Synthetic Biology Methods to Optimise T4 RNA Ligase Activities

Amy Easey

A thesis submitted for a degree of Doctor of Philosophy at the University of East Anglia.

September, 2018.

School of Biological Sciences, University of East Anglia, Norwich.

Abstract

The T4 nucleic acid (NA) ligases repair breaks in nucleic acids, which can be double-stranded DNA, RNA or a combination of both. The bacteriophage T4 has 3 NA ligases, one is a DNA ligase (T4Dnl) and two are RNA ligases: T4 RNA ligase 1 (T4Rnl1) and T4 RNA ligase 2 (T4Rnl2). These T4 DNA and RNA ligases are widely used as tools in a number of molecular biology techniques. A structural bias has been reported when the T4 RNA ligases are used in next generation sequencing (NGS) adapter ligation and there are issues with nucleic acid manipulation; these biochemical activities could be improved. Synthetic biology techniques were used to engineer a number of novel enzymes by combining the ligases with other proteins which have specific biochemical activities; and by mutating amino acid residues. A total of 22 novel enzymes have been constructed and tested to check changes in ligation activity and additional functions of the ligase fusions. This includes 8 ligases with specific amino acid changes to motif I, 4 ligases with nucleotide binding domain swaps, 3 adenylate kinase (AK)-T4 NA ligase fusions, 3 polynucleotide kinase (Pnk)-T4 NA ligase fusions and 3 fusions with truncated *Mycobacterium smegmatis* LigD (MSLigD) and the T4 NA ligases.

Ligation analysis revealed that a number of the fusion ligases were functional, and exhibited novel biochemical activities. This included utilising a new co-factor for the AK-ligase fusions, where ADP was used instead of ATP for ligation. Lower quantities of AKRnl1 were required to achieve the same ligation as T4Rnl1, an interesting feature for industrial retail. The Pnk-ligase fusions correctly ligated substrates that could not be ligated by the original T4 NA ligases, where the DNA/RNA substrate break had 3’ phosphate instead of 5’ phosphate. The amino acid mutations indicated the published sequences for the conserved polypeptide motifs are often lacking important surrounding amino acids. Fusions of T4Dnl with truncated MSLigD offered a way to use higher concentrations of ATP for ligation, in high ATP concentrations MsLigD dissociates from the DNA and is unable to complete the ligation reaction. Optimisation studies need to be completed for the new functional ligase fusions and the modified T4 NA ligases, considering temperature, buffer constituents and reaction time. The ligases detailed in this body of work offer exciting examples of creating fusion multi-function enzymes with novel activities. These new ligases may become alternative tools for molecular techniques, especially in NGS adapter ligation.

This PhD work is part of a BBSRC iCASE studentship with Inspiralis (Norwich Research Park Innovation Centre, Colney Lane, Norwich NR4 7GJ) who retail T4 nucleic acid ligases.

Contents

Acknowledgements

First and foremost, thanks go to Dr Richard Bowater. If you hadn't decided to take a chance on the person who didn't turn up to her interview none of this would have been possible. Your leadership and guidance throughout this PhD have been fantastic and I am forever grateful for this opportunity. Thanks also go to Dr Andrew Gates and Dr Nicolas Burton.

Many thanks to the lab members from Bio 2.30. Special thanks go to Dr Michael Norman, I don't think my PhD would have been anywhere near as fun without you, especially all our SciComm opportunities; Sophie Bennett, we started together, lived together and you've been nothing but a source of Ah Ha's throughout my PhD; Dr Simone Payne, coffee and cake, or a night out, you've been a much needed source of fun and support and a huge thank you to Dr Marcus Edwards, for spending far too much of your time teaching me in the lab, teaching me CCP4MG, helping me trouble shoot and letting me lead you astray with coffee breaks! I'd also like to thank all of the staff, past and present who have given me so many opportunities throughout my time at UEA, especially Dr Kay Yeoman.

Enormous thanks must go to my friends and family, my PhD adventure would have been really boring without you all. To my family, who have backed every (sometimes stupid) decision I have made, but especially to Debi, Jo, Madi and Jake thanks for keeping me grounded. To the 'family': Beckie, Harrison, Gemma, Morris, Darren and baby Lanesra, I couldn't have done this without you guys, I needed all those laugh until I cried moments. To the Lowestoft girls: Talia, Nikki, Laura, Catherine, Jade and Charlotte, thank you all for the hilarious nights out, the Harvester dinners and listening to me moan, here's to many more!

Finally, to Nathan. I'm forever grateful for all that you have done and do for me. This really doesn't express how much you've supported me while trying to write this. For making me smile and for always believing I could do this, thank you.

I'd like to dedicate this work to Horace Easey, for always making me laugh and smile. It's not your life works but you finally have your name in print.

Abbreviations

ADP	Adenosine diphosphate
AK	Adenylate Kinase
AKDnl	Adenylate Kinase T4 DNA ligase fusion
AKRnl1	Adenylate Kinase T4 RNA ligase 1 fusion
AKRnl2	Adenylate Kinase T4 RNA ligase 2 fusion
AMP	Adenosine monophosphate
ATP	Adenosine triphosphate
° C	Degrees Celsius
DNA	Deoxyribonucleic acid
IPTG	Isopropyl-β-D-thiogalactopyranoside
LB	Luria-Bertani
LigDDnl	Truncated MsLigD fused to T4Dnl
LigDRnl1	Truncated MsLigD fused to T4Rnl1
LigDRnl2	Truncated MsLigD fused to T4Rnl2
MSLigD	*Mycobacterium smegmatis* LigD
NAD	Nicotinamide adenine dinucleotide
NHEJ	Nonhomologous end joining
PCR	Polymerase chain reaction
Pnk	Polynucleotide Kinase
PnkDnl	Polynucleotide Kinase T4 DNA ligase fusion
PnkRnl1	Polynucleotide Kinase T4 RNA ligase 1 fusion
PnkRnl2	Polynucleotide Kinase T4 RNA ligase 2 fusion
RNA	Ribonucleic acid
SDS PAGE	Sodium dodecyl sulphate polyacrylamide gel electrophoresis
T4Dnl	T4 DNA Ligase
T4Rnl1	T4 RNA Ligase 1
T4Rnl2	T4 RNA Ligase 2
UV	Ultraviolet

List of Tables

List of Figures

Chapter 1

Introduction

Chapter 1: General Introduction

1.1 Molecular Biology

Molecular biology has revolutionised our understanding of living systems. The central dogma, which describes the transfer of information at a molecular level, from nucleotide bases to amino acid composition, to determine protein structure (Crick, 1970), has been described as a keystone in molecular biology. Long before the discovery of DNA and RNA, scientists had categorised behaviours and traits, without any way to determine how the information for those heritable features was stored. The binomial naming system for organisms was developed by Carolus Linnaeus (Linnaeus, 1758) and it would be another 70 years before the first theory of evolution was published by Jean-Baptiste Lamarck (reviewed in Corsi, 2015). Another 50 years passed before Darwin wrote the Origin of Species and Alfred Wallace published theories about natural selection (reviewed in Shermer, 2002). Work continued, and Mendel's experiments on hereditary traits paved the way for how DNA might be involved in heredity (reviewed in Olby, 1994).

1.1.1. The discovery of DNA

Nucleic acid discovery and characterisation was happening alongside the development of the theory of evolution. In 1869 Friedrich Miescher isolated a mixture of macromolecules, one of which was later confirmed to be DNA, from leukocytes by precipitation using acidified solutions and ether and named the substance *nuclein* (Dahm, 2008). Miescher continued to analyse the substance and revealed its composition to contain Oxygen, Nitrogen, Hydrogen and large amounts of phosphoric acid – which was dissimilar to proteins that had already been discovered (Perutz, 1995; Dahm, 2008). The function of *nuclein*, or DNA, was still under debate, however, the discovery of DNA in germ cells led Miescher to think it might be involved in fertilisation (Reviewed in Dahm, 2008). In 1889 another scientist, Altmann, separated *nuclein* from all proteins in the cell precipitate and named the substance nucleic acid – although this was the same product isolated by Miescher (Pollister and Mirsky, 1943; Dahm, 2008).

The idea that proteins carried genetic information was not challenged until 1944, although Miescher had proposed this theory when working on *nuclein*, when it was suggested that in fact DNA was the molecule responsible for carrying genetic information (McCarty and Avery, 1946; Dahm, 2008). The discovery of the composition of DNA, the bases adenine, thymine, guanine and cytosine and the variability of these bases between species (Chargaff, 1951; Dahm, 2008), led to the research that would conclusively determine that DNA was genetic material (Hershey, 1952). It was finally the discovery of the structure of DNA in 1953 by Watson and Crick (Watson and Crick, 1953; Watson *et al.*, 1953) along with work on the crystallographic data by Franklin and Wilkins (Franklin and Gosling, 1953; Wilkins, 1956; Klug, 1968) that allowed molecular biology as a research field to really take off.

