and human nucleotide excision repair enzyme systems. *Methods in enzymology*, *408*, 189-213.

Reich, C. I., McNeil, L. K., Brace, J. L., Brucker, J. K., & Olsen, G. J. (2001). Archaeal RecA homologues: different response to DNA-damaging agents in mesophilic and thermophilic Archaea. *Extremophiles*, *5*, 265-275.

Ren, B., Kuhn, J., Meslet-Cladiere, L., Myllykallio, H., & Ladenstein, R. (2007). Crystallization and preliminary X-ray analysis of a RecB-family nuclease from the archaeon Pyrococcus abyssi. *Acta Crystallographica Section F: Structural Biology and Crystallization Communications*, *63*(5), 406-408.

Ren, B., Kühn, J., Meslet‐Cladiere, L., Briffotaux, J., Norais, C., Lavigne, R., Flament, D., Ladenstein, R., & Myllykallio, H. (2009). Structure and function of a novel endonuclease acting on branched DNA substrates. *The EMBO journal*, *28*(16), 2479-2489.

Robert, X., & Gouet, P. (2014). Deciphering key features in protein structures with the new ENDscript server. *Nucleic acids research*, *42*(W1), W320-W324.

Rodríguez, A. C., & Stock, D. (2002). Crystal structure of reverse gyrase: insights into the positive supercoiling of DNA. *The EMBO journal*, *21*(3), 418-426.

Rolland, J.-l., Gueguen, Y., Persillon, C., Masson, J.-M., & Dietrich, J. (2004). Characterization of a thermophilic DNA ligase from the archaeon Thermococcus fumicolans. *FEMS microbiology letters*, *236*(2), 267-273.

Rzoska-Smith, E., Stelzer, R., Monterio, M., Cary, S. C., & Williamson, A. (2023). DNA repair enzymes of the Antarctic Dry Valley metagenome. *Frontiers in microbiology*, *14*.

Sallmyr, A., Rashid, I., Bhandari, S. K., Naila, T., & Tomkinson, A. E. (2020). Human DNA ligases in replication and repair. *DNA repair*, *93*, 102908.

Salvatore, M. R., & Levy, J. S. (2021). The McMurdo Dry Valleys of Antarctica: a geological, environmental, and ecological analog to the Martian surface and near surface. In *Mars geological enigmas* (pp. 291-332). Elsevier.

Sandigursky, M., & Franklin, W. A. (1999). Thermostable uracil-DNA glycosylase from Thermotoga maritima a member of a novel class of DNA repair enzymes. *Current biology*, *9*(10), 531-534.

Sarmiento, F., Peralta, R., & Blamey, J. M. (2015). Cold and hot extremozymes: industrial relevance and current trends. *Frontiers in Bioengineering and Biotechnology*, *3*, 148.

Schmiester, M., & Demuth, I. (2017). SNM1B/Apollo in the DNA damage response and telomere maintenance. *Oncotarget*, *8*(29), 48398.

Schröder, C., Burkhardt, C., & Antranikian, G. (2020). What we learn from extremophiles. *ChemTexts*, *6*, 1-6.

Schrödinger, L., & DeLano, W. . (2020). *PyMOL*. Retrieved from http://www.pymol.org/pymol

Seckbach, J., & Oren, A. (2005). Introduction to the extremophiles. In *Origins: Genesis, Evolution and Diversity of Life* (pp. 371-396). Springer.

Sengerová, B., Allerston, C. K., Abu, M., Lee, S. Y., Hartley, J., Kiakos, K., Schofield, C. J., Hartley, J. A., Gileadi, O., & McHugh, P. J. (2012). Characterization of the human SNM1A and SNM1B/Apollo DNA repair exonucleases. *Journal of Biological Chemistry*, *287*(31), 26254-26267.

Seo, M. S., Kim, Y. J., Choi, J. J., Lee, M. S., Kim, J. H., Lee, J.-H., & Kwon, S.-T. (2007). Cloning and expression of a DNA ligase from the hyperthermophilic archaeon Staphylothermus marinus and properties of the enzyme. *Journal of biotechnology*, *128*(3), 519-530.

Sfeir, A., & Symington, L. S. (2015). Microhomology-mediated end joining: a back-up survival mechanism or dedicated pathway? *Trends in biochemical sciences*, *40*(11), 701-714.

Sharma, J. K., Stevenson, L. J., Robins, K. J., Williamson, A. K., Patrick, W. M., & Ackerley, D. F. (2020). Methods for competitive enrichment and evaluation of superior DNA ligases. In *Methods in Enzymology* (Vol. 644, pp. 209-225). Elsevier.

Sharples, G. J., Ingleston, S. M., & Lloyd, R. G. (1999). Holliday junction processing in bacteria: insights from the evolutionary conservation of RuvABC, RecG, and RusA. *Journal of bacteriology*, *181*(18), 5543-5550.

Shi, K., Bohl, T. E., Park, J., Zasada, A., Malik, S., Banerjee, S., Tran, V., Li, N., Yin, Z., & Kurniawan, F. (2018). T4 DNA ligase structure reveals a prototypical ATP-dependent ligase with a unique mode of sliding clamp interaction. *Nucleic acids research*, *46*(19), 10474-10488.

Shi, K., Moeller, N. H., Banerjee, S., McCann, J. L., Carpenter, M. A., Yin, L., Moorthy, R., Orellana, K., Harki, D. A., & Harris, R. S. (2021). Structural basis for recognition of distinct deaminated DNA lesions by endonuclease Q. *Proceedings of the National Academy of Sciences*, *118*(10), e2021120118.

Shinagawa, H., Iwasaki, H., Kato, T., & Nakata, A. (1988). RecA protein-dependent cleavage of UmuD protein and SOS mutagenesis. *Proceedings of the National Academy of Sciences*, *85*(6), 1806-1810.

Shuman, S. (2009). DNA ligases: progress and prospects. *Journal of Biological Chemistry*, *284*(26), 17365-17369.

Shuman, S., & Glickman, M. S. (2007). Bacterial DNA repair by non-homologous end joining. *Nature Reviews Microbiology*, *5*(11), 852-861.

Shuman, S., & Lima, C. D. (2004). The polynucleotide ligase and RNA capping enzyme superfamily of covalent nucleotidyltransferases. *Current opinion in structural biology*, *14*(6), 757-764.

Simmons, L. A., Foti, J. J., Cohen, S. E., & Walker, G. C. (2008). The SOS regulatory network. *EcoSal Plus*, *2008*.

Singh, N. P., Stephens, R. E., Singh, H., & Lai, H. (1999). Visual quantification of DNA double-strand breaks in bacteria. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, *429*(2), 159-168.

Singh, O. V., & Gabani, P. (2011). Extremophiles: radiation resistance microbial reserves and therapeutic implications. *Journal of applied microbiology*, *110*(4), 851-861.

Singh, P., Jain, K., Desai, C., Tiwari, O., & Madamwar, D. (2019). Microbial community dynamics of extremophiles/extreme environment. In *Microbial diversity in the genomic era* (pp. 323-332). Elsevier.

Sinha, S., Villarreal, D., Shim, E. Y., & Lee, S. E. (2016). Risky business: Microhomology-mediated end joining. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, *788*, 17-24.

Sriskanda, V., Kelman, Z., Hurwitz, J., & Shuman, S. (2000). Characterization of an ATP-dependent DNA ligase from the thermophilic archaeon Methanobacterium thermoautotrophicum. *Nucleic acids research*, *28*(11), 2221-2228.

Sriskanda, V., Moyer, R. W., & Shuman, S. (2001). NAD+-dependent DNA ligase encoded by a eukaryotic virus. *Journal of Biological Chemistry*, *276*(39), 36100-36109.

Sriskanda, V., & Shuman, S. (2001). A second NAD+-dependent DNA ligase (LigB) in Escherichia coli. *Nucleic acids research*, *29*(24), 4930-4934.

Sriskanda, V., & Shuman, S. (2002). Conserved residues in domain Ia are required for the reaction of Escherichia coli DNA ligase with NAD+. *Journal of Biological Chemistry*, *277*(12), 9695-9700.

Srivastava, S. K., Tripathi, R. P., & Ramachandran, R. (2005). NAD+-dependent DNA ligase (Rv3014c) from Mycobacterium tuberculosis: crystal structure of the adenylation domain and identification of novel inhibitors. *Journal of Biological Chemistry*, *280*(34), 30273-30281.

Stojic, L., Brun, R., & Jiricny, J. (2004). Mismatch repair and DNA damage signalling. *DNA repair*, *3*(8-9), 1091-1101.

Stomeo, F., Makhalanyane, T. P., Valverde, A., Pointing, S. B., Stevens, M. I., Cary, C. S., Tuffin, M. I., & Cowan, D. A. (2012). Abiotic factors influence microbial diversity in permanently cold soil horizons of a maritime-associated Antarctic Dry Valley. *FEMS microbiology ecology*, *82*(2), 326-340.

Sun, Y., Seo, M. S., Kim, J. H., Kim, Y. J., Kim, G. A., Lee, J. I., Lee, J. H., & Kwon, S. T. (2008). Novel DNA ligase with broad nucleotide cofactor specificity from the hyperthermophilic crenarchaeon Sulfophobococcus zilligii: influence of ancestral DNA ligase on cofactor utilization. *Environmental microbiology*, *10*(12), 3212-3224.

Suzuki, M., Hayashi, H., Mizuki, T., Maekawa, T., & Morimoto, H. (2016). Efficient DNA ligation by selective heating of DNA ligase with a radio frequency alternating magnetic field. *Biochemistry and biophysics reports*, *8*, 360-364.

Takai, K., Nakamura, K., Toki, T., Tsunogai, U., Miyazaki, M., Miyazaki, J., Hirayama, H., Nakagawa, S., Nunoura, T., & Horikoshi, K. (2008). Cell proliferation at 122 C and isotopically heavy CH4 production by a hyperthermophilic methanogen under high-pressure cultivation. *Proceedings of the National Academy of Sciences*, *105*(31), 10949-10954.

Takemoto, N., Numata, I., Su'etsugu, M., & Miyoshi-Akiyama, T. (2018). Bacterial EndoMS/NucS acts as a clamp-mediated mismatch endonuclease to prevent asymmetric accumulation of replication errors. *Nucleic acids research*, *46*(12), 6152-6165.

Tamppari, L., Anderson, R., Archer, P., Douglas, S., Kounaves, S., McKay, C., Ming, D., Moore, Q., Quinn, J., & Smith, P. (2012). Effects of extreme cold and aridity on soils and habitability: McMurdo Dry Valleys as an analogue for the Mars Phoenix landing site. *Antarctic Science*, *24*(3), 211-228.

Tang, Q., Gulkis, M., McKenna, R., & Çağlayan, M. (2022). Structures of LIG1 that engage with mutagenic mismatches inserted by polβ in base excision repair. *Nature communications*, *13*(1), 3860.

Taylor, M. R. (2014). *The role of divalent metal ions in enzymatic DNA ligation*

Theobald, D. L., Mitton-Fry, R. M., & Wuttke, D. S. (2003). Nucleic acid recognition by OB-fold proteins. *Annual review of biophysics and biomolecular structure*, *32*(1), 115-133.

Thompson, L. R., Sanders, J. G., McDonald, D., Amir, A., Ladau, J., Locey, K. J., Prill, R. J., Tripathi, A., Gibbons, S. M., & Ackermann, G. (2017). A communal catalogue reveals Earth's multiscale microbial diversity. *Nature*, *551*(7681), 457-463.

Thompson, P. S., & Cortez, D. (2020). New insights into abasic site repair and tolerance. *DNA repair*, *90*, 102866.

Tiao, G., Lee, C. K., McDonald, I. R., Cowan, D. A., & Cary, S. C. (2012). Rapid microbial response to the presence of an ancient relic in the Antarctic Dry Valleys. *Nature communications*, *3*(1), 660.

Tomkinson, A. E., Totty, N. F., Ginsburg, M., & Lindahl, T. (1991). Location of the active site for enzyme-adenylate formation in DNA ligases. *Proceedings of the National Academy of Sciences*, *88*(2), 400-404.

Trivedi, S., Rao, S. R., & Gehlot, H. S. (2005). Nucleic acid stability in thermophilic prokaryotes: a review. *J Cell Mol Biol*, *4*, 61-69.

Tropea, J. E., Cherry, S., & Waugh, D. S. (2009). Expression and purification of soluble His 6-tagged TEV protease. *High throughput protein expression and purification: methods and protocols*, 297-307.

Truong, L. N., Li, Y., Shi, L. Z., Hwang, P. Y.-H., He, J., Wang, H., Razavian, N., Berns, M. W., & Wu, X. (2013). Microhomology-mediated End Joining and Homologous Recombination share the initial end resection step to repair DNA double-strand breaks in mammalian cells. *Proceedings of the National Academy of Sciences*, *110*(19), 7720-7725.

Ujaoney, A. K., Padwal, M. K., & Basu, B. (2021). An in vivo interaction network of DNA-repair proteins: a snapshot at double strand break repair in Deinococcus radiodurans. *Journal of Proteome Research*, *20*(6), 3242-3255.

UNSCEAR. (2000). United Nations Scientific Committee on the effects of atomic radiation. Effects and risks of ionizing radiations. In: UN New York.

Varadi, M., Anyango, S., Deshpande, M., Nair, S., Natassia, C., Yordanova, G., Yuan, D., Stroe, O., Wood, G., & Laydon, A. (2022). AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic acids research*, *50*(D1), D439-D444.

Wallace, S. S. (2014). Base excision repair: a critical player in many games. *DNA repair*, *19*, 14-26.

Wang, E., Koutsioulis, D., Leiros, H.-K. S., Andersen, O. A., Bouriotis, V., Hough, E., & Heikinheimo, P. (2007). Crystal structure of alkaline phosphatase from the Antarctic bacterium TAB5. *Journal of molecular biology*, *366*(4), 1318-1331.

Wang, L., Xi, Y., Zhang, W., Wang, W., Shen, H., Wang, X., Zhao, X., Alexeev, A., Peters, B. A., & Albert, A. (2019). 3′ Branch ligation: a novel method to ligate non-complementary DNA to recessed or internal 3′ OH ends in DNA or RNA. *DNA Research*, *26*(1), 45-53.

Warner, H. R. (1983). Base excision repair in the thermophile Thermus sp. strain X-1. *Journal of bacteriology*, *154*(3), 1451-1454.

Watford, S., & Warrington, S. J. (2017). Bacterial DNA mutations.

Watson, J. D., Baker, T. A., Bell, S. P., Gann, A., Levine, M., & Losick, R. (2004). Molecular Biology of the Gene (International Ed.).

Wei, S. T., Lacap-Bugler, D. C., Lau, M. C., Caruso, T., Rao, S., De Los Rios, A., Archer, S. K., Chiu, J. M., Higgins, C., & Van Nostrand, J. D. (2016). Taxonomic and functional diversity of soil and hypolithic microbial communities in Miers Valley, McMurdo Dry Valleys, Antarctica. *Frontiers in microbiology*, *7*, 1642.

Welsh, D. T. (2000). Ecological significance of compatible solute accumulation by micro-organisms: from single cells to global climate. *FEMS microbiology reviews*, *24*(3), 263-290.

Wilkinson, A., Day, J., & Bowater, R. (2001). Bacterial DNA ligases. *Molecular microbiology*, *40*(6), 1241-1248.

Williamson, A., Hjerde, E., & Kahlke, T. (2016). Analysis of the distribution and evolution of the ATP‐dependent DNA ligases of bacteria delineates a distinct phylogenetic group 'L ig E'. *Molecular microbiology*, *99*(2), 274-290.

Williamson, A., & Leiros, H.-K. S. (2020). Structural insight into DNA joining: from conserved mechanisms to diverse scaffolds. *Nucleic acids research*, *48*(15), 8225-8242.

Williamson, A., Rothweiler, U., & Leiros, H.-K. (2014). Enzyme–adenylate structure of a bacterial ATP-dependent DNA ligase with a minimized DNA-binding surface. *Acta Crystallographica Section D: Biological Crystallography*, *70*(11), 3043-3056.

Wong, C., Sridhara, S., Bardwell, J. C., & Jakob, U. (2000). Heating greatly speeds Coomassie blue staining and destaining. *Biotechniques*, *28*(3), 426-432.

Wood, S. A., Rueckert, A., Cowan, D. A., & Cary, S. C. (2008). Sources of edaphic cyanobacterial diversity in the Dry Valleys of Eastern Antarctica. *The ISME Journal*, *2*(3), 308-320.

Wozniak, K. J., & Simmons, L. A. (2022). Bacterial DNA excision repair pathways. *Nature Reviews Microbiology*, *20*(8), 465-477.

Wright, D., DeBeaux, A., Shi, R., Doherty, A. J., & Harrison, L. (2010). Characterization of the roles of the catalytic domains of Mycobacterium tuberculosis ligase D in Ku-dependent error-prone DNA end joining. *Mutagenesis*, *25*(5), 473-481.

Wu, D. Y., & Wallace, R. B. (1989). Specificity of the nick-closing activity of bacteriophage T4 DNA ligase. *Gene*, *76*(2), 245-254.

Wüthrich, K. (2001). The way to NMR structures of proteins. *Nature structural biology*, *8*(11), 923-925.

Wynn-Williams, D. D. (1990). Ecological aspects of Antarctic microbiology. In *Advances in microbial ecology* (pp. 71-146). Springer.

Yang, W. (2011). Nucleases: diversity of structure, function and mechanism. *Quarterly reviews of biophysics*, *44*(1), 1-93.

Yi, C., & He, C. (2013). DNA repair by reversal of DNA damage. *Cold Spring Harbor Perspectives in Biology*, *5*(1), a012575.

Yosaatmadja, Y., Baddock, H. T., Newman, J. A., Bielinski, M., Gavard, A. E., Mukhopadhyay, S. M., Dannerfjord, A. A., Schofield, C. J., McHugh, P. J., & Gileadi, O. (2021). Structural and mechanistic insights into the Artemis endonuclease and strategies for its inhibition. *Nucleic acids research*, *49*(16), 9310-9326.

Yung, C. C., Chan, Y., Lacap, D. C., Pérez-Ortega, S., de los Rios-Murillo, A., Lee, C. K., Cary, S. C., & Pointing, S. B. (2014). Characterization of chasmoendolithic community in miers valley, Mcmurdo dry valleys, antarctica. *Microbial ecology*, *68*(2), 351-359.