1.1.2. The Structure of DNA

The DNA structure published in 1953 is a right-handed DNA helix (B-DNA), the strands rotate in a clockwise direction (Franklin and Gosling, 1953; Watson and Crick, 1953; Klug, 1968). DNA is comprised of nucleotides, deoxyribose sugars and phosphate groups (Figure 1.1), the helical structure of DNA protects its nucleotides from environmental damage (Sinden, 1994). The DNA helix structure is caused by the complementarity of the four different nucleotides – adenine (A) with thymine (T); and guanine (G) with cytosine (C) (Figure 1.1) (Watson and Crick, 1953). This discovery supported previous findings in salmon sperm composition of equal amounts of A and T; and C and G (Chargraff *et al.*, 1951; Travers and Muskhelishvili, 2015). The complementary nature of the nucleotides on each strand of the helix allow for replication of 2 identical copies. These copies are necessary for information transfer to RNA which encodes protein sequences (Travers and Muskhelishvili, 2015). The outside of the DNA helix is the negatively charged phosphate backbone. The phosphate groups on this backbone are attached to the adjacent deoxyribose sugars at the 5’ carbon and 3’ carbon positions and link the nucleotides (Figure 1.1) (Sinden, 1994). DNA is synthesised in the 5’ to 3’ direction: at the 5’ end, there is a single phosphate group at this end – 5’ PO_4; the 3’ end and has a hydroxyl group – 3’ OH (Sinden, 1994). The complementarity of DNA strands results in the two strands running in opposite directions but they are only synthesised in the 5’ to 3’ direction (Figure 1.1).

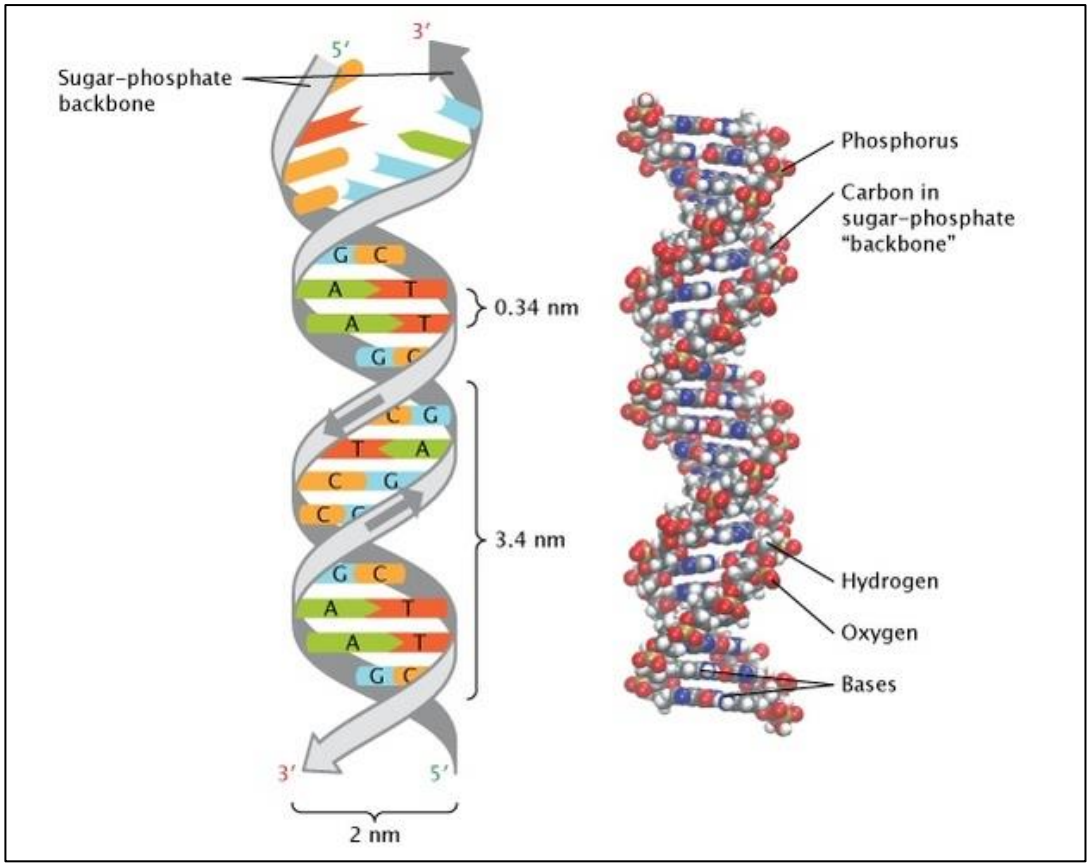


Figure 1.1. The structure of double stranded DNA. The structure of double stranded DNA published in 1953, the sugar-phosphate backbone and complementary bases A (green) and T (red), C (yellow) and G (blue) are shown (Pray, 2008).

There are other forms of helices, including a second form of the right handed helical structure: A-DNA (Sinden, 1994; Travers and Muskhelishvili, 2015). A-DNA is a shorter, more compact helix, B-DNA can revert to A-DNA at lower humidity (Travers and Muskhelishvili, 2015). DNA with a Left handed helix is referred to as Z-DNA, the strands turn in an anti-clockwise direction. Other variations to the DNA structure that have been reported, include bubbles – where the 2 strands separate forming a bubble in the structure; three stranded helices, called H-DNA and slipped loops – where a loop of nucleotides is on one of the strands (Travers and Muskhelishvili, 2015).

1.1.3. DNA sequencing

The identification of the structure of DNA steered researchers to develop the technology required to determine the order of nucleotides in this important structure. This technology did not develop quickly and in the beginning it was possible to measure the

composition of the different nucleic acids in deoxyribonucleotides (dNTPs) but not the order (Holley *et al.*, 1961). The development of a technique using labelled dNTPs meant researchers could identify sequences (Sanger, Brownlee and Barrell, 1965) and using this method the first protein coding sequence from the bacteriophage MS2 was reported (Jou *et al.*, 1972; Heather and Chain, 2016). The first complete genome, 3569 residues in length, from the bacteriophage MS2 was sequenced and crucial sequences for viral maturation, including the protein coat and a replicase subunit, were identified (Fiers *et al.*, 1976). The next step in making sequencing high throughput was the separation of nucleotides using electrophoresis and polyacrylamide gels and this technique was used to sequence the genome from the bacteriophage *Φ*X174 (Sanger and Coulson, 1975; Maxam and Gilbert, 1977; Sanger *et al.*, 1977; Hutchison, 2007; Heather and Chain, 2016). The next major advance in Sanger sequencing resulted in a more rapid and accurate method (Sanger, Nicklen and Coulson, 1977), the use of dideoxynucleotides (ddNTPS) to inhibit DNA polymerase made it possible to terminate the extended primers at specific residues. It was possible to run 4 reactions, each containing a specific ddNTP, rather than using radioactive labelled dNTPs (Figure 1.2). The use of fluorescently labelled ddNTPs made it possible to use all four in the same reaction – each base could be identified as a different colour (Sanger, Nicklen and Coulson, 1977).

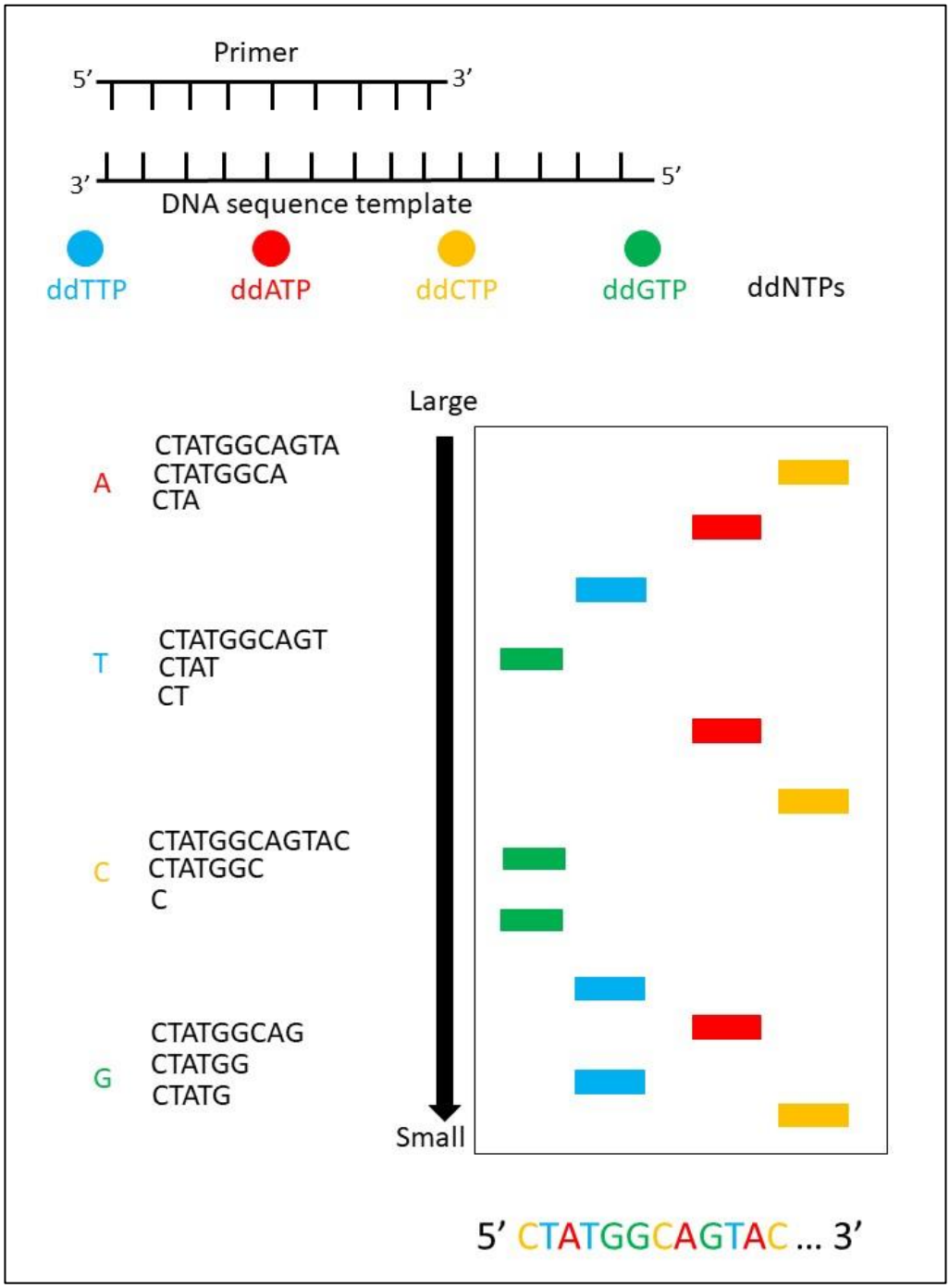


Figure 1.2. Sanger Sequencing with electrophoretic separation. Four reactions are set up containing the primer, DNA polymerase and one fluorescently labelled ddNTP to truncate the DNA fragment. Each reaction undergoes electrophoresis to separate out the fragments from top to bottom, by size, and the sequence can be determined by the bands which fluoresce.

This method won Sanger the Nobel Prize for Chemistry in 1980, this was shared with Walter Gilbert who developed a chemical sequencing method known as the Maxam-Gilbert method and Paul Berg who studied the biochemistry of nucleic acids focussing

on recombinant DNA. The Maxam-Gilbert method does not use DNA polymerase to make fragments but instead involves radiolabelling the nucleotides and then cleaving the DNA next to the radiolabelled bases (Maxam and Gilbert, 1977; Heather and Chain, 2016). In both methods the DNA fragments are run on a polyacrylamide gel to determine their length and ascertain the specific nucleotide positions (Heather and Chain, 2016).

Improvements to primer extension detection using capillary electrophoresis allowed for Sanger sequencing automation (Ansorge *et al.*, 1987; Swerdlow and Gesteland, 1990). DNA sequencing as a molecular technique has evolved from costly, time consuming methodologies to a high throughput automated technique which can deliver whole genome sequences in days and smaller bacterial genomes even quicker. The contribution of bioinformatics software and computing hardware developments has made the process more efficient; sophisticated bioinformatic software permits more complex analysis of larger genome sequences in a timely manner. Automation of the process for DNA sequencing resulted in accurate sequencing of smaller and smaller sequences, including sequencing short RNA molecules.

1.1.4. Next Generation Sequencing

Until completion of the human genome project, the sequencing technologies outlined previously were used. In the early 2000s it became clear that the Sanger approach had reached its limit and that the only way to make genome sequencing faster would be to begin making technological advances with high throughput machines. This resulted in high costs when technologies were first developed, but with optimisation and improved processes came new techniques – collectively called next generation sequencing (NGS). NGS methods are used to sequence a number of different sequences, including a type of small RNA called micro RNAs (miRNAs). Small RNA (sRNA) molecules are non-coding RNA molecules that regulate gene expression by post-transcriptional gene silencing (Zhang, 2009). Deep sequencing (where a genomic region is sequenced multiple times) is the gold standard used to profile these sRNA molecules within a genome and discover novel and closely related miRNA sequences (Jayaprakash *et al.*, 2011).

NGS is the collective term for a number of different sequencing technologies: Illumina sequencing, Roche 454 sequencing, Ion torrent (Proton/PGM sequencing) and SOLiD sequencing. These technologies developed from pyrosequencing (Heather and Chain, 2016), which is a sequencing method that measures light production proportional to the levels of pyrophosphate made (Hyman, 1988). Nucleotides are added one at a time and each time a nucleotide is incorporated pyrophosphate is released and converted proportionally to ATP by ATP sulfurylase; this ATP then catalyses the conversion of luciferin to oxyluciferin which produces light (Cummings *et al.*, 2013). The light intensity is proportional to the number of bases incorporated onto the template strand, any dNTP not incorporated is degraded before the next dNTP is added.

Each different NGS technology identifies these sequences differently but all require ligation of the miRNA to an adapter. Both Illumina sequencing and 454 sequencing read optical signals as bases are added and use annealed adapters. For Illumina sequencing these adapters are fixed to a slide while in 454 they are attached to beads (Figure 1.3 A and B). Ion torrent sequencing measures the release of H^+ ions when a dNTP is added to DNA. If the dNTP is the next base added, H^+ ions are released; the H^+ concentration is measured using a transistor and converted into a charge build up which is transmitted as a voltage charge (Merriman, R&D Team and Rothberg, 2012). An algorithm converts the voltage charge to a signal. The signals, or lack of, emitted after each dNTP wash indicate no incorporation of the base, 1-mer incorporation, 2-mer incorporation etc. (Merriman, R&D Team and Rothberg, 2012). This method is similar to 454 sequencing in that the addition of dNTPs are measured one at a time (Figure 1.3 C). The NGS technologies allow for faster sequencing, washing with dNTPs can be done relatively quickly and the instruments were designed to automate the process and they can process multiple samples at one time.

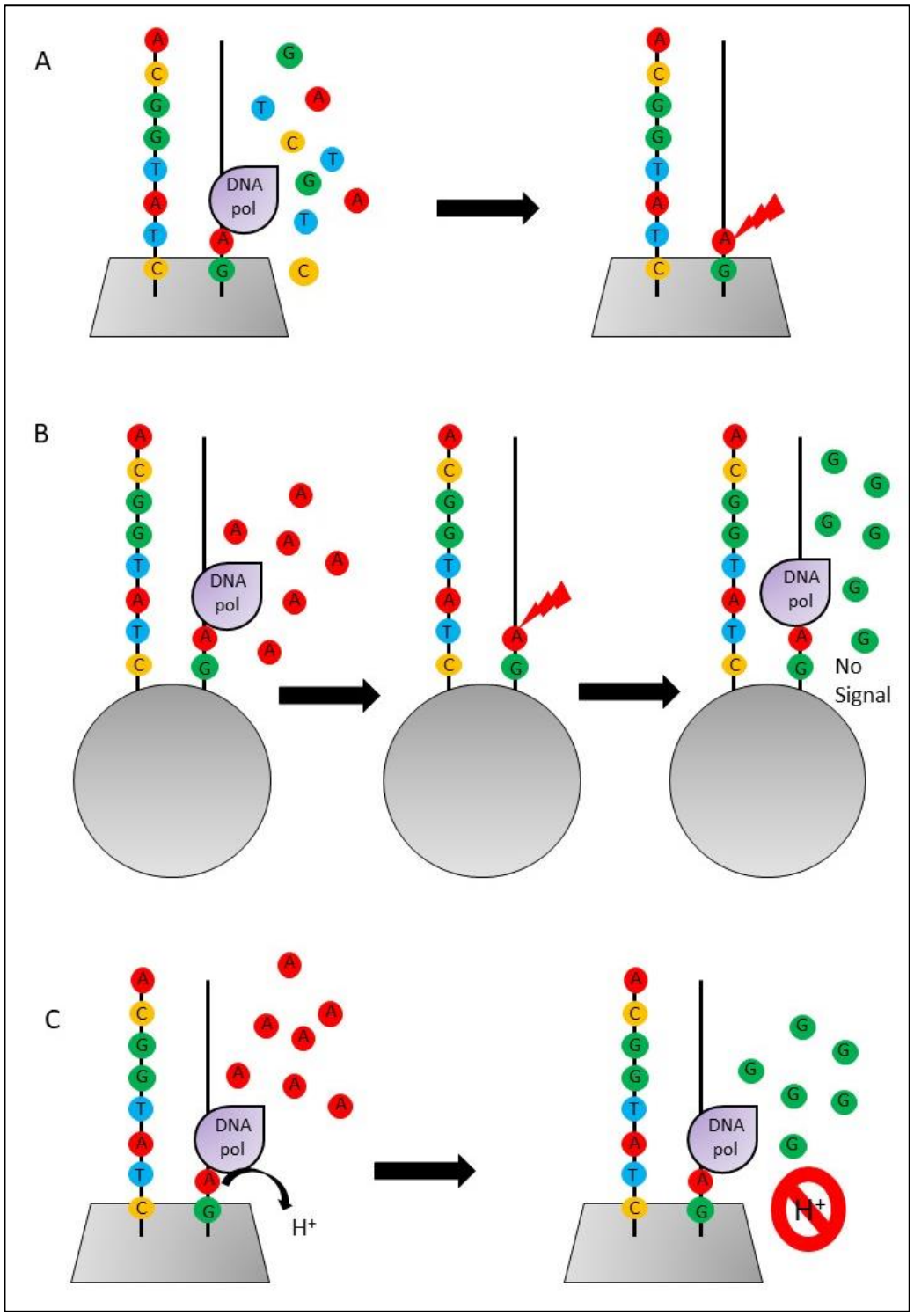


Figure 1.3. Next Generation Sequencing technologies. A) Illumina sequencing: fluorescently labelled nucleotides with a terminator attached allow detection of single nucleotide addition, the terminator and fluorescent signal are removed before the next cycle; B) 454 sequencing: single dNTPs are added separately, if incorporated a signal is released and the nucleotide order determined; C) Ion torrent sequencing: single dNTPs are added separately and the pH measured to determine if a base(s) was incorporated.