Zallot, R., Oberg, N. O., & Gerlt, J. A. (2018). 'Democratized'genomic enzymology web tools for functional assignment. *Current opinion in chemical biology*, *47*, 77-85.

Zhang, A. P., Pigli, Y. Z., & Rice, P. A. (2010). Structure of the LexA–DNA complex and implications for SOS box measurement. *Nature*, *466*(7308), 883-886.

Zhang, L., Jiang, D., Wu, M., Yang, Z., & Oger, P. M. (2020). New insights into DNA repair revealed by NucS endonucleases from hyperthermophilic archaea. *Frontiers in microbiology*, *11*, 1263.

Zhang, L., & Tripathi, A. (2017). Archaeal RNA ligase from Thermoccocus kodakarensis for template dependent ligation. *RNA biology*, *14*(1), 36-44.

Zhu, D., Adebisi, W. A., Ahmad, F., Sethupathy, S., Danso, B., & Sun, J. (2020). Recent development of extremophilic bacteria and their application in biorefinery. *Frontiers in Bioengineering and Biotechnology*, 483.

Zhu, H., & Shuman, S. (2006). Substrate specificity and structure-function analysis of the 3′ -phosphoesterase component of the bacterial NHEJ protein, DNA ligase D. *Journal of Biological Chemistry*, *281*(20), 13873-13881.

Zhu, H., & Shuman, S. (2007). Characterization of Agrobacterium tumefaciens DNA ligases C and D. *Nucleic acids research*, *35*(11), 3631-3645.

Appendices

Appendix A Methods

A.1 List of genes used in experiments

DV-metagenome gene sequences were deposited in the JGI/IMG database under the gene IDs listed in the table below.

Table A.1. DNA repair genes identified through SSN, from Dry Valley metagenomes, with gene ID and genome name for location of genes in JGI/IMG database.

Gene:	Gene ID:	Genome name:
DV-1-1-Lig-Nuc	Ga0136611_1000086013	UQ223
DV-1-2-RecA	Ga0136611_1000086014	UQ223
DV-1-3-DNA polymerase	Ga0136611_1000086015	UQ223
DV-Lig2	Ga0136636_1000055115	UQ852
DV-Lig5	Ga0136613_1000000468	UQ272
DV-Nuc3	Ga0136640_100017415	UQ864

A.2 Growth media

Growth media was used to make agar plates for plating transformations and as a growth medium for *E. coli* cultures, for plasmid extraction and for small- and large-scale protein expression. All dehydrated media used was supplied by Difco. All media was prepared using standard recipes. Liquid media was prepared in 1L glass bottles. LB agar media; was prepared, as below, with the addition of 15 g agar. Solid media was prepared in 500ml or 1L glass bottle and stored in molten form at 50 ºC until needed. TB media for expression growth was made by adding 200 mls of 5x TB media, 100 mls of 10x phosphate buffer and 700 mls of MQ H_2O.

Table A.2. Media components used in the expression of recombinant proteins

Media	Composition
LB broth	10 g Peptone, 5 g yeast extract, 10 g NaCl
5x TB media	60g Tryptone, 120 g Yeast extract, 10 % glycerol
10x Phosphate buffer	23.1 g KH_2PO_4, 125.4 g $K_2H_2PO_4$

A.3 Antibiotic stocks

Antibiotics used in growth media were diluted from 1000x concentrated stocks that were prepared in bulk and stored at -20 °C until use. Stocks were solubilised in MQ H_2O (unless otherwise stated) at concentrations of; ampicillin (100 mg/ml), and kanamycin (50 mg/ml).

A.4 Primer sequences

Table A.4. List of primers used in various PCR reactions

Primer name:	Primer sequence:
DV-1-1Nuc FD	**GAGAACCTGTATTTTCAGGGTCATCGT**
DV1-1Nuc BK	GGGGACCACTTTGTACAAGAAAGCTGGGTATAA**ACCCGGTGCGCTAGG**
DV-1-1Lig FD	GAGAACCTGTATTTTCAGGGT**GATTTTGCACGTTTTGCC**
DV1-1Lig BK	**GGGGACCACTTTGTACAAGAAAGCTGGGTCTTATTCGGTATCTGCTTT**
DV-1-1 FD 2 *	GGGGACAAGTTTGTACAAAAAAGCAGGCTTA**GAGAACCTGTATTTTCAGGGT**
DV-1-1Nuc FD 2	GAGAACCTGTATTTTCAGGGT**GATTTTGCACGTTTTGCC**
DV-1-1Nuc BK 2	GGGGACCACTTTGTACAAGAAAGCTGGGTATAA**ACCCGGTGCGCTAGG**
M13 FD	GTAAAACGACGGCCAGT
M13 BK	CAGGAAACAGCTATGACC
T7 FD	TAATACGACTCACTATAGGG
T7 BK	GCTAGTTATTGCTCAGCGG
MBP FD	GATGAAGCCCTGAAAGACGCGCAG

Forward primers (FD)
Back/reverse primers (BK)
Underline sequences in bold represent overhangs
*DV-1-1 FD 2 was used in round 2 of PCR to add attB1 sites to new constructs

A.5 Construct design for DV-1-1-Lig-Nuc, DV-1-1-Lig and DV-1-1-Nuc

Constructs were designed as described in **Section 2.2.2**. Primers used are represented in **Table A.4** above.

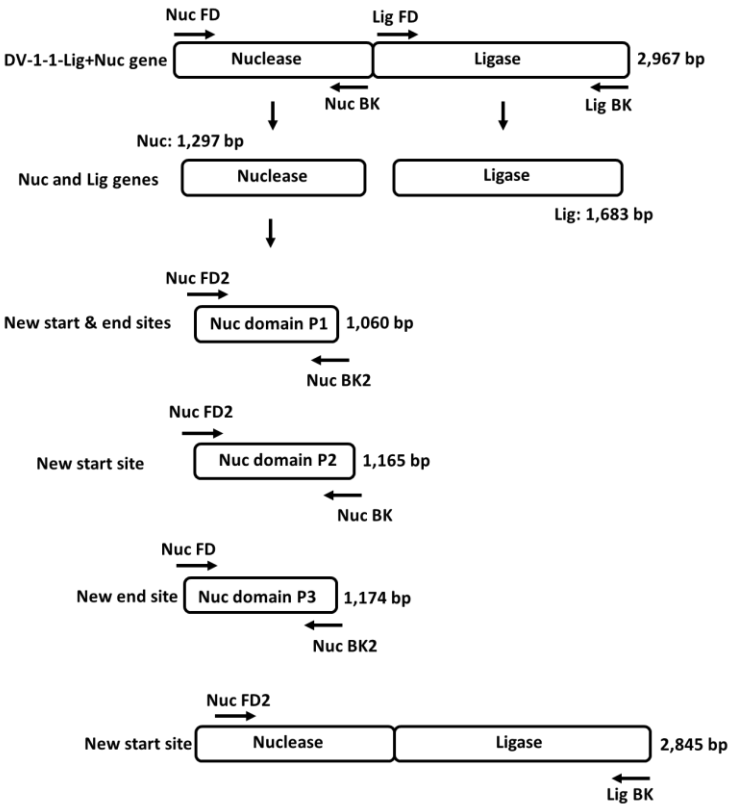


Figure A.5. Schematic of new constructs to separate the ligase and nuclease domains from the ligase nuclease fusion (DV-1-1-Lig-Nuc) and design of DV-1-1-Lig-Nuc with new start site. Primers used in PCR are indicated on the figure with arrows. Product size of new constructs are indicated on the figure.

A.6 PCR cycling conditions

PCR cycling conditions used for design of new constructs and colony PCR.

	Steps	Temperature (°C)	Length (minutes)
	Pre-denaturation	95	15:00
x 29	Denaturation	95	00:30
	Annealing	-	00:20
	Extension	72	00:30
	Final extension	72	10:00

Annealing T_m for required PCR are detailed in **Section 2.2.3.**

A.7 Composition of buffers used for purification of recombinant proteins

Buffers	Components
Lysis buffer	50 mM Tris pH 8.0, 750 mM NaCl, 1mM MgCl2, 5% glycerol

Buffer A	50 mM Tris pH 8.0, 750 mM NaCl, 5% glycerol, 10 mM imidazole
Buffer B	50 mM Tris pH 8.0, 750 mM NaCl, 5% glycerol, 500 mM imidazole
Buffer C	50 mM Tris pH 8.0, 200 mM NaCl, 1mM DTT, 5% glycerol
Wash buffer	50 mM Tris pH 8.0, 800 mM NaCl, 40 mM imidazole, 30 % glycerol
MBP binding buffer	20 mM Tris-HCl, 200 mM NaCl, 1 mM EDTA, pH 7.4
MBP elution buffer	20 mM Tris-HCl, 200 mM NaCl, 1 mM EDTA, 10 mM maltose, pH 7.4

A.8 Results from cloning and protein expression trials

Table A.8. Proteins involved in this study and results from cloning and protein expression trials.

Protein	**Successful cloning**	**Soluble expression**
DV-1-1-Lig-Nuc	Yes	No
DV-1-2-RecA	No	N/A
DV-1-3 DNA polymerase	Yes	No
DV-1-1-Lig	Yes	Yes
DV-1-1-Nuc	Yes	No
DV-1-1-Nuc construct 1	Yes	Yes
DV-1-1-Nuc construct 2	Yes	Yes
DV-1-1-Nuc construct 3	Yes	No
DV-1-1-Nuc construct 1 mutant	Yes	Yes
DV-1-1-Lig-Nuc (new start site)	Yes	Yes
DV-Nuc3	Yes	Yes
DV-Nuc3 mutant	Yes	Yes
DV-Nuc3 N-terminal truncation	Yes	No
DV-Lig2	Yes	Yes
DV-Lig5	Yes	Yes

A.9 Optimisation results

Table A.9. Optimized expression conditions used for protein production in this study.

Protein	**Expression conditions**			**Purification conditions**	
	Plasmid	***E. coli* strain**	**Temperature**	**Tev**	**Gel filtration**

				cleavage	column
DV-1-1-Lig-Nuc*	pHMGWA	Origami (DE3)	15 ^{o}C	No	S200
DV-1-1-Lig	pDEST17	Origami (DE3)	15 ^{o}C	No	S200
DV-1-1-Nuc P1	pHMGWA	Origami (DE3)	25 ^{o}C	Yes and No	S200/S75
DV-1-1-Nuc mut	pHMGWA	Origami (DE3)	25 ^{o}C	Yes	S200/S75
DV-Nuc3	pHMGWA	BL21 pLysS	15 ^{o}C	Yes	S200
DV-Nuc3 mut	pHMGWA	BL21 pLysS	15 ^{o}C	Yes	S200
DV-Lig5	pHMGWA	Origami (DE3)	25 ^{o}C	Yes	S200
DV-Lig2	pHMGWA	Origami (DE3)	15 ^{o}C	No	S200

* New start site

pDEST17 plasmid contains a His-tag

pHMGWA contains both His-tag and MBP tag

A.10 Gene and protein information

All sequences were codon optimised using software available from Twistbioscience.com. Gateway attL cloning sites are underlined. TEV cleavage sites are in red text. The start position of the gene sequence is indicated in bold. Protein translated products follow each gene sequence.

DV-Lig2 gene sequence in pDONR221 plasmid (1,640 bp, excluding attL sites)

CAAATAATGATTTTATTTTGACTGATAGTGACCTGTTCGTTGCAACAGATTGATGAGCAATGCTTTTTTATAAT
GCCAACTTTGTACAAAAAAGCAGGCTATCATCATCATCATCATCATGAGAATCTCTATTTTCAAGGC**AAG**GC
CTTTGCAGCACTCTATGCCGAGCTTGATGCCAGCACTGCCACGCGCGATAAGGTGGCCGCCATGGCAGCATA
TTTTACGCATGCAGCCGCCGCAGATGCAGCATGGGCCTTGTGGTTTCTCGCTGGCGAGCGCCTCAAGCGCATT
GCCGGCTCGGCACTTTTACGTGAGTTGCTCGGTCGCGCATCCGGATATCCTGCCTGGCTTGTTGAGGATAGCT
ATGCCCATGTGGGCGATTTAGCCGAGACGATTACGTTGCTTACGCATGGCAATCCACAAGATGTGCCTGAGC
GCCCATTACATGCATGGGTCGCTGCCCTTCGCGAGCTTCCCACATTAGCACCCGAGGTGCGCGATGAGCGCA
TTCTCGATTGGTGGAAGACGTTGGGCGATGAGCAACGCTATGTCCTTAATAAGTTATTGACTGGCGGGTTAC
GCGTGGGAGTAGCACAACGCCTCGTGGTGCTTGCTATTGCCCAAGCCTTTGATTTACCTGCAGATCGCATTGC
CCAACGCCTTGCCGGCAGTTGGGAGCCCACGCCTGCATCGTGGGCCCGCCTCACGGCTGATGCCGCTACTGA
TGATGATCGCAGTGATCAACCATATCCCTTCTTCCTTGCCAGCGCCCTTGAGCAACCAGTAGAGGCCTTGGGC
ACTTGTACGGCATGGCTCGCAGAGTGGAAGTGGGATGGCATTCGCGCCCAACTCATTCGCCGTGGCGATACT
GTTGCCGTCTGGAGCCGCGGTGAGGAACGTCTTGATGGCCGCTTTCCTGAAATTGAGCGCGCTGCACTCGCC
CTTCCCGGCGGGTGTGTCTTAGATGGCGAGATTTTGGCATGGCGCGATGGCCGCCCTCTCCCCTTTAATTTGC
TCCAAAAGCGCATTGGGCGCTTACGCCCCGGCGCCCGCTCGTTGACAGAGGCTCCTGTAGCCTTTGTAGCTTA
TGATTTACTTGAGTTTGATGGAACTGATCGCCGCACTTTGCCCTTAGATGAGCGCAAGCGCTTACTTGGCATT
GCATTACAAGATGCAGCCAATCCTAGCGTACTCTTAGCTAGTCCAACGATTGAGGCAGATGATTGGCAAACG
CTCGTAGCAGCACGCGAGCAAGCCCGCACACAAGGCGTCGAGGGCTTAATGCTTAAGCGCCGCGATAGCGC
ATATCAAACTGGACGCCGCCGCGGGGATTGGTATAAGTGGAAGGTGAGTCCTTTTACGTTGGATGCAGTGCT
TATTTATGCTCAAGCCGGCCATGGCCGCCGCAGCAATTTGTATACGGATTATACGTTTGCCGTATGGGATGGC
GAGACACTCGTACCAGTCGCCAAGGCATATAGCGGCTTAGATGATGCAGAGATTGCCCGCTTGGATCGCTGG
ATTCGCGCCCATACTCGCGAGCGCTTTGGACCTGTACGCTCGGTGGAGCCTTTACAAGTGTTTGAGTTGGCAT
TTGAGGGCGTGGCACGCAGCACGCGCCATAAGTCAGGCGTCGCTGTCCGCTTTCCTCGCATTCTCCGCTGGCG
CGAGGATAAGCCAGCTACGCAAGCAGATCGCCTTCAAACTCTCCAATCAATGGCGTCCGCCGGCGCCTGACA
GCTTTCTTGTACAAAGTTGGCATTATAAGAAAGCATTGCTTATCAATTTGTTGCAACGAACAGGTCACTATCA
GTCAAAATAAAATCATTATTTG

DV-Lig2 translated protein product (63.51 kDa)

MSYYHHHHHHLESTSLYKKAGYHHHHHHENLYFQG**K**AFAALYAELDASTATRDKVAAMAAYFTHAAAADAA
WALWFLAGERLKRIAGSALLRELLGRASGYPAWLVEDSYAHVGDLAETITLLTHGNPQDVPERPLHAWVAALRE
LPTLAPEVRDERILDWWKTLGDEQRYVLNKLLTGGLRVGVAQRLVVLAIAQAFDLPADRIAQRLAGSWEPTPAS

WARLTADAATDDDRSDQPYPFFLASALEQPVEALGTCTAWLAEWKWDGIRAQLIRRGDTVAVWSRGEERLDGR
FPEIERAALALPGGCVLDGEILAWRDGRPLPFNLLQKRIGRLRPGARSLTEAPVAFVAYDLLEFDGTDRRTLPLDE
RKRLLGIALQDAANPSVLLASPTIEADDWQTLVAAREQARTQGVEGLMLKRRDSAYQTGRRRGDWYKWKVSPF
TLDAVLIYAQAGHGRRSNLYTDYTFAVWDGETLVPVAKAYSGLDDAEIARLDRWIRAHTRERFGPVRSVEPLQV
FELAFEGVARSTRHKSGVAVRFPRILRWREDKPATQADRLQTLQSMASAGA*

DV-Lig5 gene sequence in pDONR221 plasmid (1,557 bp, excluding attL sites)