To discover mutations and new miRNAs it is necessary to sequence cDNA libraries using these NGS methodologies. The miRNA is isolated and fractionated by size, adapters are ligated to the 3' end first and then the 5' end. The miRNA undergoes reverse transcription and PCR amplification converting the adapter ligated miRNAs into cDNA clones required for NGS. The sequencing protocol varies depending on the sequencing technology used (Figure 1.3) (Hafner *et al.*, 2008). It is possible to carry out the quantitative polymerase chain reaction (qPCR) to check the levels of the miRNAs; it should be noted that qPCR can only be used to quantify known miRNAs (Hafner *et al.*, 2008; Sorefan *et al.*, 2012). In some cases there is under or over representation of miRNA quantification with NGS when compared to qPCR or microarray data. For example, a pool of miRNAs in 3 different breast tissue samples were sent to 6 microarray platforms and compared to both NGS and qPCR, the miRNA expression and sequencing results varied depending on the platform used and the sensitivity of the platforms (Git *et al.*, 2010). Sequence bias, when using ligases for adapter attachment has become a recent focus for researchers implementing these techniques. Ligase bias has been noted when using both of the T4 RNA ligases and a modified version of T4Rnl2. The bias was more prominent with the 5' adapter ligation step (Hafner *et al.*, 2011). Such bias of the ligases is usually with regards to DNA/RNA structure and not the sequence (Zamore and Haley, 2005). Despite these issues the NGS technologies are considered a cheaper, higher through put option compared to traditional Sanger sequencing.

Knowledge about DNA and the associated DNA sequencing technologies has come a long way: sequencers are shrinking, nanopore sequencers are commercially available (Branton *et al.*, 2008; Deamer, Akeson and Branton, 2016) and research has started looking at using DNA as a storage molecule for digital information (Grass *et al.*, 2015). DNA sequencing technologies are developing constantly and nowadays determining a DNA sequence is often the first, easy step in the identification of proteins. However, the discovery of proteins and research into their function had started before the discovery of DNA. The term protein is now common place in molecular biology with some proteins, specifically enzymes, utilised as research tools.

1.1.5. Protein Discovery

The first proteins, discovered in the late 1700s by Fourcroy, were albumin, fibrin and gelatin (Tanford and Reynolds, 2003). These were isolated from animal sources and sparked interest in understanding the composition of these molecules; researchers found different levels of nitrogen, oxygen, hydrogen, carbon and sulphur (Kossel, 1910; Tanford and Reynolds, 2003; Stanford, 2005). It was this compositional data that revealed similarities between the substances and they were classified under the name protein (Tanford and Reynolds, 2003). Work began to separate proteins from one another and elicit more information about their structures, and the discovery of the peptide bond and the different amino acids continued throughout the early 1900s (Kohler, 1973). In 1886, Zinnofsky discovered that for every iron atom in haemoglobin there are 600 carbon atoms – a much more complex molecule than previously studied proteins. In 1958, the first 3 dimensional (3D) structure of Myoglobin revealed proteins were much more complex than originally thought. The first myoglobin structure, with a resolution of 7 Å, showed densely packed regions of electrons and surprisingly lacked symmetry (Kendrew *et al.*, 1958; Tanford and Reynolds, 2003; Balasubramanian and Ponnuraj, 2010). Protein symmetry is thought to provide stability to the protein, the majority of membrane-bound and soluble proteins are symmetrical oligomeric complexes with at least two identical subunits (Goodsell and Olson, 2000). The technologies required to visualise 3D protein structures needed to become more sophisticated and by 1960 structures with a resolution of 2 Å were achieved. The range of current protein identification technologies is vast and often researchers make choices about the technologies used based on cost, time constraints and the type of protein being purified. While fundamental work to look at naturally occurring proteins and understand their functions is incredibly important, it is also possible to make modifications to proteins by altering the DNA sequence that encodes them resulting in recombinant proteins.

1.1.6. Modifying proteins

Protein mutations can affect protein stability, function and interaction (Reva, Antipin and Sander, 2011). Often mutations include addition, removal or substitution of single amino acids, but molecular research has evolved to look at how changing protein

domains by altering specific amino acids affects the function of the protein. This can be especially important when characterising proteins involved in pathways that are incorrectly regulated in diseases.

Site directed mutagenesis (SDM) is used to change specific amino acids by making base substitutions, insertions and deletions in the relevant gene (Kunkel, 1985). SDM can be used to study the effect of changes to nucleic acid sequence and how these changes impact the structure and function of the encoded proteins. The most commonly used SDM approach requires specific primers with the desired base modification (Metzenberg, 2002). These primers are used in PCR, whereby the sequence change in the primers replaces the original sequence and is incorporated into the amplicon. The PCR product can be treated with a mix containing a kinase, nucleic acid ligase (usually DNA but can RNA ligases can also be used) and the restriction enzyme *DpnI*. The kinase is used to firstly phosphorylate the DNA, so that the ligase is able to circularise the PCR product and then *DpnI* cleaves only the methylated template DNA (the newly synthesised DNA from PCR will not be cleaved), and after this treatment the product is ready for transformation into competent *Escherichia coli* (*E. coli*) cells (Metzenberg, 2002). While a relatively simple process, SDM is now used considerably less due to the reduced cost of whole plasmid synthesis (Carrigan, Ballar and Tuzmen, 2011). SDM, similar to synthetic biology techniques (see Section 1.2.), usually requires the use of synthetic oligonucleotides for PCR primers.

Understanding the central dogma and how the order of nucleotides in DNA translate into proteins has contributed to vital, fundamental and innovative research and has helped in developing knowledge about living organisms (Heather and Chain, 2016). The ability to make modifications to DNA for the synthesis of novel proteins has transformed molecular biology into a more engineering based discipline - synthetic biology; this new branch of molecular biology allows for a more complex and innovative type of science.

1.2. Synthetic Biology

Developments in molecular biology, including more advanced sequencing technologies and better protein identification methods paved the way for synthetic biology. Synthetic biology is described as the engineering, design and assembly, of systems and the redesigning of existing biological systems, devices and biological parts (Check, 2005; Endy, 2005) with a focus on solving specific problems. The principals involved in synthetic biology techniques are comparable to those used in engineering, and permit the design of novel biological systems in a systematic way (Serrano, 2007). Synthetic biology looks to assemble parts. The term 'parts' can be used to describe proteins, genes, amino acids and bases, and this can be achieved using different enzymatic molecular tools. It is worth remembering that the parts may need to be specific to the organisms. For example, differences in cell biology of eukaryotes and prokaryotes may mean that what works for one cell type may not work for the other. The most common techniques in synthetic biology utilise specific restriction enzymes and DNA ligases to achieve sequential joining of parts, to create a recombinant molecule. Synthetic biology is still a relatively new concept within biology and focus is often placed on the design of novel parts and their function (Knight, 2003; Serrano, 2007).

The different synthetic biology methodologies are often described as genetic engineering, falling into two categories of experiments: either there is the construction of non-natural building blocks that mimic natural functions or the building blocks already available are used to generate combinations not found in nature (Khalil and Collins, 2010). Different approaches can be used to achieve both of these goals, including BioBrick assembly, Gibson cloning and Golden Gate Cloning.

1.2.1. Golden Gate Cloning

A method with high success of ligating multiple molecular parts (DNA, RNA, ribosome binding sites, regulatory, protein coding, terminators and conjugates) is Golden Gate Cloning (GGC). A plasmid containing 10 parts was successfully ligated (Engler *et al.*, 2009) using one restriction enzyme and one DNA ligase. GGC is another single tube method that yields nearly 100 hundred percent recombination and provides the means

to make larger constructs from multiple smaller fragments, e.g. 9 fragments were successfully ligated using the Golden Gate method to form a 1.17 kb linear product (Engler, Kandzia and Marillonnet, 2008). Current research using GGC has resulted in highly accurate cloning and has mostly been achieved using plant systems (Patron *et al.*, 2015). This technique relies on the use of type IIS restriction enzymes to create constructs without the scarring seen with other methodologies such as BioBrick™ assembly (Shetty, Endy and Knight, 2008). GGC uses type IIS restriction enzymes that are able to cut 4 nucleotides outside of their recognition site to create sticky ends. These ends can be flanked by any nucleotide sequence, which gives more than 200 different combinations for these 4 nucleotide overhang regions (Engler, Kandzia and Marillonnet, 2008). The overhang regions can be specifically designed and as such can be arranged in an order that has already been pre-determined, this can be quite time consuming though as it has to be done carefully.