CAAATAATGATTTTATTTTGACTGATAGTGACCTGTTCGTTGCAACAGATTGATGAGCAATGCTTTTTTATAAT
GCCAACTTTGTACAAAAAAGCAGGCTATCATCATCATCATCATCATGAGAATCTTTATTTTCAAGGC**TTA**AAT
GAGTTAGTCAGCACATCAGCCGAGGTCACGAAGACGAGCAGCCGCAAGGCCAAGGCCGCACTTCTTGCAAA
GTTACTTCGCCAACTTGAGACGGGCGAGGTGGCCACTGCCGTGGGCCTTTTAATTGGCCAACCTCGCCAACG
CCGCTTAGGAGTGGGGTTTGGCAGCGTGTTTAATCTTGATATTGAGCCAGCCACTCATCCTACGCTTACGATT
GATGAGGTGGATGCCGCTTTCTCATCATTGCAACGCGCTGGCGGAGCTGGCAGCCAAAAGTCCCGCAATGAT
TTACTCACGGGCCTTATGAAGCGCGCCACGTCGAGCGAGCAAGCATTTCTCCGCCAAGTCCTCACTGGGGAA
GTCCGCCAAGGCGCTCTCGGCGGCGTGCTTACTGAGGCTGTGGCATTGGGCTTTGAGGTACCGCCCGAAGTA
GTGCGTCGCGCATCAATGCTTCGCGGCGATTTGGGCCAAGTGGCAGAGGTAGCTGCTGTGGGCGGCGTGGTG
GGGTTAGAGTCAATTGGCTTACGCGTGTTAACGCCTATTCAACCTATGCTCGCCTCGCCAGGCGCCGCTCTTC
CAAATATCTTTCCTGAGGTGACTTCAATTGAGTGGAAGCTTGATGGAGCTCGCATTCAAGTGCATCGCTTAAA
TGATGAGGTAGCCATTTTCACTCGCAATCTCAATAATATTACGGAGCGCATGACAGAGGTGGTGGAAGCAGC
CTTATCATTTCGCGCTAAGGCCTTTGTGCTTGATGGCGAGGCCATGGCATTACGCGATGATGGGACACCACA
ACCCTTTCAAGAGACTATGTCGCGCTTTGGCACTGAGGAGCGCGTGTTTGCAGAGGTGCCTGTGCTTGGATTC
TTCTTTGATCTCTTACATCTCGATGGCGTAGATCTCATTGATGAGCCTCTTCATCGCCGCCAAGAGCTCTTAGA
TGAGTTGGTGCCCTTAGCTCAACTCATTCCCCGCGTCCTTACATCCAATGCCGATGAGGCCGCCACGTTTGCC
CAAGGAGCATTGGCAGCCGGCCATGAGGGTGTGATGTTGAAGGATCCCGAGTCCCGCTATGAGGCCGGACG
TCGCGGGAAGAGTTGGCTCAAGGTGAAGCCTGTACATACTTATGATCTTGTCGTATTAGCAGCCGAGTGGGG
ACATGGACGCCGCTCGGGCTATCTCTCAAATATTCATTTAGGCGCTCGCGATCCCGCTACTGGCGGGTTTGTG
ATGGTCGGAAAGACGTTTAAGGGAATGACAGATGAGATGCTTGGCTGGCAAACGGAGCATTTTCCTACGTTG
GAGACACATCGCGATCGCTGGGCAGTGTATCTCCGCCCAGAGCAAGTGGTGGAGATTGCCTTAGATGGAGTG
CAAGCGTCCACACGCTATCCCGGCGGAGTGGCCTTACGCTTTGCCCGCGTGAAGCGCTATCGCTTTGATAAG
GCCCCTGCAGAGGCAGATACGATTCAAACGTTACAAGCCTTACTTCCAGGCCATACACCCAGCTGAACCCAG
CTTTCTTGTACAAAGTTGGCATTATAAGAAAGCATTGCTTATCAATTTGTTGCAACGAACAGGTCACTATCAG
TCAAAATAAAATCATTATTTG

DV-Lig5 translated protein product (59.3 kDa)

MSYYHHHHHHLESTSLYKKAGYHHHHHHENLYFQG**L**NELVSTSAEVTKTSSRKAKAALLAKLLRQLETGEVAT
AVGLLIGQPRQRRLGVGFGSVFNLDIEPATHPTLTIDEVDAAFSSLQRAGGAGSQKSRNDLLTGLMKRATSSEQAF
LRQVLTGEVRQGALGGVLTEAVALGFEVPPEVVRRASMLRGDLGQVAEVAAVGGVVGLESIGLRVLTPIQPMLA
SPGAALPNIFPEVTSIEWKLDGARIQVHRLNDEVAIFTRNLNNITERMTEVVEAALSFRAKAFVLDGEAMALRDDG
TPQPFQETMSRFGTEERVFAEVPVLGFFFDLLHLDGVDLIDEPLHRRQELLDELVPLAQLIPRVLTSNADEAATFAQ
GALAAGHEGVMLKDPESRYEAGRRGKSWLKVKPVHTYDLVVLAAEWGHGRRSGYLSNIHLGARDPATGGFVM
VGKTFKGMTDEMLGWQTEHFPTLETHRDRWAVYLRPEQVVEIALDGVQASTRYPGGVALRFARVKRYRFDKAP
AEADTIQTLQALLPGHTPS*

DV-Nuc3 gene sequence in pDONR221 plasmid (1,145 bp, excluding attL sites)

CAAATAATGATTTTATTTTGACTGATAGTGACCTGTTCGTTGCAACACATTGATGAGCAATGCTTTTTTATAAT
GCCAACTTTGTACAAAAAAGCAGGCTTAGAGAATTTATATTTTCAAGGC**AAA**CTGGTTCGTGTTGCGCTGAA
CAAAGAAAACGGTCGTTTCGCGCAGATCATCTTCGAAAACGCGTCTAAATCTCAGATCGCGGTTCTGGCGGA
CGTTTCTGAAATCGCGGCGCCGGAAAACCTGCTGACCGCGGCGATCCTGTGGTTCGCGCGTCTGCAGTACCG
TAAAAAAAACCCGATCAACGCGGTTTGGATCCTGGCGGAAAAAAAACTGTGCAAAAACCTGCAGAAACTGC
ACGCGCTGCTGCTGGAAAACTGGCAGCGTAACATCCTGATCAAAGAAGTTTGCCGTGACGTTAAAACCCAGA
AACGTGAAATCACCGAAGCGCCGGCGATCACCTTCGGTAACCTGTGGCGTGAAAAACCGCCGGCGATCTCTC
TGGCGAACTCTGAAATGTCTCGTACCGCGAACGAAATCGTTAAACTGGCGCCGGAAAAAATCGACATCATCT
ACACCCGTCACGGTGAAACCCTGCGTTTCTTCGGTCTGCCGTTCGCGCGTGTTCGTAAAATCGGTGACGCGGA
AAAAGCGTGGTTCGGTACCGAACGTGAAAAACGTATCCTGAACGAAAACACCCGTGCGGAATTCTTCGCGCT
GCTGGAAAACCTGGAAACCTACCGTCGTTTCGACTCTGCGAACAAACGTCACGACTTCTCTCGTCTGGCGCC
GGAAGCGTGGCTGGAAGCGATCCTGCGTCGTAACATCAAACTGCTGGACGGTAACCTGATCCTGTCTCCGAT
CTACAACCAGTTCCGTGCGGCGAACGACAAAATCGACCTGCTGGCGCTGCGTACCGACGGTCGTCTGGTTGT
TATCGAACTGAAAGTTGAACCGGACCGTGAAATGATCTTCCAGGCGGCGGACTACTGGCGTAAAATCGAACT
GCAGCGTCGTTCTCGTAACCTGCGTCGTGCGAAAATCTTCGGTGACCTGGAAATCGCGGACGTTCCGACCCT
GGTTTACCTGGTTGCGCCGACCCTGTCTTTCCACCGTGACTTCACCTTCCTGTCTAAAACCGTTTCTCCGCAGA
TCGAAATCTACCGTTTCGACCTGAACGAAAACTGGCGTGAAAACCTGAAAGTTATGAAAGTTGGTGAAGTTC
GTTCTGAAAAATAATACCCAGCTTTCTTGTACAAAGTTGGCATTATAAGAAAGCATTGCTTATCAATTTGTTG
CAACGAACAGGTCACTATCAGTCAAAATAAAATCATTATTTG

DV-Nuc3 translated protein product (60.3 kDa)

MGSSHHHHHHENLYFQG**G**GSVDKISDEEDEMNDENAFSEIIESLASGNEWLLIHSSGNAFALKRDEIEITFERGRIIL
GFLDEKGFQIWRVAGRKIEREKLTLDLTRNFGRREREKINLVPRLSAKESGAAVELARLEKANQLAGLIVTENPKSK
LVRVALNKENGRFAQIIFENASKSQIAVLADVSEIAAPENLLTAAILWFARLQYRKKNPINAVWILAEKKLCKNLQ
KLHALLLENWQRNILIKEVCRDVKTQKREITEAPAITFGNLWREKPPAISLANSEMSRTANEIVKLAPEKIDIIYTRH
GETLRFFGLPFARVRKIGDAEKAWFGTEREKRILNENTRAEFFALLENLETYRRFDSANKRHDFSRLAPEAWLEAI
LRRNIKLLDGNLILSPIYNQFRAANDKIDLLALRTDGRLVVIELKVEPDREMIFQAADYWRKIELQRRSRNLRRAKI
FGDLEIADVPTLVYLVAPTLSFHRDFTFLSKTVSPQIEIYRFDLNENWRENLKVMKVGEVRSEK*

DV-1-1-Ligase full length gene sequence in pDONR221 plasmid (2,916 bp, excluding attL sites)

CAAATAATGATTTTATTTTGACTGATAGTGACCTGTTCGTTGCAACACATTGATGAGCAATGCTTTTTTATAAT
GCCAACTTTGTACAAAAAAGCAGGCTTCCATCATCATCATCACCACGAGAACCTGTATTTTCAGGGT**CAT**CG
TAGCGAACTGGGTAAACTGAATAGCGCAATTGATCGTCCGACACGTTGTGATATTCATGCACGTAGCAATAG
CCGTAGCCGTCTGGAACTGAGCACCATTGCATTTCCGCTGCTGCCGATTCGTTTTCATCGTGGTGTTGAACTG
CCGGAACAGAGCCTGTGGCTGGATCCGCATGATCCGAAACCGTTTGCATTTGTTAGCCATGCACATAGCGAT
CATCTGGGCACCCATGCAGAAATTATCACCAGCAAAGGCACCAGCGCACTGATGCGTGAACGTCTGCCTGGT
GAACGTATTGAACATGTGCTGGAATTTGATAGTCCGGCAACCATTCGTGGTCTGAATGTTACCCTGCTGCCTG
CAGGTCATGTTTTTGGTAGCGCACAGCTGTTTCTGCAGACCGAAAATGAAAGCCTGCTGTATACCGGTGATTT
TAAACTGCGTCGCGGTCTGAGCGCAGAACCGACCGGTTGGCGTCATGCAGATACCCTGATTATGGAAACCAC
CTATGGTCTGCCGAAATATGCAATGCCTCCGACCGAAGAAACCCTGGCACGTATGATTGCATTTTGTCAAGA
GGCACAAGAAGAAGGCGCAGTTCCGGTTCTGCTGGGTTATAGCCTGGGTAAAGCACAAGAAATTCTGTGTGC
ACTGGTTCAGGCAGGTCTGACCCCGATGCTGCATGGTGCAGTTTGGAATATGACCGAAGTTTATCGTAAATT
GCGTCCGGATTTTCCGTGTGGTTATGAACGTTATGCAGCCGGTGAAACCGCAGGTAAAGTTCTGGTTTGTCCT
CCGAGCGCAATTCGTATGAAAATGGTTACCCAGATTAAACAGCGTCGTGTTGCAGTTCTGACCGGCTGGGCA
TTAGATCCGGGTGCAATTTATCGTTATCAGTGTGATGCAGCCTTTCCGCTGACCGATCATGCCGATTATCCGG
ATCTGCTGCGCTATGTGGAACTGGTGCAGCCGAAACGTGTTCTGACCCTGCATGGTTTTGCAGCAGAATTTGC
ACGTGATCTGCGCGAACGTGGTGTGGAAGCATGGGCACTGAGCGAAGAAAATCAGCTGGAATTAACACTGG
CACGTCCGACCGCACGTCAAGAAAAACCGCAGCTGACCCGTACACCGGATAGCGGTGATCCGGCACCGCAG
CCTCCGGTACAGGCAGCACCTAGCGCACCGGGTGATTTTGCACGTTTTGCCGCAATTGGTGAAGAAATTGCA
AAAACCACCAGTAAACTGGCAAAAATTGCACTGCTGAGCGATTATCTGCGTAGCCTGGCAGCAGATGAACTG
CCTAGCGCAGCCACCTTTCTGACAGGTCGTGCGTTTCCGCAGAATGATGGTCGTGTGCTGCAGACAGGTTGG
AGCGTGATTCATCGTGCCCTGCTGGCAGCAAGCGGTGTTGGTGAAGCACGTCTGCGTGAAGCAGGTCGTACC
TATGCAGATGCAGGTAAAACCGCATTTGAAGTGCTGCTGGGTCGTACCACACCGGCACCGTTTAGCCTGATT
GATGCCCGTGATTTTTTTGCGGCACTGGCAGCCGCACGTGGTCCGCTGCGTAAAACCGAACTGCTGACCCAG
CGTCTGGCAACCCTGACACCGATTGAAGCAAGCTATGTGGTTAAAATTCTGACCAGCGATCTGCGTATTGGT
CTGAAAGAAGGTCTGGTTGAAGAAGCAATTGCAGCAGCCTTTGAAGCACCGGCAGATGATGTTCGTGAAGC
AAATATGCTGGTTGGTGATCTGGGTGAAGTTGCAGCCCTGGCAGCCCGTAAAGCACTGGAAGAAGCGACCCT
GCACCTGTTTCGTCCGATTAAATGTATGCTGGCAAGTCCGGAACCGACCAGCGAAGCAATTTGGAGCCGTAT
TGAAAATACCGATCATCATAGCCCGATTACAGATCATAGCAGCAGTCCGCATTGGGCTGAAGATAAATTTGA
TGGTATTCGTGCACAGCTGCATCTGGCAGATGGTCGCGTTGAAATCTTTACCCGTGATCTGAAATGTGTTACC
GGTCAGTTTGCAGATCTGGCAGGCAAAGCACGTGCATGGCCTGGTCGTGCCATTTTTGATGGTGAAATTCTG
GCATTTGCCGAGGGTAAAAAACTGAGCTTTTTTGATTTACAGAAACGCCTGGGTCGCAAAACCGAAGATGAC
CTGTTTTTAGGTGGTGGTAGTGATGTTCCGGTTATTTTTCAGGCATTTGATCTGCTGTGGTTAGATGGTGAATC
ACTGCTGAAACAGCCGCTGCGTGATCGTCGCGATCTGCTGGAAATGCTGAGTCTGCCAGAACCTTTTGCACT
GGCCGAACGTTATCCGATTTGTAGCGCAGATGAAATTGAAGCAGCATTTCGTGCAGCACGTGCCCGTCGTAA
TGAAGGTCTGATTATCAAAGATGCAGAAAGCGCATATACACCGGGTCGTCGTGGCCTGAGCTGGCTGAAACT
GAAAAAAGATTTTGCAACCCTGGATGTTGTTGTTGTGGCAGCCGAACAAGGTCATGGTAAACGTAGCCATGT
TCTGAGTGATTATACCTTTGCAGTTCGTGATGAAGAAACAGGCGCTCTGCTGACCATTGGTAAAGCATATAG
CGGTCTGACCGATGATGAAATCGAAGATCTGACCGAACATTTTACCCGTACCACCATTGCGCAGCATGGTCA
TTATCGTGAAGTTACACCGGAAATTGTTCTGGAAATTGCCTTTGATAGCCTGCAGCCGAGCACACGTCATGCA
AGCGGTCTGGCAATGCGTTTTCCGCGTATTAAAGCAATTCGTCGTGATAAAACTCCGGCAGAAATTGATACC
CTGGCATATGCACGTTCACTGGTTGTTAGCGAATTTGGCAAAGCAGATACCGAATAAGACCCAGCTTTCTTGT
ACAAAGTTGGCATTATAAGAAAGCATTGCTTATCAATTTGTTGCAACGAACAGGTCACTATCAGTCAAAATA
AAATCATTATTTG

DV-1-1-Ligase full length translated protein product (110 kDa)

MSYYHHHHHHLESTSLYKKAGFHHHHHHENLYFQG**H**RSELGKLNSAIDRPTRCDIHARSNSRSRLELSTIAFPLLP
IRFHRGVELPEQSLWLDPHDPKPFAFVSHAHSDHLGTHAEIITSKGTSALMRERLPGERIEHVLEFDSPATIRGLNVT
LLPAGHVFGSAQLFLQTENESLLYTGDFKLRRGLSAEPTGWRHADTLIMETTYGLPKYAMPPTEETLARMIAFCQ
EAQEEGAVPVLLGYSLGKAQEILCALVQAGLTPMLHGAVWNMTEVYRKLRPDFPCGYERYAAGETAGKVLVCP
PSAIRMKMVTQIKQRRVAVLTGWALDPGAIYRYQCDAAFPLTDHADYPDLLRYVELVQPKRVLTLHGFAAEFAR
DLRERGVEAWALSEENQLELTLARPTARQEKPQLTRTPDSGDPAPQPPVQAAPSAPGDFARFAAIGEEIAKTTSKL
AKIALLSDYLRSLAADELPSAATFLTGRAFPQNDGRVLQTGWSVIHRALLAASGVGEARLREAGRTYADAGKTAF
EVLLGRTTPAPFSLIDARDFFAALAAARGPLRKTELLTQRLATLTPIEASYVVKILTSDLRIGLKEGLVEEAIAAAFE
APADDVREANMLVGDLGEVAALAARKALEEATLHLFRPIKCMLASPEPTSEAIWSRIENTDHHSPITDHSSSPHWA
EDKFDGIRAQLHLADGRVEIFTRDLKCVTGQFADLAGKARAWPGRAIFDGEILAFAEGKKLSFFDLQKRLGRKTE
DDLFLGGGSDVPVIFQAFDLLWLDGESLLKQPLRDRRDLLEMLSLPEPFALAERYPICSADEIEAAFRAARARRNE
GLIIKDAESAYTPGRRGLSWLKLKKDFATLDVVVVAAEQGHGKRSHVLSDYTFAVRDEETGALLTIGKAYSGLTD
DEIEDLTEHFTRTTIAQHGHYREVTPEIVLEIAFDSLQPSTRHASGLAMRFPRIKAIRRDKTPAEIDTLAYARSLVVS
EFGKADTE

DV-1-1-Ligase domain gene sequence in pDONR221 plasmid (1,704 bp, excluding attL sites)

CAAATAATGATTTTATTTTGACTGATAGTGACCTGTTCGTTGCAACACATTGATGAGCAATGCTTTTTTATAAT
GCCAACTTTGTACAAAAAAGCAGGCTTAGAGAACCTGTATTTTCAGGGT**GAT**TTTGCACGTTTTGCCGCAATT
GGTGAAGAAATTGCAAAAACCACCAGTAAACTGGCAAAAATTGCACTGCTGAGCGATTATCTGCGTAGCCTG
GCAGCAGATGAACTGCCTAGCGCAGCCACCTTTCTGACAGGTCGTGCGTTTCCGCAGAATGATGGTCGTGTG
CTGCAGACAGGTTGGAGCGTGATTCATCGTGCCCTGCTGGCAGCAAGCGGTGTTGGTGAAGCACGTCTGCGT
GAAGCAGGTCGTACCTATGCAGATGCAGGTAAAACCGCATTTGAAGTGCTGCTGGGTCGTACCACACCGGCA
CCGTTTAGCCTGATTGATGCCCGTGATTTTTTTGCGGCACTGGCAGCCGCACGTGGTCCGCTGCGTAAAACCG
AACTGCTGACCCAGCGTCTGGCAACCCTGACACCGATTGAAGCAAGCTATGTGGTTAAAATTCTGACCAGCG
ATCTGCGTATTGGTCTGAAAGAAGGTCTGGTTGAAGAAGCAATTGCAGCAGCCTTTGAAGCACCGGCAGATG
ATGTTCGTGAAGCAAATATGCTGGTTGGTGATCTGGGTGAAGTTGCAGCCCTGGCAGCCCGTAAAGCACTGG
AAGAAGCGACCCTGCACCTGTTTCGTCCGATTAAATGTATGCTGGCAAGTCCGGAACCGACCAGCGAAGCAA
TTTGGAGCCGTATTGAAAATACCGATCATCATAGCCCGATTACAGATCATAGCAGCAGTCCGCATTGGGCTG
AAGATAAATTTGATGGTATTCGTGCACAGCTGCATCTGGCAGATGGTCGCGTTGAAATCTTTACCCGTGATCT
GAAATGTGTTACCGGTCAGTTTGCAGATCTGGCAGGCAAAGCACGTGCATGGCCTGGTCGTGCCATTTTTGAT
GGTGAAATTCTGGCATTTGCCGAGGGTAAAAAACTGAGCTTTTTTGATTTACAGAAACGCCTGGGTCGCAAA
ACCGAAGATGACCTGTTTTTAGGTGGTGGTAGTGATGTTCCGGTTATTTTTCAGGCATTTGATCTGCTGTGGTT
AGATGGTGAATCACTGCTGAAACAGCCGCTGCGTGATCGTCGCGATCTGCTGGAAATGCTGAGTCTGCCAGA
ACCTTTTGCACTGGCCGAACGTTATCCGATTTGTAGCGCAGATGAAATTGAAGCAGCATTTCGTGCAGCACGT
GCCCGTCGTAATGAAGGTCTGATTATCAAAGATGCAGAAAGCGCATATACACCGGGTCGTCGTGGCCTGAGC
TGGCTGAAACTGAAAAAAGATTTTGCAACCCTGGATGTTGTTGTTGTGGCAGCCGAACAAGGTCATGGTAAA
CGTAGCCATGTTCTGAGTGATTATACCTTTGCAGTTCGTGATGAAGAAACAGGCGCTCTGCTGACCATTGGTA
AAGCATATAGCGGTCTGACCGATGATGAAATCGAAGATCTGACCGAACATTTTACCCGTACCACCATTGCGC
AGCATGGTCATTATCGTGAAGTTACACCGGAAATTGTTCTGGAAATTGCCTTTGATAGCCTGCAGCCGAGCA
CACGTCATGCAAGCGGTCTGGCAATGCGTTTTCCGCGTATTAAAGCAATTCGTCGTGATAAAACTCCGGCAG
AAATTGATACCCTGGCATATGCACGTTCACTGGTTGTTAGCGAATTTGGCAAAGCAGATACCGAATAAGACC
CAGCTTTCTTGTACAAAGTTGGCATTATAAGAAAGCATTGCTTATCAATTTGTTGCAACGAACAGGTCACTAT
CAGTCAAAATAAAATCATTATTTG

DV-1-1-Ligase domain translated protein product (65 kDa)

MSYYHHHHHHLESTSLYKKAGLENLYFQG**D**FARFAAIGEEIAKTTSKLAKIALLSDYLRSLAADELPSAATFLTGR
AFPQNDGRVLQTGWSVIHRALLAASGVGEARLREAGRTYADAGKTAFEVLLGRTTPAPFSLIDARDFFAALAAAR
GPLRKTELLTQRLATLTPIEASYVVKILTSDLRIGLKEGLVEEAIAAAFEAPADDVREANMLVGDLGEVAALAARK
ALEEATLHLFRPIKCMLASPEPTSEAIWSRIENTDHHSPITDHSSSPHWAEDKFDGIRAQLHLADGRVEIFTRDLKC
VTGQFADLAGKARAWPGRAIFDGEILAFAEGKKLSFFDLQKRLGRKTEDDLFLGGGSDVPVIFQAFDLLWLDGES
LLKQPLRDRRDLLEMLSLPEPFALAERYPICSADEIEAAFRAARARRNEGLIIKDAESAYTPGRRGLSWLKLKKDF
ATLDVVVVAAEQGHGKRSHVLSDYTFAVRDEETGALLTIGKAYSGLTDDEIEDLTEHFTRTTIAQHGHYREVTPEI
VLEIAFDSLQPSTRHASGLAMRFPRIKAIRRDKTPAEIDTLAYARSLVVSEFGKADTE

DV-1-1-Nuclease domain gene sequence in pDONR221 plasmid (1,218 bp, excluding attL sites)

CAAATAATGATTTTATTTTGACTGATAGTGACCTGTTCGTTGCAACACATTGATGAGCAATGCTTTTTTATAAT
GCCAACTTTGTACAAAAAAGCAGGCTTAGAGAACCTGTATTTTCAGGGT**CAT**CGTAGCGAACTGGGTAAACT
GAATAGCGCAATTGATCGTCCGACACGTTGTGATATTCATGCACGTAGCAATAGCCGTAGCCGTCTGGAACT
GAGCACCATTGCATTTCCGCTGCTGCCGATTCGTTTTCATCGTGGTGTTGAACTGCCGGAACAGAGCCTGTGG
CTGGATCCGCATGATCCGAAACCGTTTGCATTTGTTAGCCATGCACATAGCGATCATCTGGGCACCCATGCAG
AAATTATCACCAGCAAAGGCACCAGCGCACTGATGCGTGAACGTCTGCCTGGTGAACGTATTGAACATGTGC
TGGAATTTGATAGTCCGGCAACCATTCGTGGTCTGAATGTTACCCTGCTGCCTGCAGGTCATGTTTTTGGTAG
CGCACAGCTGTTTCTGCAGACCGAAAATGAAAGCCTGCTGTATACCGGTGATTTTAAACTGCGTCGCGGTCT
GAGCGCAGAACCGACCGGTTGGCGTCATGCAGATACCCTGATTATGGAAACCACCTATGGTCTGCCGAAATA
TGCAATGCCTCCGACCGAAGAAACCCTGGCACGTATGATTGCATTTTGTCAAGAGGCACAAGAAGAAGGCGC
AGTTCCGGTTCTGCTGGGTTATAGCCTGGGTAAAGCACAAGAAATTCTGTGTGCACTGGTTCAGGCAGGTCT
GACCCCGATGCTGCATGGTGCAGTTTGGAATATGACCGAAGTTTATCGTAAATTGCGTCCGGATTTTCCGTGT
GGTTATGAACGTTATGCAGCCGGTGAAACCGCAGGTAAAGTTCTGGTTTGTCCTCCGAGCGCAATTCGTATG
AAAATGGTTACCCAGATTAAACAGCGTCGTGTTGCAGTTCTGACCGGCTGGGCATTAGATCCGGGTGCAATT
TATCGTTATCAGTGTGATGCAGCCTTTCCGCTGACCGATCATGCCGATTATCCGGATCTGCTGCGCTATGTGG
AACTGGTGCAGCCGAAACGTGTTCTGACCCTGCATGGTTTTGCAGCAGAATTTGCACGTGATCTGCGCGAAC
GTGGTGTGGAAGCATGGGCACTGAGCGAAGAAAATCAGCTGGAATTAACACTGGCACGTCCGACCGCACGT
CAAGAAAAACCGCAGCTGACCCGTACACCGGATAGCGGTGATCCGGCACCGCAGCCTCCGGTACAGGCAGC
ACCTAGCGCACCGGGTTAATACCCAGCTTTCTTGTACAAAGTTGGCATTATAAGAAAGCATTGCTTATCAATT
TGTTGCAACGAACAGGTCACTATCAGTCAAAATAAAATCATTATTTG

DV-1-1-Nuclease domain translated protein product (47.7 kDa)

MSYYHHHHHHLESTSLYKKAGLENLYFQG**H**RSELGKLNSAIDRPTRCDIHARSNSRSRLELSTIAFPLLPIRFHRGV
ELPEQSLWLDPHDPKPFAFVSHAHSDHLGTHAEIITSKGTSALMRERLPGERIEHVLEFDSPATIRGLNVTLLPAGH
VFGSAQLFLQTENESLLYTGDFKLRRGLSAEPTGWRHADTLIMETTYGLPKYAMPPTEETLARMIAFCQEAQEEG
AVPVLLGYSLGKAQEILCALVQAGLTPMLHGAVWNMTEVYRKLRPDFPCGYERYAAGETAGKVLVCPPSAIRM
KMVTQIKQRRVAVLTGWALDPGAIYRYQCDAAFPLTDHADYPDLLRYVELVQPKRVLTLHGFAAEFARDLRER
GVEAWALSEENQLELTLARPTARQEKPQLTRTPDSGDPAPQPPVQAAPSAPG*

A.11 DNA sequences and duplex design for use in ligase and nuclease activity assays

A.10.1 DNA substrate oligonucleotide sequences

Table A.10.1. DNA substrate oligonucleotide sequences used to construct assay substrates in **A.9.2**. Abbreviations: 5’ 6-carboxyfluorescien (5’FAM), 8-Oxo-deoxyguanosine (8OxodG), abasic tetrohydrofuran (dSpacer).

Name	Sequence (5’ to 3’)	Modifications
NL1	AGGCCATGGCTGATATCGCA	5' FAM
NL1	AGGCCATGGCTGATATCGCA	No label
NL2	TAGGCATTCGAGCTCCGTCG	5’ phosphate
NL3	CGACGGAGCTCGAATGCCTATGCGATATCAGCCATGGCCT	
NL5	AGGCCATGGCTGATATCGCATAGGCATTCGAGCTCCGTCG	5' FAM
NL6	CGACGGAGCTCGAATGCCTA	
NL7	TGCGATATCAGCCATGGCCT	
NL8	ATATCAGCCATGGCCT	
NL9	CGACGGAGCTCGAATGCCTATGCG	
NL10	CGACGGAGCTCGAATGCCTACGCGATATCAGCCATGGCCT	
NL11	CGACGGAGCTCGAATGCctAGTGCGATATCAGCCATGGCCT	
MD5	AGGCCATGGCTGATATC**X**CATAGGCATTCGAGCTCCGTCG	5' FAM, X=8-Oxo-dG
MD6	AGGCCATGGCTGATATCGCA**X**AGGCATTCGAGCTCCGTCG	5' FAM, X=dSpacer
MD9	AGGCCATGGCTGATATCGCA**U**AGGCATTCGAGCTCCGTCG	5' FAM
MD10	CGACGGAGCTCGAATGCCTGTGCGATATCAGCCATGGCCT	
HJ5	ATCATAGCTAACATGACTAGTGCGATATCAGCCATGGCCT	
HJ6	CTAGTCATGTTAGCTATGAT	
UB14	TGGCTGATATC**PPP**	5' FAM
UB15	AGCTCGAATGCCTA**ZZZ**GATATCAGCCA	
UB16	**ZZZ**GCATTCGAGCT	5’ phosphate
UB17	AGCTCGAATGC**PPP**TGCGATATCAGCCA	
UB18	AGCTCGAATGCPPP**ZZZ**GATATCAGCCA	

UBSB2	**B**AGGCATTCGAGCT	5’ phosphate
UBSB4	AGCTCGAATGCCT**S**TGCGATATCAGCCA	
UBSB14	TGGCTGATAT**SBB**C	5' FAM
UBSB15	AGCTCGAATGCCTAG**SSB**ATATCAGCCA	
UBSB16	**BSS**GCATTCGAGCT	5’ phosphate
UBSB17	AGCTCGAATGC**BBS**TGCGATATCAGCCA	
UBSB18	AGCTCGAATGC**BBSSSB**ATATCAGCCA	

Damages and UBPs are indicated in bold and underlined.

A.10.2 DNA oligomer combinations

Table A.10.2. DNA substrate combinations used to construct various assay substrates. Oligonucleotide sequences are given in Supplementary x.

Assay	Substrate name	Oligonucleotide combinations
DNA damage	8-Oxo guanine	MD5*, $NL3_a$
	Abasic	MD6*, $NL3_a$
	Uracil match	MD9*, $NL3_a$
	Uracil mismatch	MD9*, $MD10_a$
	A/C mismatch	NL5*, $NL10_a$
	T/G mismatch	NL5*, $MD10_a$
Flapped/splayed	Flap 3’	NL15*, $HJ6_a$, $HJ5_a$,
	Flap 5’	NL15*, $NL7_a$, $HJ5_a$
	Splayed	NL15*, $HJ6_a$
Ligase	Nick	NL1*, $NL2_b$, $NL3_b$
	Mismatch	NL1*, $NL2_b$, $NL10_b$
	Gapped	NL1*, $NL3_b$, $NL11_b$
	Blunt	NL1*, $NL6_b$ / $NL2_b$, $NL7_b$
	Overhang	NL1*, $NL8_b$ / $NL2_b$, $NL9_b$
Ligase UBPs	UB_duplex 13	UB14*, $NL2_a$, $UB15_b$
	UB_duplex 14	NL1*, $UB16_a$, $UB17_b$
	UB_duplex 15	UB14*, $UB16_a$, $UB18_b$
	UBSB_duplex 2	NL1*, $UBSB2_a$, $UBSB4_b$
	UBSB_duplex 13	UBSB14*, $NL2_a$, $UBSB15_b$
	UBSB_duplex 14	NL1*, $UBSB16_a$, $UBSB16_b$
	UBSB_duplex 15	UBSB14*, $UBSB16_a$, $UBSB18_b$
Nuclease	Single strand	NL5*
	3’ tail	NL5*, $NL6_a$
	5’ tail	NL1*, $NL3_a$
Controls	Double strand	NL5*, $NL3_a$
	Double strand (20+20)	NL1*, $NL6_a$

*5’ Fam (Labelled oligo)

[a] 5’ phosphate oligo

[b] Long compliment oligo

Appendix B Introduction

B.1 DV-DNA ligase SSNs and Pfam identification

Table B.1. Nodes used in Sequence Similarity Network construction for different Pfam assignments.

Family names	Length threshold (aa)[a]	Total nodes in SSN[b]	# MG nodes[c]	# NCBI nodes (as %)[d]	Sequence identity (%)[e]	# Connected (Fragmentation %)
DNA_ligase_A_M	300	2,998	2,477	521 (17.4)	64	190 (7)
DNA_ligase_A_M + DNA_ligase_A_N	450	1,145	700	445 (38.9)	50	203 (7)
DNA_ligase_A_M + LigD_N +/− PrimaseS	500	1,005	650	355 (35.3)	54	78 (7)

[a]Length cut-off applied to sequences before SSN constructuion; [b]Number of nodes above the length threshold used in final SSN; [c]Number of DV-metagenome nodes in the SSN; [d]Number of NCBI nodes and their percentage in the SSN; [c]Percentage identity threshold for edges retained in the SSN; [d]Number of connected componenes (i.e., clusters or sinlge nodes) and percentage fragmentation calculated as total # nodes/# connected components.

Appendix C Results

C.1 AlphaFold structural predictions

AlphaFold2 was used to predict the 3D protein structures of DV-1-1-Nuc, DV-1-1-Lig, DV-1-1-Lig-Nuc, DV-Nuc3, DV-Lig5 and DV-Lig2. Along with a 3D protein structure output, AlphaFold also generates confidence metrics to evaluate the model performance. Looking at the Predicted Aligned Error (PAE) plot, the shade of green indicates expected distance error in Ångströms. Along the diagonal of the heat map, most elements are expected to be close to 0 Å. The following Figures (**C.1.1-C.1.5**) show the results of a 3D structural predication for each protein and its confidence metrics.

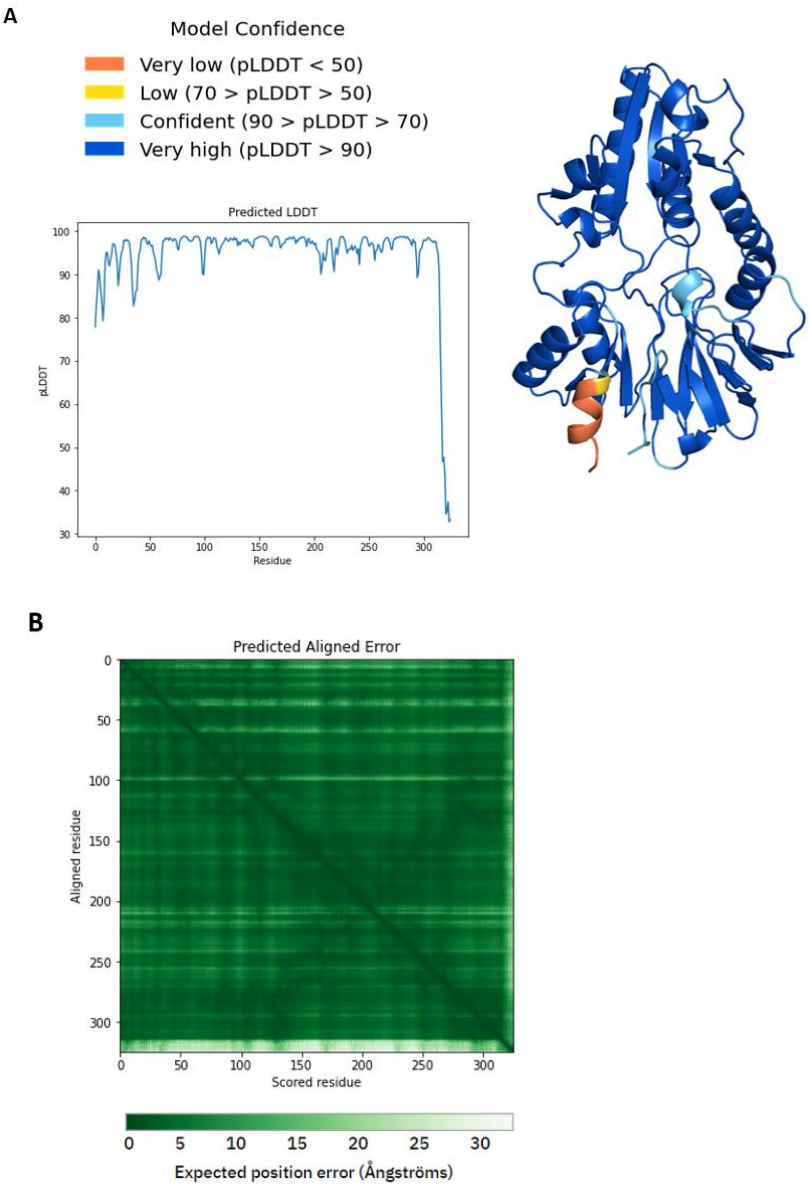


Figure C1.1. Outputs from Alpha fold run for DV-1-1-Nuc showing prediction confidence. **A**) A per-residue confidence metric (pLDDT) plot showing its confidence on a scale from 0-100 and corresponds to the model's predicted score on the LDDT Cα metric. This pLLDDT was used to colour code residues of the model, a key for model confidence is colour coded and protein model was coloured according based on the pLDDT confidence. **B**) Predicted Aligned Error plot used to assess confidence in the domain packing and large-scale topology of the protein. 3D structural prediction using alpha fold. Protein model and prediction profiles were generated by AlphaFold2, from Google Colab, version v2.3.1 (Jumper, Evans et al. 2021). Model was presented using pymol (Schrödinger, 2020).