GGC relies on digestion and then ligation to create parts that can be assembled into a vector (Figure 1.4). The parts are designed so that the type IIS restriction sites are directed towards the parts to give the predefined sticky ends. The parts and vector then undergo restriction digestion, which results in sticky ends and removes the type IIS cut site. The digested parts undergo ligation and if the cloning has been successful a transformation experiment should confirm the correct clones (Engler, Kandzia and Marillonnet, 2008). The primary advantage for this technique is that it is successful in avoiding the expensive and more time consuming efforts involved with molecular cloning, and can also be carried out as a one tube reaction, like Gibson cloning, rather than multiple cloning steps. All reagents are heated and cooled between 10 and 50 times to ensure the DNA parts are digested and ligated repeatedly; this methodology is highly successful because once ligated the fragments are held in the constructs and cannot be re-digested after they have been assembled (Engler *et al.*, 2009). This whole process takes only 30 minutes and results in nearly 100 % cloning efficiency. The type IIS enzymes will cut at their recognition site regardless of the adjacent nucleotides; selection of appropriate type IIS enzymes will ensure the sequences only fit together one way, which means that in one reaction there is the possibility for the construction of multiple sticky end fragments (Engler *et al.*, 2009).

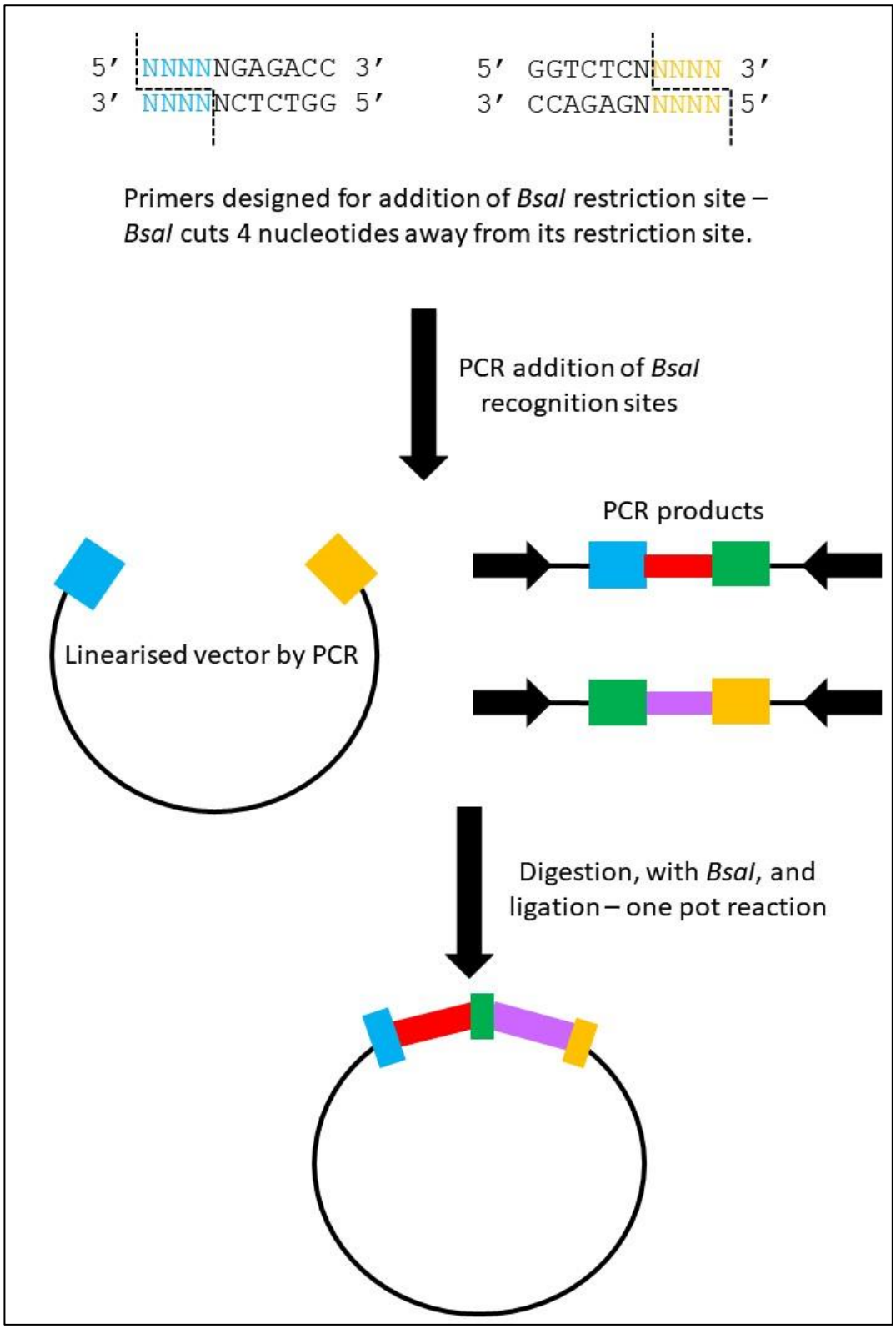


Figure 1.4. GGC of gene fragments into a vector. Primers are designed for the addition of the *BsaI* restriction site. The restriction enzyme *BsaI* can be used to create 4 bp overhangs in the forward or reverse direction for ligation of multiple genes (red and purple) into a vector. The small arrows (pointing left and right) indicate the direction of the *BsaI* recognition sites (5'-3') for sticky end creation. Once ligated the inserts lose the *BsaI* sites and the plasmid is correctly assembled.

GGC has been used to successfully create Transcription Activator-Like Effector nucleases (Cermak, Starker and Voytas, 2014) and to ligate 12 eukaryote transcriptional parts from plants (Patron *et al.*, 2015). GCC offers a time saving technique to create recombinant DNA plasmids for protein analysis. The type IIS restriction enzymes are just one molecular tool used in GGC. All of the techniques that come under the umbrella term synthetic biology use a multitude of enzymes as tools.

1.3. Enzymes as tools for molecular and synthetic biology

Enzymes can be classified into six categories, which are determined by the reaction they catalyse (Boyce and Tipton, 2001). The six categories are defined as 1) oxidoreductases, 2) transferases, 3) hydrolases, 4) lyases, 5) isomerases and 6) ligases (Boyce and Tipton, 2001). Over the last 60 years, the specific properties of these different types of proteins have been exploited for molecular and synthetic biology purposes. A quick search on any number of scientific commercial companies with biotechnology products provides a multitude of molecular tools for cloning, protein expression and purification as well as DNA modifying enzymes.

1.3.1. Protein Tags for purification

A wide variety of protein fusions have been developed for the purpose of purification or visualisation of protein interactions and networks in the cell (Terpe, 2003; Kimple and Sondek, 2004; Michelson-Horowitz *et al.*, 2005; Lowder *et al.*, 2011). When selecting a purification tag for a protein it is important to consider its effects on protein yield and cost. The type of tag selected will depend on the protein being purified and the host organism used to synthesise large amounts of protein. Some of the common host systems used are *E. coli*, yeast, human HeLa cells or *Drosophila*, but other host systems are available (Michelson-Horowitz *et al.*, 2005). The different positive and negative factors for each tag/protein and host system will impact tag/protein choice for purification of proteins.

One common tag used for purification from *E. coli* host cells is the Poly HIS tag system. This technique was developed in the 1980s (Hochuli *et al.*, 1988) and still remains one

of the most common forms of protein tagging for purification due to its low cost, ease of use and low impact on the protein (Kimple and Sondek, 2004). The tag itself is a chain of at least six histidine amino acids at either the N- or C-terminus of the protein that will be purified and relies on interactions between the histidine tag and transition metal ions in the purification matrix (Porath *et al.*, 1975).

Glutathione S-transferase (GST) is another common form of tagging often used for purification of polypeptides (Smith and Johnson, 1988; Smith, 2000; Terpe, 2003). An attractive feature of GST tagging is the ability to cleave the tag from the protein of interest while bound to glutathione agarose, providing a convenient separation method (Smith, 2000). Tag cleavage can also be achieved with His-tagged proteins but is usually not thought to be necessary as the His-tag has no impact on the protein structure (Carson *et al.*, 2007). However, there are problems with GST tagging, the glutathione matrix has a finite number of uses, usually between 4 and 20 regenerations, making GST tagging a less favourable option for large yield protein purifications as the finite number of uses increases the associated cost.