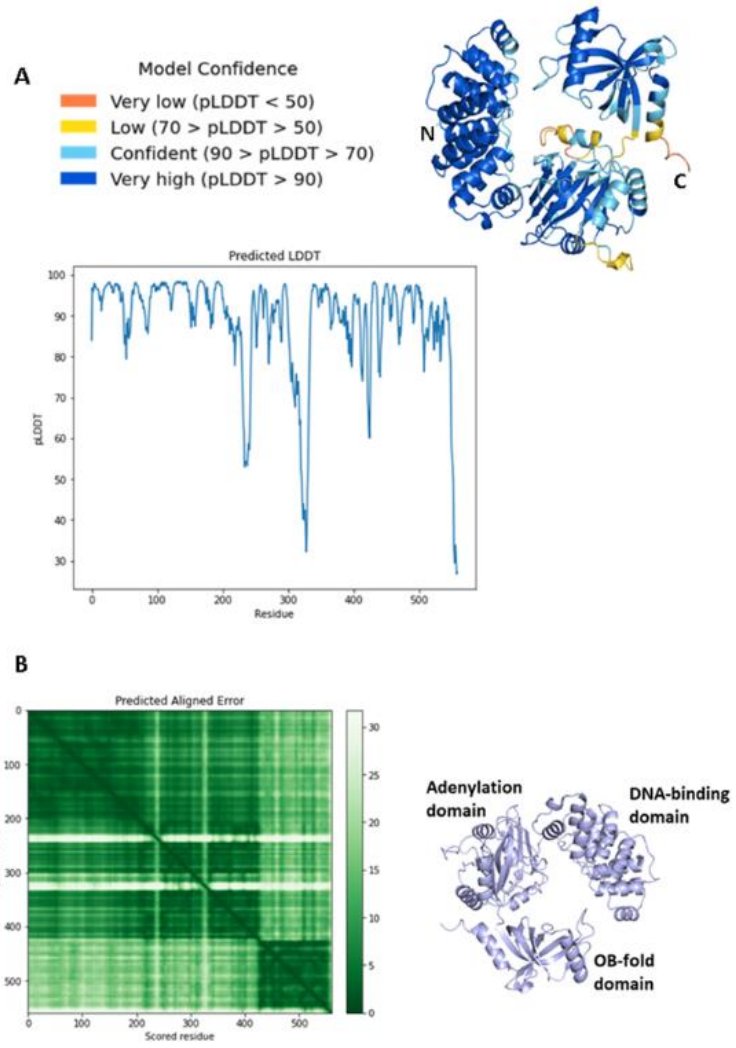


Figure C.1.2. Outputs from Alpha fold run showing prediction confidence for DV-1-1-Lig protein structural model **A**) A per-residue confidence metric (pLDDT) plot showing its confidence on a scale from 0-100 and corresponds to the model's predicted score on the LDDT Cα metric. This pLLDDT was used to colour code residues of the model, a key for model confidence is colour coded and protein model was coloured according based on the pLDDT confidence. **B**) Predicted Aligned Error plot used to assess confidence in the domain packing and large-scale topology of the protein. Here the shade of green indicates expected distance error in Ångströms. The colour at (x,y) corresponds to the expected distance error in residues x's position, when the prediction and true structure are aligned on residue y. Dark green, represents low error, while light green represents high error. Protein model and prediction profiles were generated by AlphaFold2, from Google Colab, version v2.3.1 (Jumper, Evans et al. 2021). Model was presented using pymol (Schrödinger, 2020).

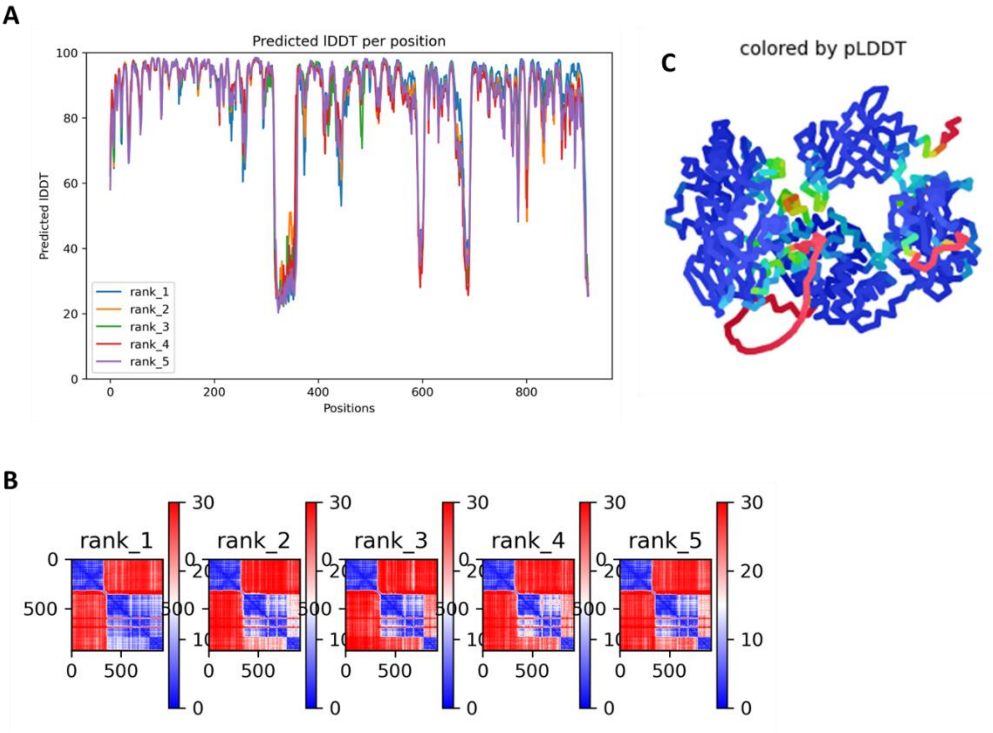


Figure C.1.3. Outputs from Alpha fold run showing prediction confidence for DV-1-1-Lig-Nuc protein structural model **A**) A per-residue confidence metric (pLDDT) plot showing its confidence on a scale from 0-100 and corresponds to the model's predicted score on the LDDT Cα metric. This pLLDDT was used to colour code residues of the model, a key for model confidence is colour coded and protein model was coloured according based on the pLDDT confidence. **B**) Predicted Aligned Error plot used to assess confidence in the domain packing and large-scale topology of the protein. Here the colours blue to red indicates expected distance error in Ångströms. The colour at (x,y) corresponds to the expected distance error in residues x's position, when the prediction and true structure are aligned on residue y. Dark blue, represents low error, while dark red represents high error. **C**) 3D structural prediction of DV-1-1-Lig-Nuc using alpha fold. Protein model and prediction profiles were generated by AlphaFold2, from Google Colab, version v2.3.1 (Jumper, Evans et al. 2021). Model was presented using pymol (Schrödinger, 2020).

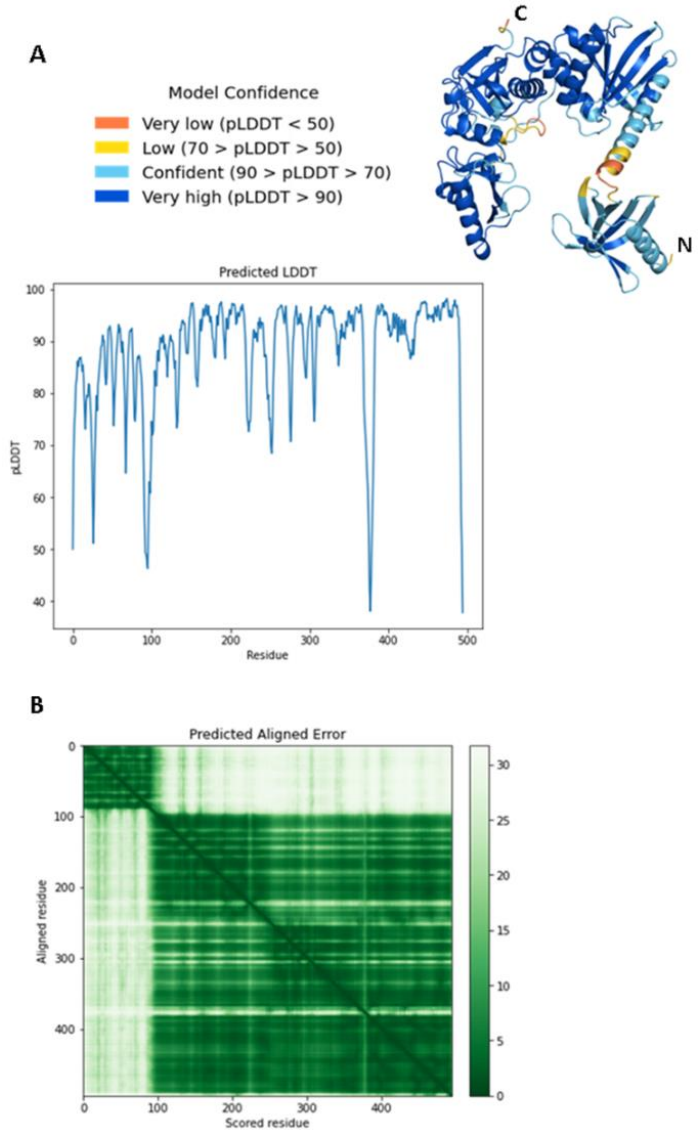


Figure C.1.4. Outputs from Alpha fold run showing prediction confidence for DV-Nuc3 protein structural model **A**) A per-residue confidence metric (pLDDT) plot showing its confidence on a scale from 0-100 and corresponds to the model's predicted score on the LDDT Cα metric. This pLLDDT was used to colour code residues of the model, a key for model confidence is colour coded and protein model was coloured according based on the pLDDT confidence. **B**) Predicted Aligned Error plot used to assess confidence in the domain packing and large-scale topology of the protein. Here the shade of green indicates expected distance error in Ångströms. The colour at (x,y) corresponds to the expected distance error in residues x's position, when the prediction and true structure are aligned on residue y. Dark green, represents low error, while light green represents high error. Protein model and prediction profiles were generated by AlphaFold2, from Google Colab, version v2.3.1 (Jumper, Evans et al. 2021). Model was presented using pymol (Schrödinger, 2020).

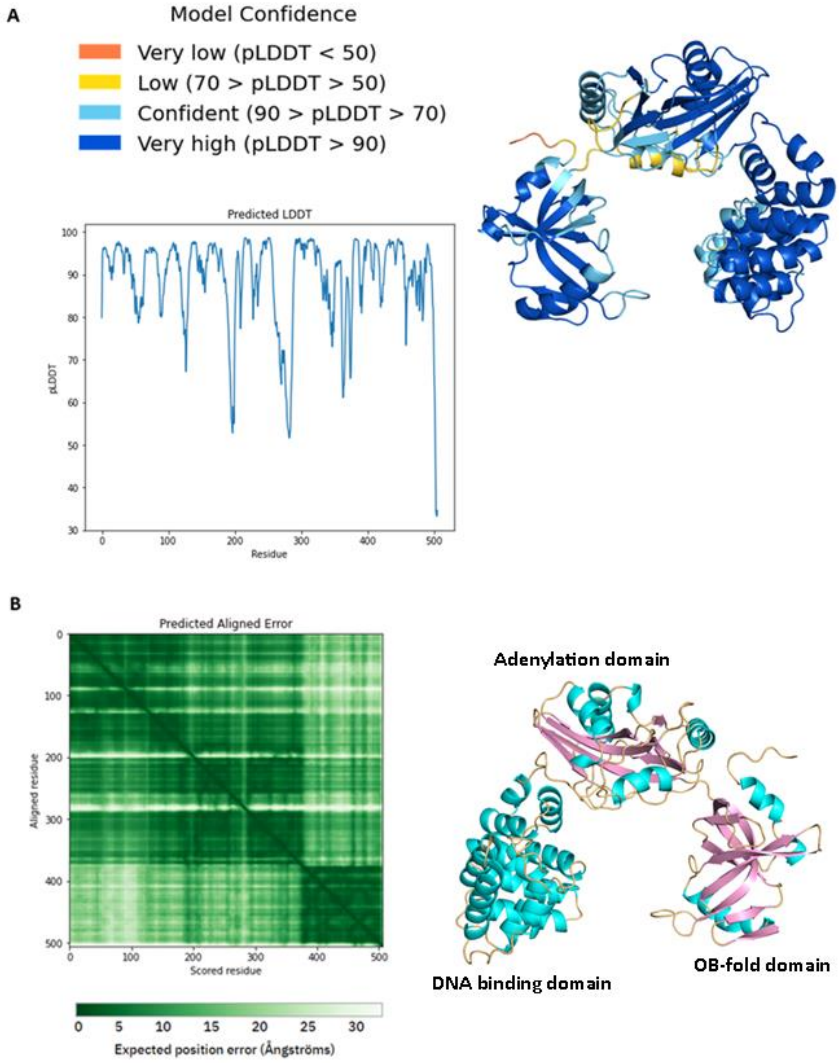


Figure C.1.5. Outputs from Alpha fold run showing prediction confidence for DV-Lig5 protein structural model **A**) A per-residue confidence metric (pLDDT) plot showing its confidence on a scale from 0-100 and corresponds to the model's predicted score on the LDDT Cα metric. This pLLDDT was used to colour code residues of the model, a key for model confidence is colour coded and protein model was coloured according based on the pLDDT confidence. **B**) Predicted Aligned Error plot used to assess confidence in the domain packing and large-scale topology of the protein. 3D structural prediction using alpha fold. Protein model and prediction profiles were generated by AlphaFold2, from Google Colab, version v2.3.1 (Jumper, Evans et al. 2021). Model was presented using pymol (Schrödinger, 2020).

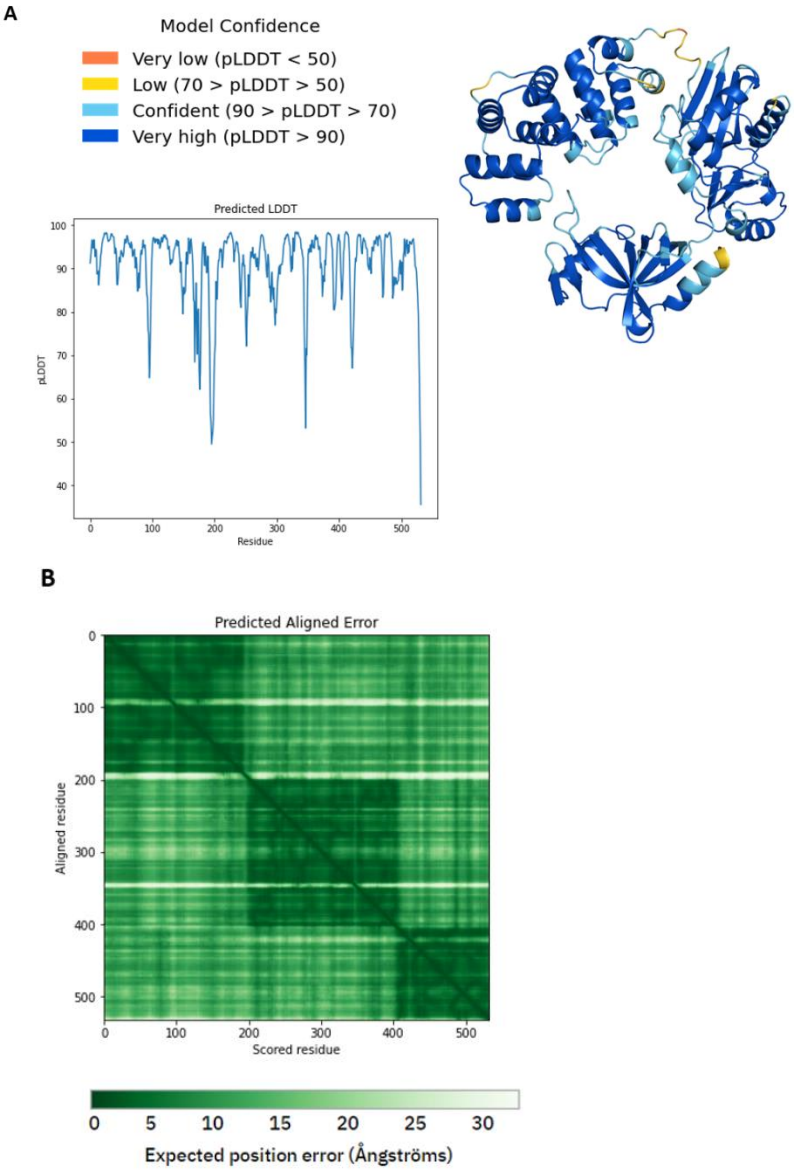


Figure C.1.6. Outputs from Alpha fold run showing prediction confidence for DV-Lig2 protein structural model **A**) A per-residue confidence metric (pLDDT) plot showing its confidence on a scale from 0-100 and corresponds to the model's predicted score on the LDDT Cα metric. This pLLDDT was used to colour code residues of the model, a key for model confidence is colour coded and protein model was coloured according based on the pLDDT confidence. **B**) Predicted Aligned Error plot used to assess confidence in the domain packing and large-scale topology of the protein. The colour at (x,y) corresponds to the expected distance error in residue x's position, when the prediction and true structure aligned on residue y. Dark green corresponds to low error and light green corresponds to high error. Protein model and prediction profiles were generated by AlphaFold2, from Google Colab, version v2.3.1 (Jumper, Evans et al. 2021). Model was presented using pymol (Schrödinger, 2020).

C.2 Protein sequence alignments against homologous proteins

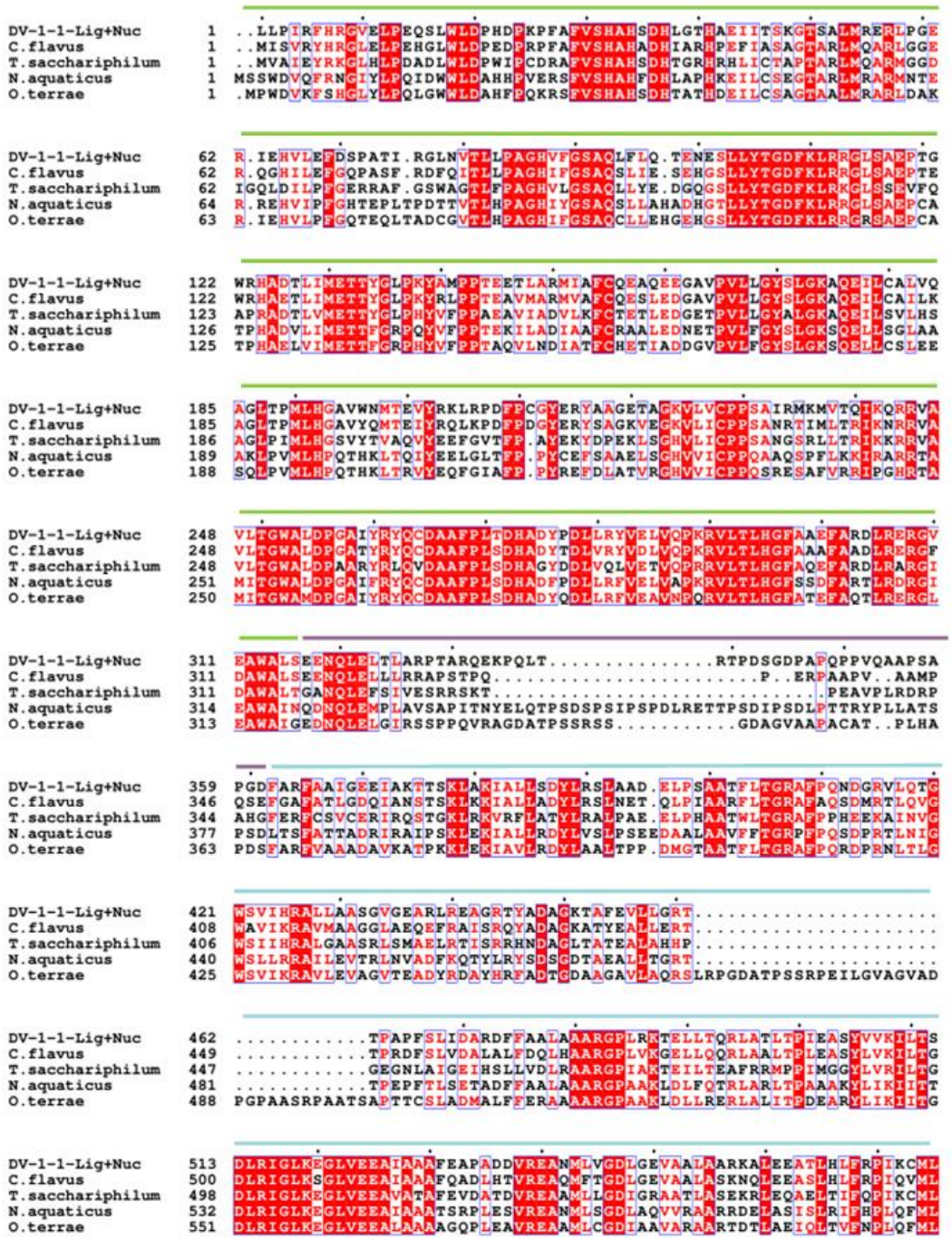


Figure C.2.1. Structural arrangement and sequence alignments of DV-1-1-Lig+Nuc, against homologous Lig+Nuc predicted proteins. A multiple peptide sequence alignment between DV-1-1-Lig+Nuc protein and homologous Lig+Nuc proteins encoded by *C. flavus Ellin428*, *T. saccharihilum*, *N. aquaticus* and *O. terrae*. Highly conserved residues are highlighted in red, with a white text and less conserved residues are colored red, with a white background. Domains are indicated by a coloured line above alignment. Nuclease domain (green), linker region (purple) and ligase domain (blue). Sequence alignment was created using Clustal Omega version 1.2.4. **Figure C.2.1.** was created using ESPript 3 (Robert & Gouet, 2014).

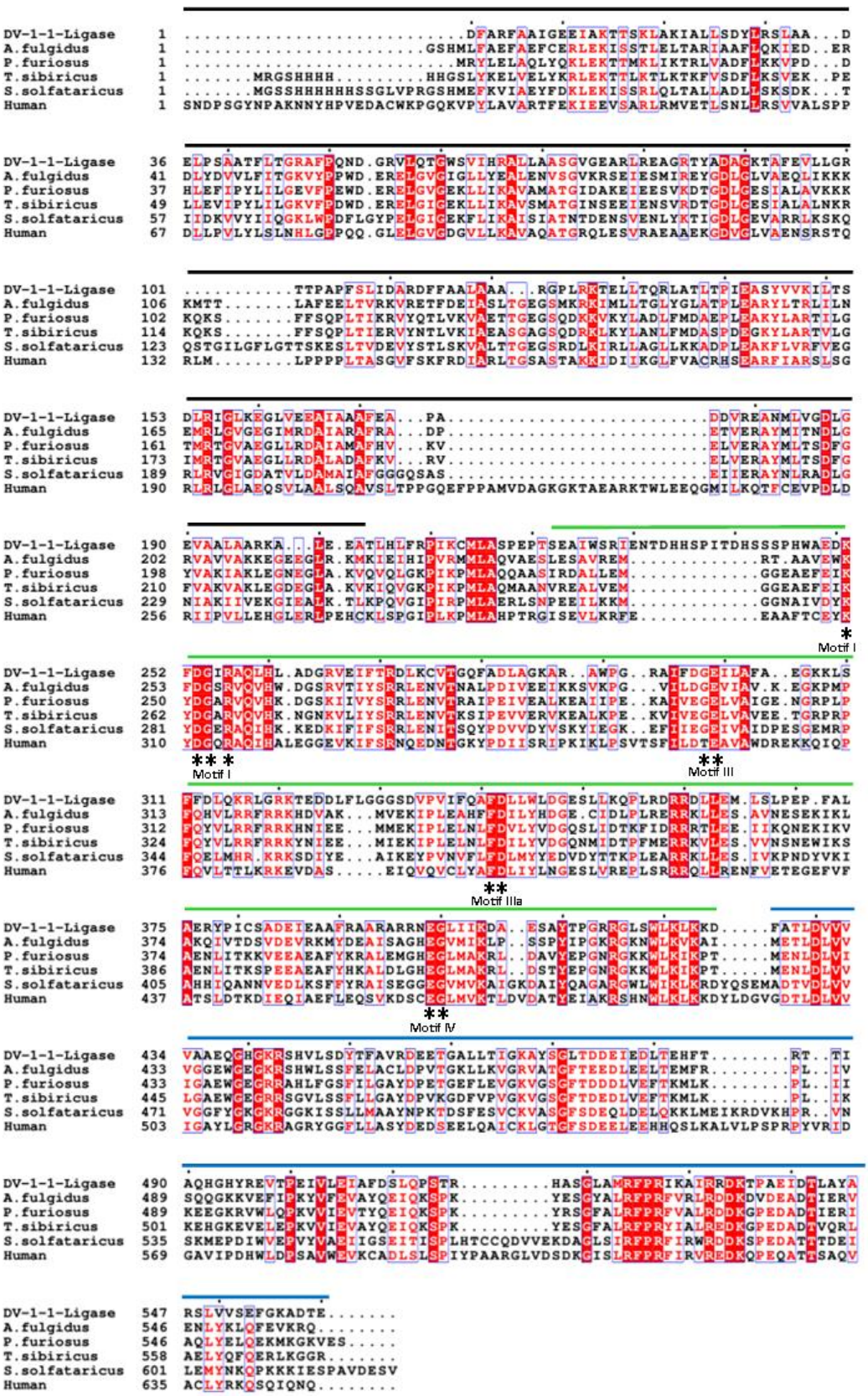
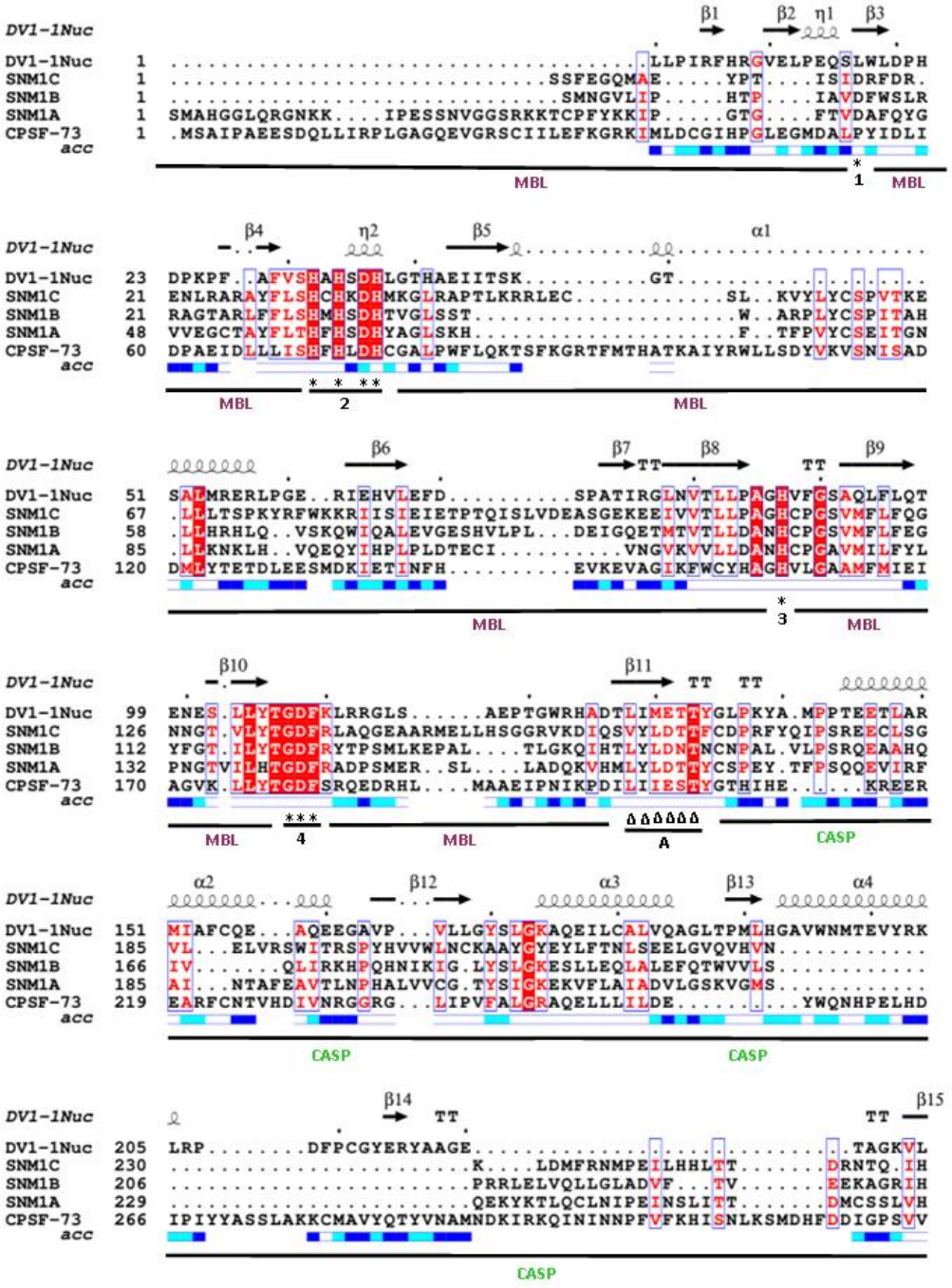


Figure C.2.2. Sequence alignments of DV-1-1-Ligase, against homologous ATP dependent DNA ligases. A multiple peptide sequence alignment between DV-1-1-Ligase domain and homologous ATP DNA ligases encoded by *A. fulgidus*, *P. furiosus*, *T. sibiricus, S. solfataricus* and human. Highly conserved residues are highlighted in red, with a white text and less conserved residues are colored red, with a white background. Nucleotidyl transferase motifs I, III, IIIa, IV, V, and IV represented by asterisks (*). Domain arrangement for

DV-1-1-Ligase domain is represented by colored lines above the sequences; DNA binding domain (black), adenylation domain (green) and OB-fold domain (blue). Sequence alignment was created using Clustal Omega version 1.2.4. **Figure C.2.2**, was created using ESPript 3 (Robert & Gouet, 2014).



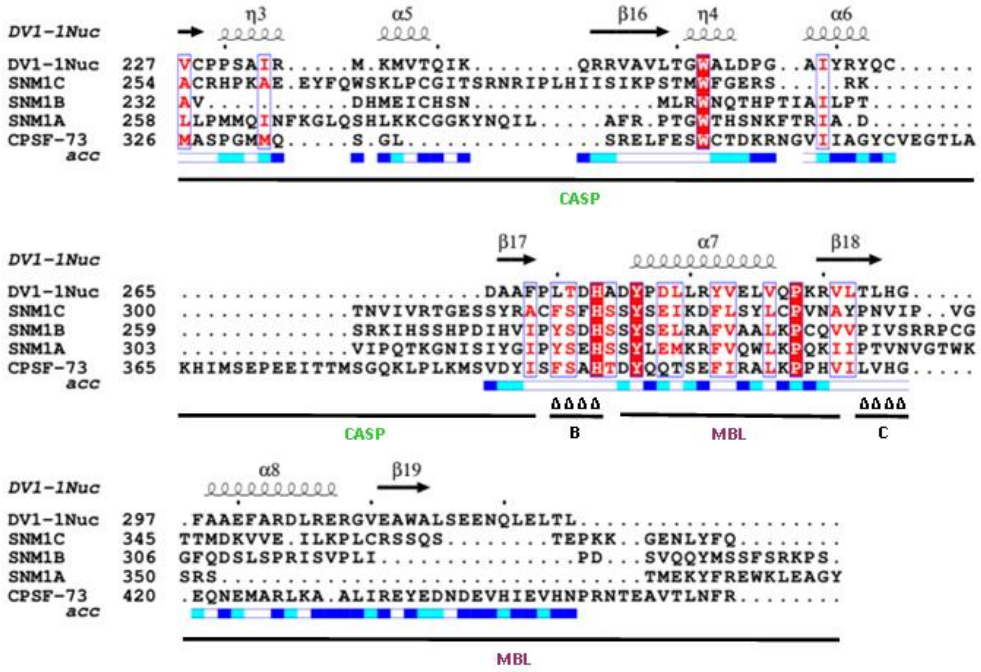


Figure C.2.3. Multiple peptide sequence alignment between DV-1-1-Nuclease domain and homologous MBP-β-CASP nuclease proteins encoded by human (CPSF-73, SNM1C/Artemis, SNM1B/Apollo and SNM1A. Highly conserved residues are highlighted in red, with a white text and less conserved residues are colored red, with a white background. MBL motifs are indicated by asterisks (*) and β-CASP motifs are indicated by triangles. Domain arrangement for DV-1-1-Nuclease domains are represented by colored text above the sequences; MBL domain (purple) and β-CASP (green). Sequence alignment was created using Clustal Omega version 1.2.4. and figure was created using ESPript 3 (Robert & Gouet, 2014).