A highly utilised tagging system is the Streptavidin-Biotin interaction. Streptavidin is a protein that can bind 4 biotin molecules, the biotin molecule is bound to the molecule of interest and acts as a tag with a high affinity for the streptavidin column (Terpe, 2003). The tag is located at either the N- or C-terminus of the protein for purification and offers not only a method for purification but also a way to immobilise proteins to media, for example streptavidin-coated beads (Terpe, 2003; Michelson-Horowitz *et al.*, 2005). This system, when compared to other tagging methods, yields higher quantities of purified protein but the resin is costly and would become a large expenditure when carrying out large scale purifications on an industrial scale (Michelson-Horowitz *et al.*, 2005).

A different type of tagging for proteins is the use of reporter tags which are not commonly used for protein purification. Reporter tags enable protein-protein interaction to be studied as well as protein synthesis (Kimple and Sondek, 2004). While not an ideal protein purification tag, β-galactosidase is an ideal protein colour marker. It

catalyses the hydrolysis of its substrate 5-bromo-4-chloro-3-indolyl-β-D-galactoside (commonly referred to as X-gal) to produce 5-bromo-4-chloro-indoxyl, which dimerizes spontaneously creating a detectable blue colour (Kimpton and Emerman, 1992; Kimple and Sondek, 2004; Padmanabhan, Banerjee and Mandi, 2011). Horseradish peroxidase (HRP) is a highly used protein that is commonly used for Western blotting and immunoassays to amplify the photometric signals in these assays (Krainer and Glieder, 2015). HRP can be bound to antibodies and proteins and is detectable using a spectrometer (Kimple and Sondek, 2004). Alkaline phosphatase (ALP) is another useful reporter tag as it produces a purple colour when reacting with 5-bromo-4-chloro-3-indolyl phosphate (Kimple and Sondek, 2004). ALP is commonly used in enzyme linked immunosorbent assays (ELISA) and blotting methods. ALP is not commonly used for purification as yield is usually low and the matrix for purification cannot be easily reused, which increases cost for purification (Michelson-Horowitz *et al.*, 2005). There are a number of commercial kits readily available for conjugating ALP to a protein for reporting and it is often conjugated to primary antibodies for immunoblotting (Kimple and Sondek, 2004). A number of other tags are readily available which utilise different purification matrices, for example the maltose binding protein, c-myc tag and the FLAG peptide tag (Terpe, 2003; Kimple and Sondek, 2004). Different tags offer ways to increase protein solubility and improve protein stability, although this is often dependent on the protein of interest for purification.

1.3.2 DNA modifying enzymes

There are several different enzymes used for DNA modification, including but not limited to kinases, nucleases and ligases. These enzymes can be used in combination or on their own to achieve specific modifications to DNA and are common place tools in most molecular biology laboratories.

1.3.2.1. Endonucleases

Endonucleases, which cut within a DNA molecule that is provided, were first detected in the 1950s. Different rates of bacteriophage growth were noted in different bacteria infected with bacteriophages (Luria and Human, 1952). This role of restriction enzymes,

specifically Type I, in bacteriophage modification was shown at a biochemical level (Arber, 1965). There are four different groups (Type) of restriction endonuclease: Type I – multi subunit enzymes that randomly cut DNA away from the recognition site and are not used for DNA analysis (Roberts *et al.*, 2003). Type II – cut at the specific sequence to produce distinct DNA fragments and are routinely used for DNA analysis (Roberts *et al.*, 2003). Type III – need two sequences in opposite directions on the same piece of DNA to cut (Type III cut outside of the recognition site sequence and often do not fully digest the DNA) (Roberts *et al.*, 2003). Type IV – cleave modified DNA, these cleave 30 bases away from the recognition site and are not well defined (Roberts *et al.*, 2003). Type I and Type II restriction endonucleases can be further sub-categorised, with focus on the Type IIS and Type IIG subgroup for molecular and synthetic biology methodologies. The first Type II restriction enzyme, *HindII*, was purified from *Haemophilus influenza*, and could cleave T7 DNA (Smith and Welcox, 1970). The discovery of the recognition site sequence (Kelly and Smith, 1970) led to the many practical uses for type II restriction enzymes that are in place today.

1.3.2.2. Exonucleases

Exonucleases work differently from endonucleases and cleave nucleotides from the end of a polynucleotide chain. Each nucleotide is usually removed one at a time and exonucleases are vital for genome stability (Mason and Cox, 2012). There are a number of types of exonuclease which are named after their specific activities, this includes 5'-3', 3'-5' and poly (A) specific 3'-5' exonucleases (Mukherjee *et al.*, 2004). Exonucleases are also valuable tools in molecular biology, specifically after PCR to degrade primers or to degrade single stranded DNA from nucleic acid mixtures (Werle *et al.*, 1994; Nabavi and Nazar, 2005).

1.3.2.3. Kinases

Bacteriophage T4 polynucleotide kinase (Pnk) is another DNA modification enzyme that is routinely used in molecular biology. Pnk was first discovered in the 1960s when purified from *E. coli* infected with the bacteriophage T4 (Richardson, 1965). Pnk carries out the removal of 3' phosphate groups and the addition of phosphate to the 5' OH

terminus (Wang and Shuman, 2001; Wang, Lima and Shuman, 2002). Historically Pnk was used during cloning, its ability to remove phosphate and correctly prime ends of DNA for ligation has made it a useful biotechnology tool and this will be discussed in more detail in Chapter 5.

1.3.2.4. Nucleic Acid Ligases

Nucleic Acid (NA) Ligases are used in cloning for recircularization of plasmids and insertion of genes into plasmids to create recombinant DNA. Discovery of DNA (see Section 1.1.1) sparked the beginning of molecular biology and while this discovery is one of the most important, it is not the only discovery that has contributed to the molecular biology field and cloning techniques used today. The study of *E. coli* cells infected with λ phages, enterobacteria phages that infect *E. coli*, showed that a repair system was able to circularise the DNA (Gellert, 1967) and, thus, DNA ligases were discovered although at this time their uses in biotechnology were limited. The T3, T4, T7 and *E. coli* DNA ligases are all commercially available for DNA modification. The T4 ligases are usually used for blunt and sticky end ligation, nick repair, cloning and adapter ligation. One of the most common uses for T4 DNA ligase is recircularization of plasmids or addition of new DNA into plasmids as recombinant DNA. More recently the T4 RNA ligases isolated from bacteriophage T4 have been used in NGS for adapter ligation to miRNAs (Thomson, Bracken and Goodall, 2011). The T4 ligases *in vivo* are involved in DNA and RNA repair systems by repairing nicks in DNA and RNA by phosphodiester bond formation (Bullard and Bowater, 2006).

1.4. DNA and RNA repair systems

DNA and RNA repair are fundamental molecules involved in a number of reactions within an organism or cell (Tomkinson and Levin, 1997; Wilkinson, Day and Bowater, 2001). DNA damage can be caused either by metabolic activity or by external factors, and repair systems aim to mend the damaged DNA or RNA often ensuring cell survival (Hustedt and Durocher, 2017). There are different types of damage that can happen to the DNA structure; exogenous damage may be caused by oxidation, alkylation or base hydrolysis while external damage may be caused by chemicals, UV light, ionising radiation or increased temperature. Each of these damage the DNA in different ways;

oxidation, alkylation and ionising radiation result in either altered bases, abasic sites (no purine or pyrimidine at the position) or a single stranded break and are repaired via base excision repair (BER) (Houtgraaf, Versmissen and van der Giessen, 2006; Jackson and Bartek, 2009). Polycyclic aromatic hydrocarbons and UV light cause strand cross-linking and DNA adducts (where the DNA is bound to a cancer causing chemical) this damage is repaired by nucleotide excision repair (NER) (Houtgraaf, Versmissen and van der Giessen, 2006). A third type of repair pathway, mismatch repair (MMR) occurs when replication errors result in base mismatches or small insertions or deletions (Houtgraaf, Versmissen and van der Giessen, 2006). Ionising radiation and alkylation can also lead to cross links or double stranded breaks (DSBs) which require recombinational repair, either homologous recombination (HR) or non-homologous end joining (NHEJ) (Drabløs *et al.*, 2004; Houtgraaf, Versmissen and van der Giessen, 2006; Jackson and Bartek, 2009). HR results in accurate DSB repair while NHEJ is mutagenic. HR uses a DNA template that is not damaged and uses nucleotides to repair the original DNA sequence; NHEJ modifies the broken DNA ends and ligates the break, often generating insertions or deletions (Mao *et al.*, 2008)

Some of the types of damage that may happen to DNA can also happen to RNA, including alkylation and radiation damage (Bellacosa and Moss, 2003). RNA alkylation, also termed N- alkylation, where adducts are formed at the N- atoms in RNA bases, are repaired by BER or by oxidative demethylases (Drabløs *et al.*, 2004). One enzyme from *E. coli*, AlkB, and its human homologs have been identified as playing a key role in RNA alkylation repair by directly reversing the damage by demethylating methylated RNA bases (Bellacosa and Moss, 2003). The pathways for RNA repair are not as well studied or established in the literature, but one study noted it was possible to complement the RNA repair process in yeast by using T4 RNA ligase 1 and T4 polynucleotide kinase (Schwer *et al.*, 2004) indicating RNA repair in different organisms have shared evolutionary mechanisms.