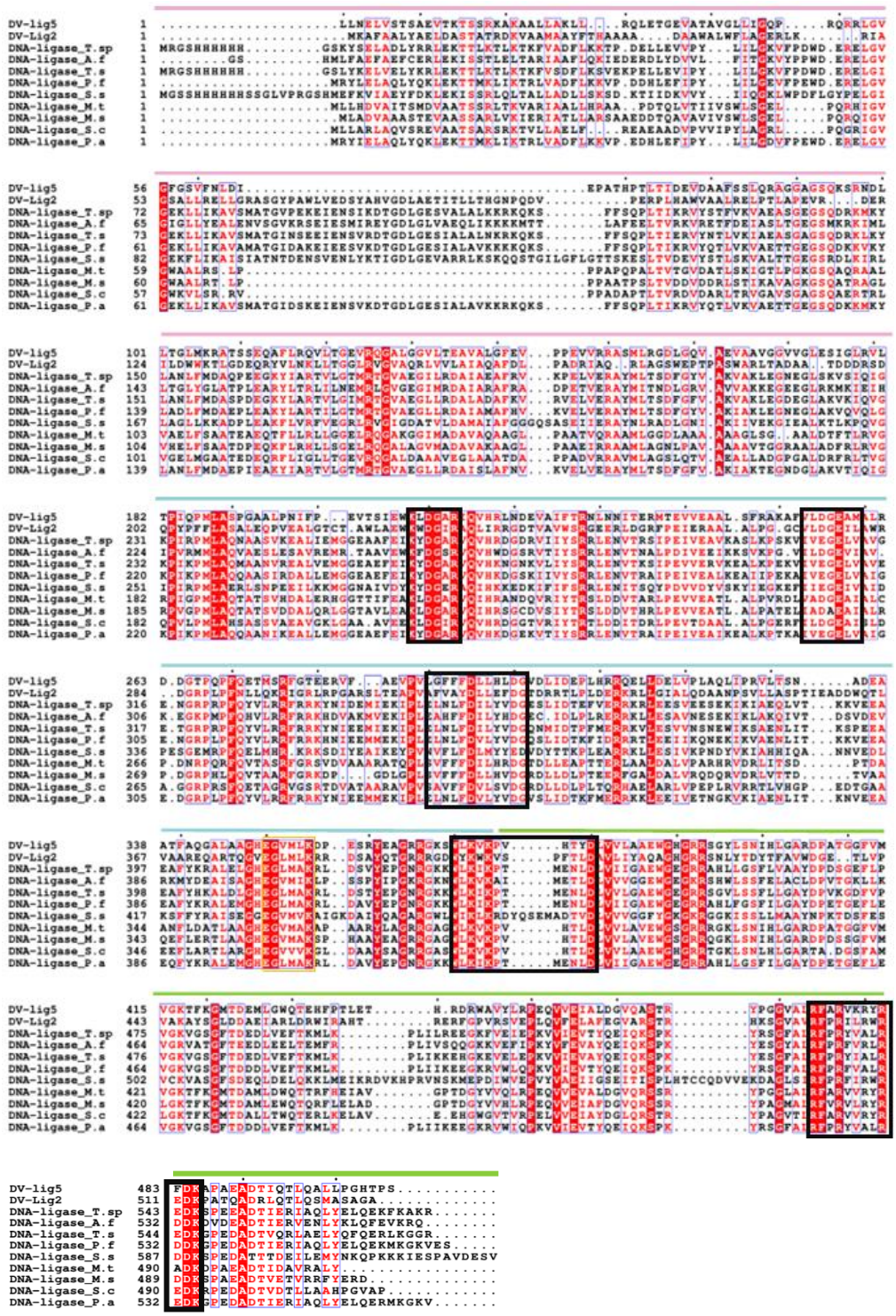


Figure C.2.4. Multiple peptide sequence alignment between DV-Lig, DV-Lig2 and other homologous ATP dependent DNA ligases encoded by T.sp (*Thermococcus sp*. 1519), A.f (*Archaeoglobus fulgidus*), T.s (*Thermococcus sibiricus*), P.f (*Pyrococcus furiosus*), S.s (*Sulfolobus solfataricus*), M.t (*Mycobacterium tuberculosis*), M.s (*Mycobacterium smegmatis*), S.c (*Streptomyces coelicolor*) and P.a (*Pyrococcus abyssi*). Highly conserved residues are highlighted in red, with a white text and less conserved residues are colored red, with a white background. Nucleotidyl transferase motifs I, III, IIIa, IV, V, and IV represented consecutively within black boxes. Sequence alignment was created using Clustal Omega version 1.2.4. Figure was created using ESPript 3 (Robert & Gouet, 2014).

C.3 PCR results of new DV-1-1-Nuc constructs

Following methods from **Section 2.2.2**, three new constructs were designed for DV-1-1-Nuc, to remove residues from the N and C terminus of the protein. Constructs were run on 2 % agarose gels to confirm correct size of PCR product. All constructs were at the expected size (**Figure C.3**).

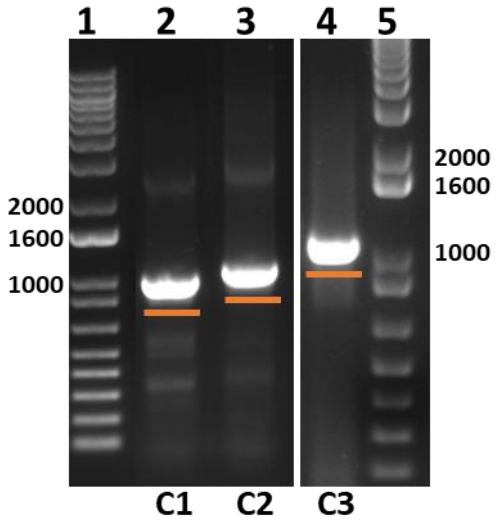


Figure C.3. PCR results of three new DV-1-1-Nuc constructs, run on a 2 % agarose gel. Construct 1 (C1) is at the expected band size of 1,060 bp, construct 2 (C2) is at the expected band size of 1,165 bp and construct 3 (C3) is at the expected band size of 1,174 bp. A 1 Kb^{+} ladder was used was used as a marker to confirm band sizes.

C.4 Protein expression and purification

C.4.1 Large scale purification of MBP-tagged DV-1-1-Nuc

IMAC and gel filtration chromatography produced soluble expressing, active DV-1-1-Nuc$_{MBP}$ for use in characterisation experiments (**Figure C.4.1**). An IMAC purification resulted in elution of DV-1-1-NucMBP protein from the IMAC column, with the addition of 10 mM imidazole. Protein was highly expressed and in fractions with several *E. coli* contaminating proteins. Target protein was further purified using an Amylose Resin column and protein eluted off column in a single peak along with another highly expressed *E. coli* contaminant protein. A gel filtration purification saw the separation of target protein from most of the *E. coli* contaminant protein.

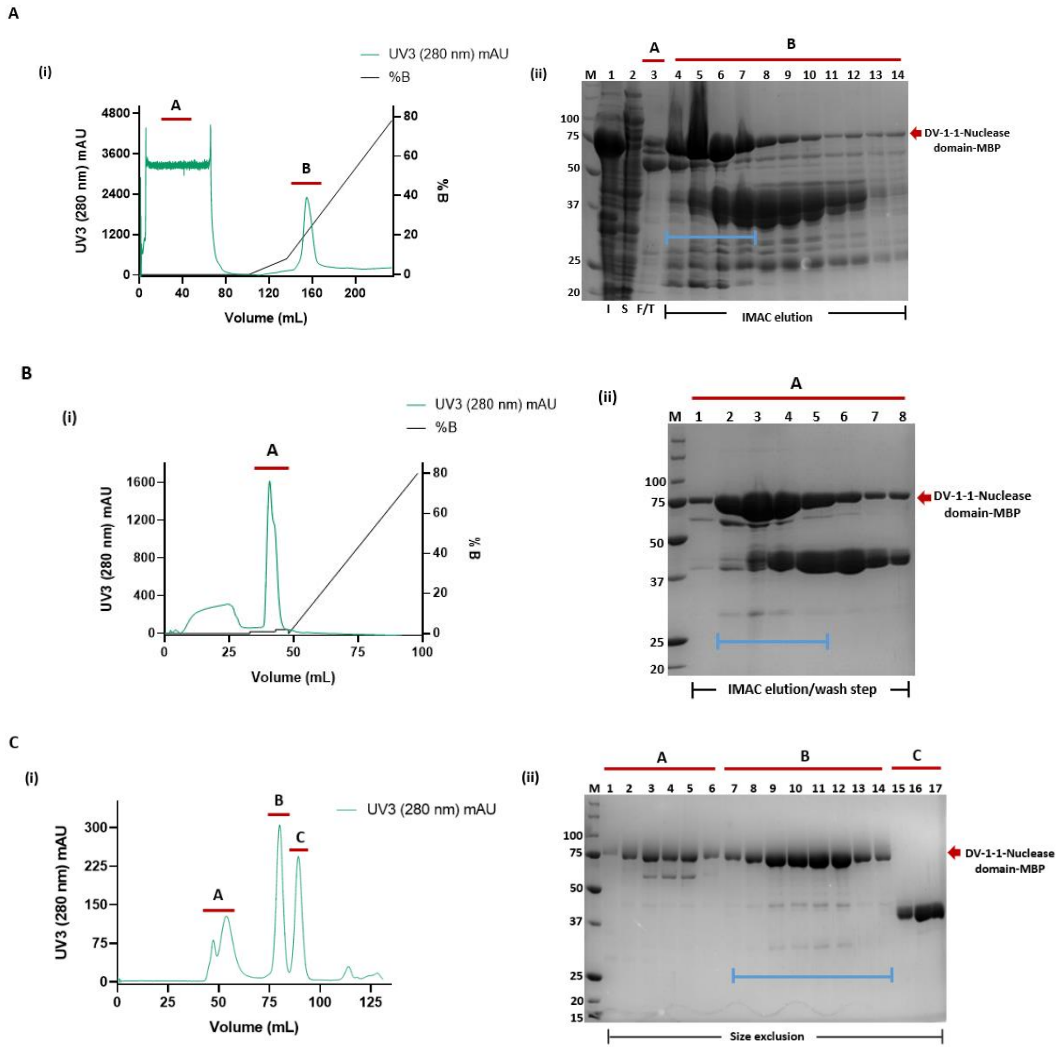


Figure C.4.1. IMAC, MBP and gel filtration chromatograms (i) and SDS PAGE gels for production of DV-1-1 Nuc_{MBP} domain protein from *E. coli* (DE3) Origami (ii). **A**) IMAC purification of DV-1-1-Nuc_{MBP}. (i) Peak A represents flow through during IMAC purification, peak B represents where the protein of interest (80 kDa) eluted during the elution step of the IMAC purification. Lanes 1-3 are; insoluble (I), soluble (S) and flowthrough (F/T), lanes 4-14 are fractions eluted during the imidazole gradient in the first IMAC step. **B**) MBP purification of DV-1-1-$Nuc_{MBP.}$ (i) Peak A represents where the protein of interest (80 kDa) eluted during the elution/wash step of the MBP purification. Lanes 1-8 are fractions eluted during the maltose gradient in the MBP purification. **C**) gel filtration purification of DV-1-1-Nuc_{MBP}. (i) Peak A represents where the protein of interest (DV-1-1-Nuclease domain-MBP, 80 kDa) eluted along with an *E. coli* contaminant, which may be forming a complex with the nuclease protein, causing it to come off the column early. Peak B represent also where the protein of interest (DV-1-1-Nuclease domain-MBP, 80 kDa) eluted. Small *E. coli* contaminants are also present in these fractions. Peak C represent fractions containing an *E. coli* contaminant, which was very strongly expressed and was visually observed through out the first two purifications steps. (ii) Lanes 1-6 are fractions eluted in peak A (i), lanes 7-14 are fractions eluted in peak B (i), lanes 15-17 are fractions eluted in peak C (i).

C.4.2 Small scale expression trials of DV-1-1-Nuc mutant

DV1-1-Nuc mutant (D37A/H38A) protein was recombinantly expressed in multiple *E. coli* strains; origami (DE3) BL21 pLysS (DE3) and arctic express (DE3). All *E. coli* strains were grown in small scale cultures, at 15 ºC and 25 ºC overnight. Protein was expressed in His-tagged (pDEST17) and MBP-tagged

(pHMGWA) plasmids. The following figure shows results from the trials that gave the best expression of protein. The his-tagged protein had very low expression in all strains, at both temperatures, while the MBP tagged protein showed very high level of soluble protein expression, particularly at 25 ºC (**Figure C.4.2**).

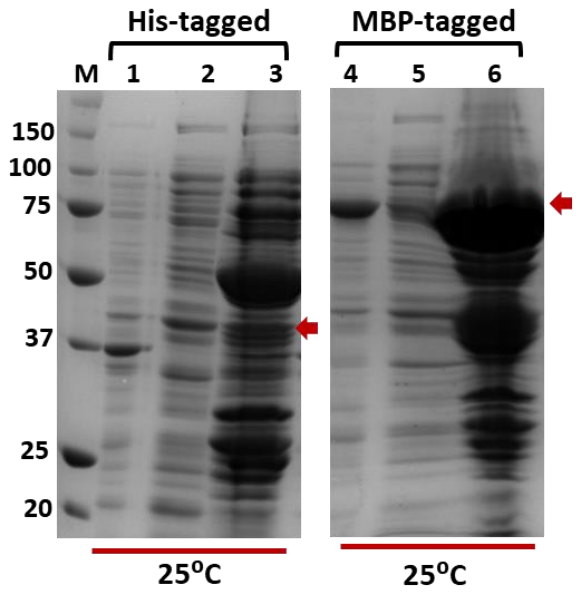


Figure C.4.2. SDS PAGEs of small-scale expression trials for DV-1-1-Nuclease mutant Lanes 1,4, represent insoluble pellet samples, lanes 2,5, represent soluble samples and lanes 3,6, represent samples bound to Ni beads. His-tagged protein (pDEST17 plasmid) should be 40. kDa and MBP-tagged (pHMGWA plasmid) protein should be 80.1 kDa. Red arrows indicate expressed protein at the correct size. M in each gel stands for a precision plus 250 kDa protein marker. Proteins were recombinantly expressed in *E. coli* origami cells and grown at 25ºC overnight.

C.4.3 Large scale purification of DV-1-1-Nuc mutant

IMAC and gel filtration chromatography produced soluble expressing DV-1-1-Nuc mutant protein for use in characterisation experiments (**Figure C.4.3**). An IMAC purification resulted in DV-1-1-Nuc mutant protein eluting early off the IMAC column, along with several *E. coli* contaminant proteins. Fractions containing target protein were pooled, up concetrated and incubated overnight with TEV protease. Following incubation with TEV, the protein sample was purified by reverse IMAC chromatography. A pre and post TEV sample shows that there was successful removal of the MBP tag from DV-1-1Nuclease domain mutant, with a new band visuable in the post tev sample (36 kDa). Protein was further purified by gel filtration chromatography, where DV-1-1-Nuc mutant protein was present in Peak D.

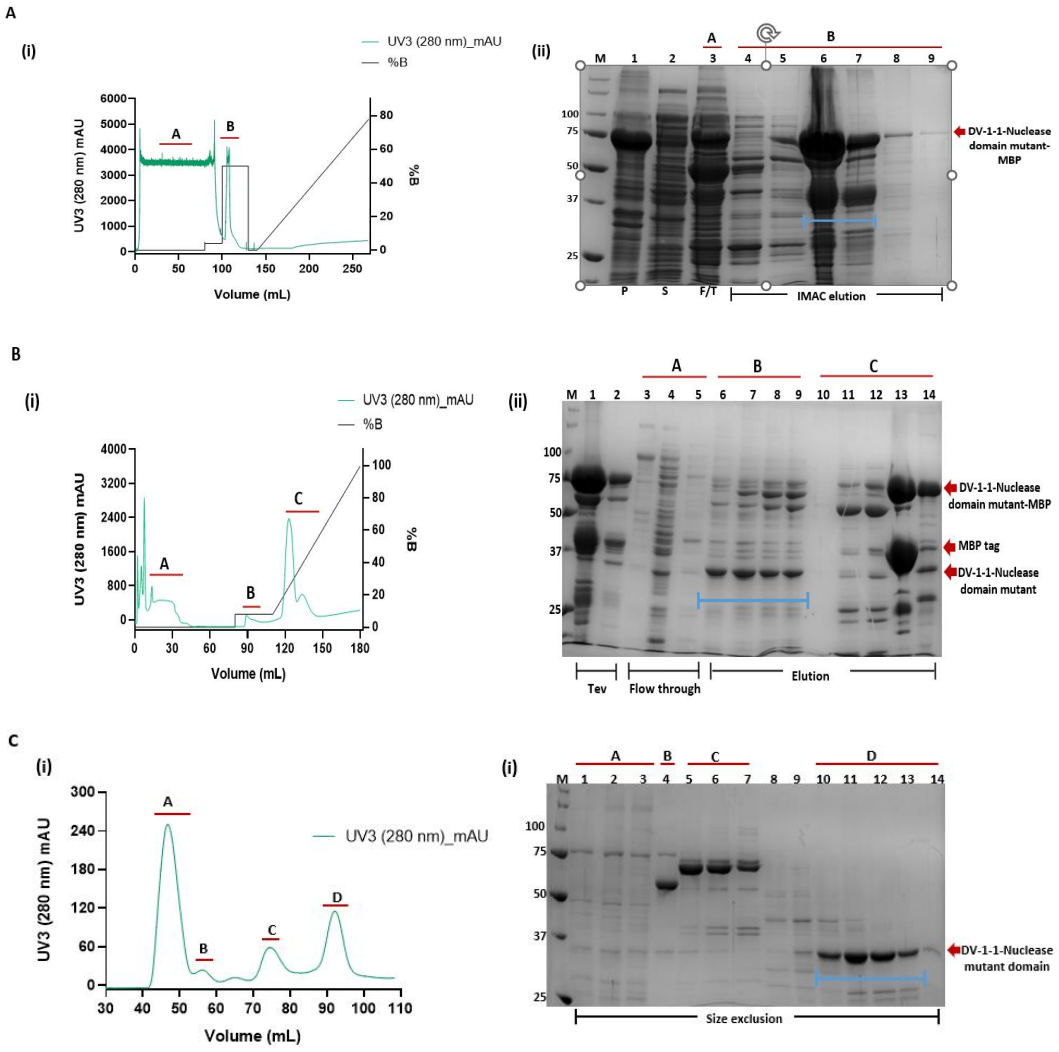


Figure C.4.3. IMAC and gel filtration chromatograms (i) and SDS PAGE gels for production of DV-1-1-Nuclease domain mutant from *E. coli* (DE3) origami (ii). **A**) IMAC purification of DV-1-1-Nuc$_{MBP}$ mutant (i) Peak A represents flow through during IMAC purification, peak B represent fractions that eluted early off the IMAC column, during the wash buffer step. Lanes 1-3 are; insoluble (I), soluble (S) and flowthrough (F/T), lanes 4-9 are fractions that eluted early off the IMAC column, during the wash buffer step. DV-1-Nuclease mutant protein (80 kDa) eluted off the IMAC column during the wash buffer step, along with *E. coli* protein contaminants. **B**) reverse IMAC purification of DV-1-1-Nuc mutant. (i) Peak A represents flow through during IMAC purification, peak B represents fractions that contain the de-tagged protein of interest (36 kDa), that eluted from IMAC column, with a low (8%) imidazole concentration, peak C represents fractions eluted during the elution step of the IMAC purification. (ii) Lanes 1-2 are pooled IMAC fractions before the addition of TEV (1) and fractions after an overnight incubation with TEV (2), Lanes 3-5 are fractions from the flowthrough during the reverse IMAC purification. Lanes 6-9 are fractions that eluted off the IMAC column as a small peak, with a low imidazole (8) wash. Lanes 10-14 are fractions eluted during the imidazole gradient step. **C**) gel filtration purification of DV-1-1-Nuc mutant. (i) Size exclusion purification resulted in four peaks, with DV-1-1 Nuclease mutant domain protein (36 kDa) represented in peak D. (ii) Lanes 1-6 are fractions eluted in peak A (i), lanes 7-14 are fractions eluted in peak B (i), lanes 15-17 are fractions eluted in peak C (i). The blue bar indicates fractions that were pooled and used in the next purification step. Chromatogram graphs were designed in GraphPad Prism, version 9.0.0.