The repair pathways mentioned for both DNA repair and RNA repair use a number of different enzymes, notably ligases in the final phosphodiester bond formation step, for

repair (Jackson and Bartek, 2009). These phosphodiester bond forming enzymes can be categorised into three different enzyme sub-classes based on the molecule repaired: DNA ligases, RNA ligases and mRNA capping enzymes. These ligases and mRNA capping enzymes are also part of a super family of nucleotidyl transferases. Ligases were originally classified as either DNA or RNA ligases, but as new enzymes have been identified it appears this classification system requires more work and often ligases are being grouped by their substrate specificity (Popow, Schleiffer and Martinez, 2012). In this body of work, groups of NA ligases will be described as Dnl-like, Rnl1-like or Rnl2-like families and are discussed in more detail below.

1.4.1. Nucleotidyl transferase conserved motifs

DNA ligases, RNA ligases and mRNA capping enzymes are often described as a super family of nucleotidyl transferases. Nucleotidyl transferases are characterised by 6 conserved peptide motifs: motifs I, III, IIIa, IV, V and VI, which play a role in the catalytic activity of the enzymes in this family (Shuman and Schwer, 1995; Doherty, 1999). These motifs vary significantly between the DNA ligases, RNA ligases and mRNA capping enzymes, including the T4 ligases (Table 1.1), which show little motif similarity. Motif I (K-x-H/D/N-G, where x refers to any amino acid) is essential for ligase function, it contains the lysine residue that binds the AMP that is essential for the ligation activity, and is found in all members of this family. The Shuman research group have carried out a number of mutagenesis studies (see section 3.1.3) comparing each of the T4 ligases and their motifs to homologous ligases. They identified conservation of all motifs in T4Dnl, motifs I, IV and V in T4Rnl1 and motifs I-V in T4Rnl2 (Table 1.1)(Shuman and Schwer, 1995; Ho and Shuman, 2002; Wang *et al.*, 2003; Yin, Ho and Shuman, 2003; Nandakumar *et al.*, 2004). DNA ligases and mRNA capping enzymes also have the addition of an oligonucleotide binding (OB) fold in the C-terminal region, also described as motif VI, which is missing in all members of the Rnl1-like and Rnl2-like families (Pascal, 2008).

Table 1.1. Conserved polypeptide motifs of the T4 NA ligases. The amino acid sequences for each conserved motif in T4Dnl, T4Rnl1 and T4Rnl2. Where there is no sequence (-), the ligase does not contain this motif. These sequences are conserved when the ligases are aligned against other Dnl- Rnl1- or Rnl2- like ligases.

Ligase	Motif I	Motif III	Motif IIIa	Motif IV	Motif V	Motif VI
T4Dnl	KADGAR	VLIDGELV	KFQVWDYVP	EGIIL	LYKFK	PIAIRLREDK
T4Rnl1	KEDGSL	-	-	EGYVA	HFKIK	-
T4Rnl2	KIHGTN	YQVFGEFA	FYVFDIIV	EGYVL	AIKCK	-

1.4.2. The ligation reaction

The ligation reaction carried out by DNA and RNA ligases has 3 main steps: enzyme adenylation, transfer of AMP to the 5' end of the DNA and formation of the phosphodiester bond (Figure 1.5) (Nandakumar, Shuman and Lima, 2006). Ligases require cofactors to join the strands of nucleic acids, with the cofactor either NAD^+ or ATP providing the adenylate group. Current literature suggests that functional DNA ligases using NAD^+ are only found in bacteria, while ATP-dependent ligases are abundant in all domains of life. The first step of the ligation reaction is enzyme adenylation. This step involves the formation of the ligase-AMP intermediate and requires magnesium as a cation for the reaction. The DNA/RNA ligase binds to the α-phosphorus of ATP (if ATP is the cofactor required by the ligase) (Shuman, 2009), generating pyrophosphate (Figure 1.5, Step 1)(Ho *et al.*, 2004; Chauleau and Shuman, 2013). The second step is AMP transfer, which can at times be problematic because it is the step closest to equilibrium and, as such, has a highly reversible nature (Figure 1.5, Step 2) (Cherepanov and De Vries, 2003). This step leads to the formation of the DNA/RNA-adenylate intermediate by transferring the AMP from the ligase-adenylate to the 5'-PO_4 of the DNA/RNA (Ho *et al.*, 2004; Shuman, 2009). The final step is referred to as strand joining, whereby the ligase catalyses the formation of the phosphodiester bond and releases AMP (Figure 1.5 Step 3) (Ho *et al.*, 2004; Popow, Schleiffer and Martinez, 2012).

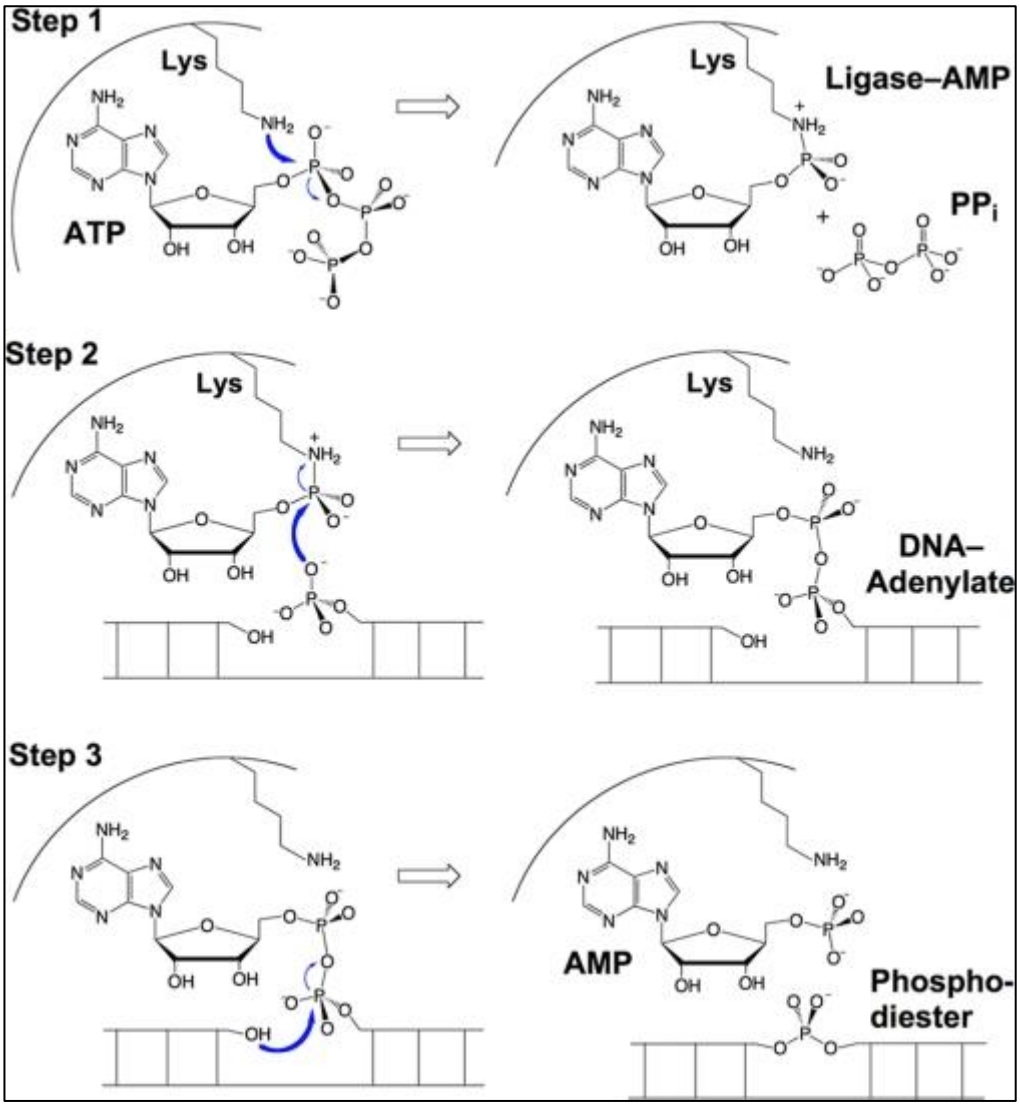


Figure 1.5. The 3-step ligation reaction pathway of T4 DNA and RNA ligases with ATP as the co-factor. Step 1: Ligase-AMP formation via the Lysine residue in motif I, pyrophosphate is released; Step 2: AMP transfer to create the 5' phosphate of the broken DNA/RNA strand, this is DNA adenylate; Step 3: nick sealing catalysis of the 5' phosphate and 3' OH via phosphodiester bond formation (Shuman, 2009).

1.5. Bacteriophage T4 Ligases

The bacteriophage T4 is a renowned source for commercially available enzymes, including polynucleotide kinase, DNA polymerase and DNA and RNA ligases. These enzymes have wide spread applications across molecular microbiology, genetics and gene expression, demonstrating how studying the T4 phage has been vital in developing biological concepts. The bacteriophage T4 has a genome of 168,903 bases, encoding 300 probable genes (Miller *et al.*, 2003). The number of genes coding for proteins is thought to be 289, although some genes code for more than one protein, and of the 289

possible proteins 156 have been characterised (Armstrong, Brown and Tsugita, 1983; Miller *et al.*, 2003). Three of these proteins are NA ligases: T4 DNA ligase (T4Dnl), T4 RNA ligase 1 (T4Rnl1) and T4 RNA ligase 2 (T4Rnl2).