C.4.4 Large scale purification of DV-1-1-Nuc$_{MBP}$ mutant

IMAC and gel filtration chromatography produced soluble expressing, DV-1-1-Nuc$_{MBP}$ mutant for use in characterisation experiments (**Figure C.4.4**). An IMAC purification resulted in elution of DV-1-1-NucMBP protein from the IMAC column, with the addition of 20 mM imidazole. Protein was highly expressed and in fractions with several *E. coli* contaminating proteins. Target protein was further purified using an Amylose Resin column and protein eluted off column in a single peak along with another highly expressed *E. coli* contaminant protein. A gel filtration purification saw the separation of target protein from most of the *E. coli* contaminant protein, with target protein present in Peak B.

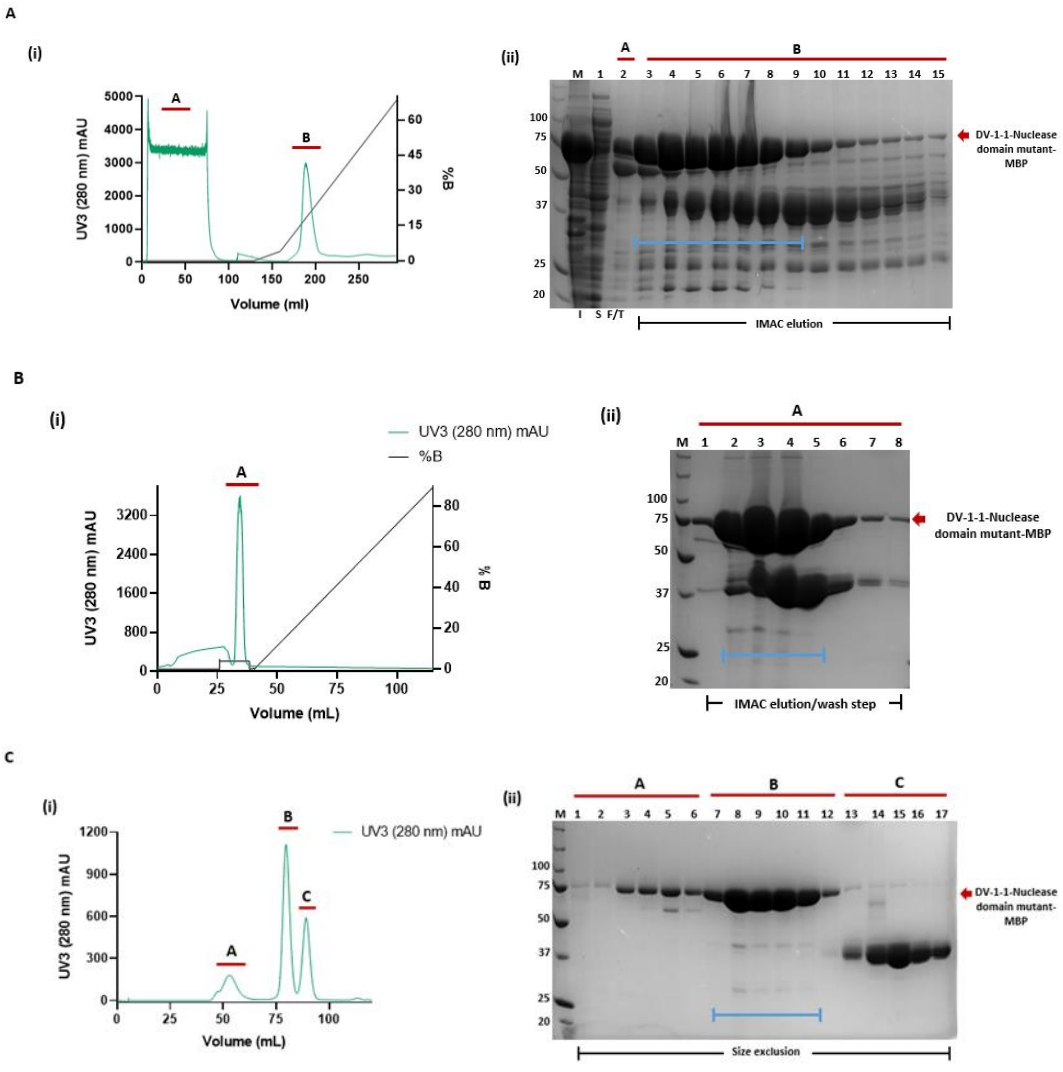
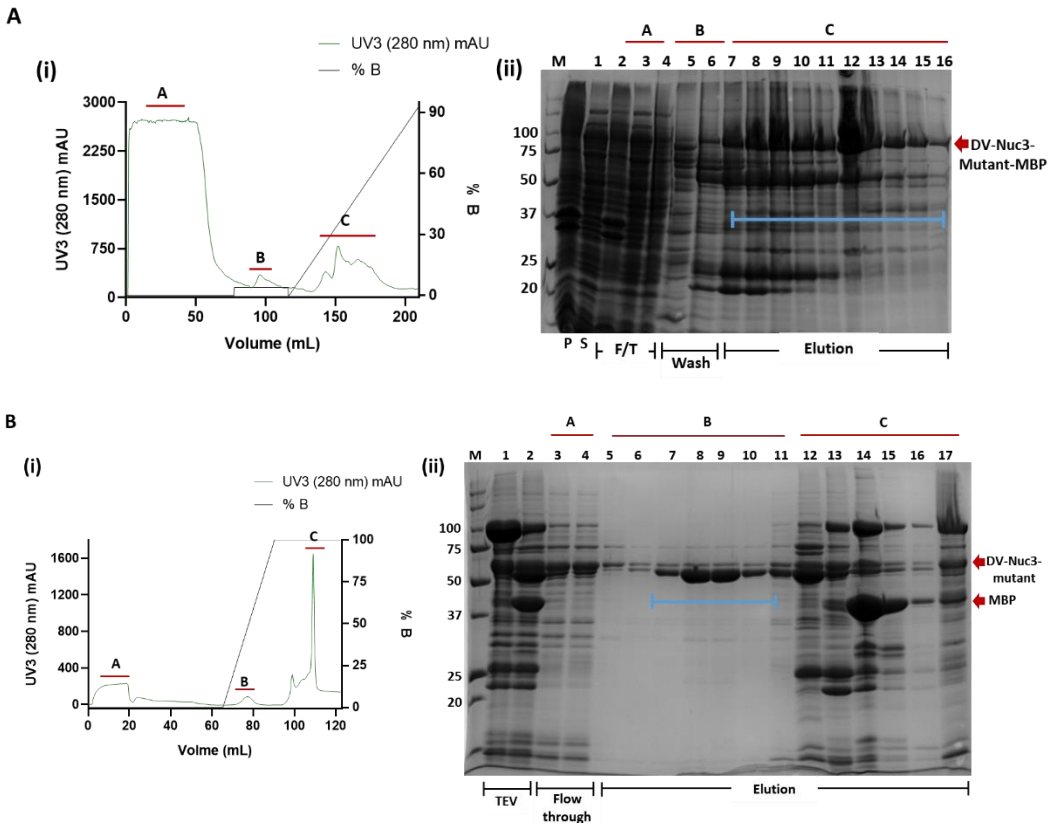


Figure C.4.4. IMAC, MBP and gel filtration chromatograms (i) and SDS PAGE gels for production of DV-1-1-Nuc_{MBP} mutant protein from *E. coli* (DE3) Origami (ii). **A**) IMAC purification of DV-1-1-Nuc_{MBP} mutant. (i) Peak A represents flow through during IMAC purification, peak B represents where the protein of interest (80 kDa) eluted during the elution step of the IMAC purification. Lanes 1-3 are insoluble (I), soluble (S) and flowthrough (F/T), lanes 4-15 are fractions eluted during the imidazole gradient in the first IMAC step. **B**) MBP purification of DV-1-1-Nuc_{MBP} mutant. (i) Peak A represents where the protein of interest (80 kDa) eluted during the elution/wash step of the MBP purification. Lanes 1-8 are fractions eluted during the maltose gradient in the MBP purification. C) gel filtration purification of DV-1-1-Nuc_{MBP} mutant. (i) Peak A represents where the protein of interest (DV-1-1-Nuclease domain $mutant_{MBP}$ protein, 80 kDa) eluted along with an *E. coli* contaminant, which may be forming a complex with the nuclease protein, causing it to come off the column early. Peak B represent also where the protein of interest (DV-1-1-Nuclease domain mutant-MBP, 80 kDa) eluted. Small *E. coli* contaminants are also present in these fractions. Peak C represent fractions containing an *E. coli* contaminant, which was very strongly expressed and was visually observed through out the first two purifications steps. (ii) Lanes 1-6 are fractions eluted in peak A (i), lanes 7-12 are fractions eluted in peak B (i), lanes 13-17 are fractions eluted in peak C (i). The blue bars indicates fractions that were pooled and used in the next purification step. Chromatogram graphs were designed in GraphPad Prism, version 9.0.0.

C.4.5 Large scale purification of DV-Nuc3 mutant



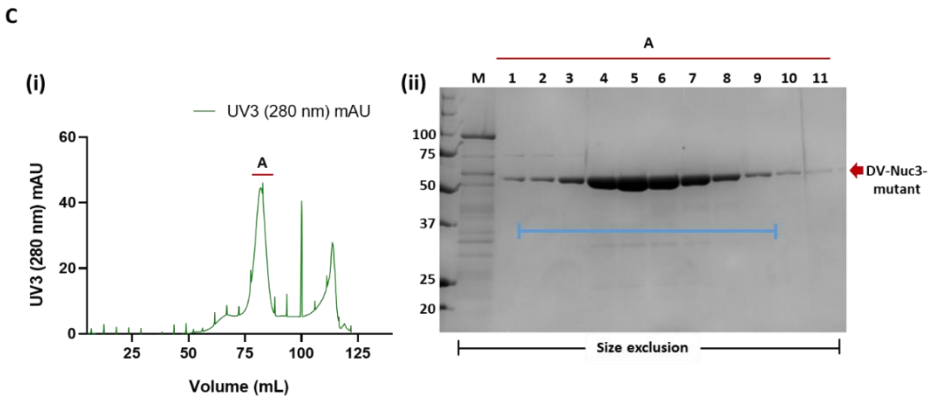


Figure C.4.5. IMAC and gel filtration chromatograms (i) and SDS PAGE gels for production of DV-Nuc3 mutant from *E. coli* (DE3) origami (ii). **A**) IMAC purification of DV-Nuc3 mutant (i) Peak A represents flow through during IMAC purification, peak B represent fractions that eluted early off the IMAC column, during the wash buffer step, peak C represents fractions containing DV-Nuc3 mutant. Lanes 1-4 are; insoluble (I), soluble (S) and flowthrough (F/T), lanes 5-6 are fractions that eluted early off the IMAC column, during the wash buffer step, lanes 7-16 represent proteins that eluted off the IMAC column with the addition of imidazole. **B**) reverse IMAC purification of DV-Nuc3 mutant. (i) Peak A represents flow through during IMAC purification, peak B represents fractions that contain the de-tagged protein of interest (58.3 kDa), that eluted from IMAC column, with a low (8 %) imidazole concentration, peak C represents fractions eluted during the elution step of the IMAC purification. (ii) Lanes 1-2 are pooled IMAC fractions before the addition of TEV (1) and fractions after an overnight incubation with TEV (2), Lanes 3-4 are fractions from the flowthrough during the reverse IMAC purification. Lanes 5-11 are fractions that eluted off the IMAC column as a small peak, with a low imidazole (8) wash. Lanes 12-17 are fractions eluted during the imidazole gradient step. **C**) gel filtration purification of DV-Nuc3 mutant. (i) gel filtration purification resulted in 1 main peak, with DV-Nuc3 mutant domain protein (58.3 kDa) represented in peak A. (ii) Lanes 1-11 are fractions eluted in peak A (i). The blue bar indicates fractions that were pooled and used in the next purification step. Chromatogram graphs were designed in GraphPad Prism, version 9.0.0.

C.5 Temperature dependence of DV-Nuc3

Temperature dependence activity assays were carried out with uracil matched DNA substrate, from 0 ºC up to 75 ºC. These reactions were carried out in a shorter time period (4 hours), compared to the previous assays, to observe the specific cutting pattern at different temperatures. Here nuclease activity is observed from 5 ºC to 45 ºC.

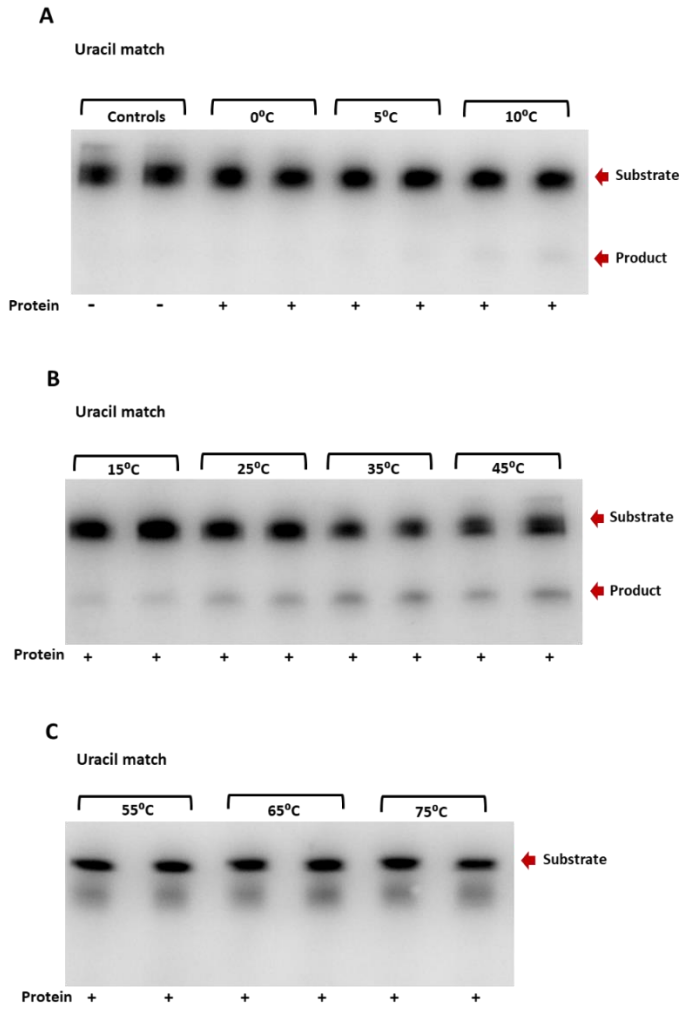


Figure C.5.1. Urea PAGEs of nuclease activity assays on uracil match DNA substrate by DV-Nuc3 protein, at different temperatures. Samples containing protein are annotated (+) and samples with no protein are annotated (-). Temperatures range from 0 ^{o}C up to 75 ^{o}C. All reactions are run in duplicates of 2. Substrate and products are indicated by red arrows. Reactions were carried out for 4 hours at the annotated temperature, with 2 µM final DV-Nuc3 protein concentration and 10 mM final concentration of magnesium. Results of activity assays were visualized using iBright™ CL750 Imaging System, Invitrogen™.

C.6 DSF thermal melts for DV-Lig5 and DV-Nuc3

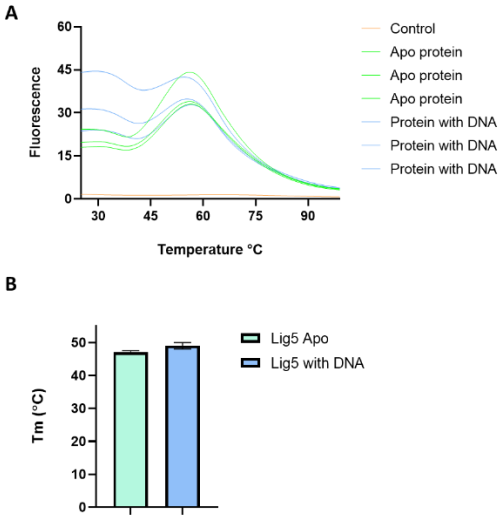


Figure C.6.1. DSF thermal melt curves of DV-Lig5 protein. **A**) DSF thermal melt of DV-Lig3 protein with no DNA and the addition of nick DNA substrate. **B**) Quantification of T_ms derived from first derivate of thermal melt curves. Each reaction was carried out in triplicate. Graphs were designed using Prism version 8 (GraphPadSoftware).

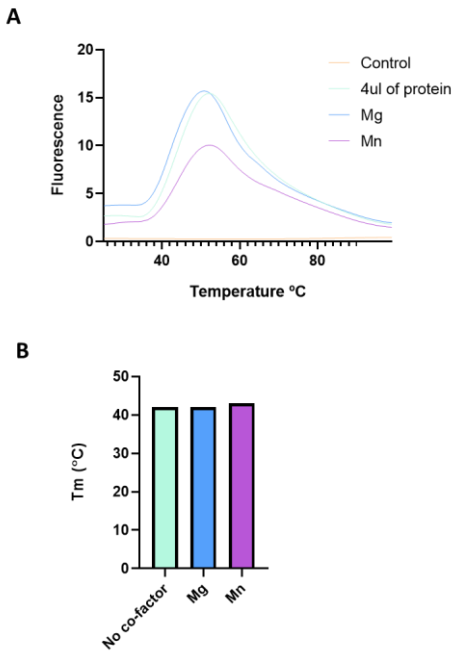


Figure C.6.2. DSF thermal melt of DV-Lig3 protein. **A**) DSF thermal melt of DV-Lig3 protein with the addition of 10 mM magnesium and manganese metal ions. **B**) quantification of T_ms derived from first derivate of thermal melt curves. Graphs were designed using Prism version 8 (GraphPadSoftware).