1.5.1. T4 DNA ligase

The bacteriophage T4 has one DNA ligase (T4Dnl) which is coded for by gene 30, *gp30*. The molecular weight of T4Dnl is 55.3 kDa and it uses ATP as its co-factor (Armstrong, Brown and Tsugita, 1983). When T4Dnl was first purified, its capacity as a ligase was identified and subsequently it is now a highly used tool in molecular biology techniques (Weiss and Richardson, 1967). Mutational studies, where *E. coli* was infected with T4 phages containing a defective copy of T4Dnl, demonstrated that T4Dnl is required for phage growth (Gellert, 1967; Karam and Barker, 1971). In *E. coli* cells deficient in the bacterial host ligase, complementation of T4Dnl resulted in recovery of ligase function (Wilson and Murray, 1979). The crystallographic structure of a number of DNA ligases are known but there is no crystallographic structure available for T4Dnl (at the time of submission of this work a structure for T4Dnl was published, discussed in section 7.2); however, DNA sequencing has helped in identifying the conserved polypeptide motifs for this ligase. The OB fold in T4Dnl appears to be the region that provides the enzyme specificity for DNA ligases and mRNA capping enzymes and its conformational change is essential for step I of the enzymatic reaction in some DNA ligases. While the OB fold is absent in Rnl1 and Rnl2 families it is thought they may undergo a similar conformational change to T4 DNA ligase allowing the lysine in motif I to interact with ATP (Doherty, 1999; Sriskanda, 2002). T4Dnl can seal nicks in single stranded breaks when both strands are DNA, or RNA and DNA when the 3'-OH part of the break is RNA and the 5'-phosphate is DNA (Bullard and Bowater, 2006).

1.5.2. T4 RNA Ligase 1

T4Rnl1 is one of two RNA ligases from the bacteriophage T4, it has a molecular weight of 43.5 kDa and was first purified from *E. coli* infected with the bacteriophage T4 (Silber, Malathi and Hurwitz, 1972). It is part of a sub group of proteins called Rnl1-like proteins, which can be found across the plant and fungi domains (Ho and Shuman, 2002). T4Rnl1

is encoded by *gene 63* in the T4 genome and is involved in catalysing tail fibre attachment (TFA) during phage formation and nick sealing during infection of *E. coli* (Snopek *et al.*, 1977; Hall *et al.*, 1980). TFA involves a number of different proteins, T4Rnl1 performs the last step in attaching the long tail fibre to the baseplate (Wood *et al.*, 1978; Leiman *et al.*, 2003). The role of T4Rnl1 in this assembly process was further confirmed when attachment of tail fibres to tail-fibreless T4 particles increased 50-fold when T4Rnl1 was present compared to mutants absent in T4Rnl1 (Wood *et al.*, 1978; Leiman *et al.*, 2010). Its ligation activity is separate to its function in TFA (Miller *et al.*, 2003) and appears to be an adapted response mechanism utilised when the phage infects *E. coli* (El Omari *et al.*, 2006).

During bacteriophage infection with T4, *E. coli* cells activate two restriction enzymes – *Eco*prrI and a tRNA lysine specific anticodon nuclease (Penner *et al.*, 1995). These two enzymes are involved in recognition of the *stp* polypeptides found on the surface of the T4 bacteriophage (Penner *et al.*, 1995; Blanga-Kanfi *et al.*, 2006), resulting in cleavage of the tRNA anti-codon loop at the lysine residue, preventing further T4 infection (Amitsur, Levitz and Kaufmann, 1987). The T4 phage has evolved a system to counter this: T4 Pnk repairs the ends of the nicked tRNA and then T4Rnl1 seals the break in the anti-codon loop, allowing for further maturation of viral proteins and continued phage infection (Penner *et al.*, 1995; Uzan and Miller, 2010; Popow, Schleiffer and Martinez, 2012).

T4Rnl1 is notably different from the other T4 ligases, in that it is missing two of the conserved polypeptide motifs, III and IIIa (Table 1.1) (Figure 1.5). Site directed mutagenesis of specific conserved amino acids in T4Rnl1 showed essential residues located in the active site, notably Lys-99, Gly-102, Lys-118, Glu-227, Gly-228, Lys-240 and Lys-242. These residues are also conserved in the motifs in the active sites of DNA ligases, mRNA capping enzymes and T4Rnl2 (Wang *et al.*, 2003). Specifically the Lysine residues (99, 240 and 242) are essential for ligation but not for the phosphodiester bond formation, indicating their role in the first step of ligation – adenylation (Wang *et al.*, 2003). T4Rnl1 is able to ligate a range of DNA/RNA substrates, but in all instances the 3'-OH strand must be RNA (Bullard and Bowater, 2006).

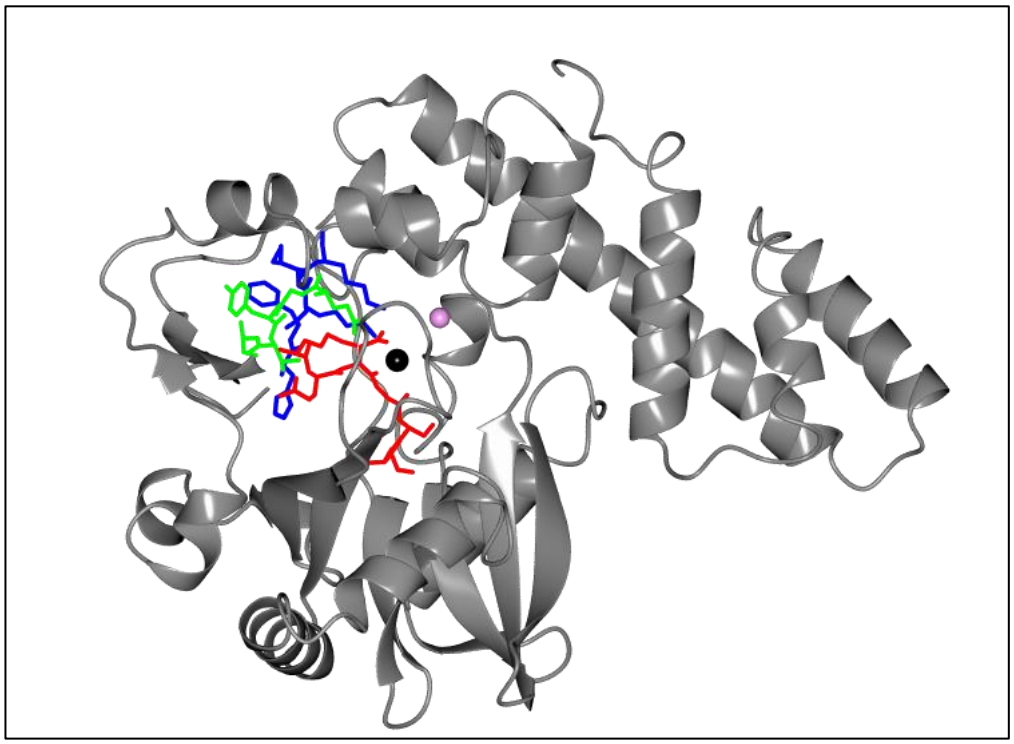


Figure 1.6. Crystallographic structure of T4Rnl1. T4Rnl1 with the nucleotidyl transferase conserved polypeptide motifs highlighted, motif I (red), motif IV (blue) and motif V (green), which form the nucleotide binding pocket. The metal ions calcium (pale pink sphere) and magnesium (black sphere) are located in the nucleic acid binding pocket. Illustrated using CCP4MG software (PDB 2C5U).

1.5.3. T4 RNA Ligase 2

T4Rnl2 is the second RNA ligase from the bacteriophage T4 and Rnl2-like proteins can be found in all phylogenetic domains (Shuman and Lima, 2004). It is encoded by the gene 24.1 and has a molecular weight of 37 kDa. The *in vivo* function of T4Rnl2 is still unknown, however, it is thought to be involved in guided RNA repair in mRNAs, encompassing mRNA repair as part of RNA ligation (Ho et al., 2004). T4Rnl2 is more similar to T4Dnl as it has all of the conserved polypeptide motifs for the nucleotide binding pocket, I, III, IIIa, IV and V (Figure 1.6) (Yin, Ho and Shuman, 2003; Shuman and Lima, 2004). Mutagenesis of conserved amino acids identified a number of essential residues: Glu-34, Asn-40, Arg-55, Glu-99, Phe-119, Asp-120, Glu-20 and Lys-209 (Yin, Ho and Shuman, 2003). Each of these amino acids are found in the conserved polypeptide motifs, or close to the nucleotide binding pocket. The N- terminal region is responsible

for steps 1 and 3 in the ligation reaction, but the C- terminal region is required for step 2 (Yin *et al.*, 2004).

The DNA/RNA substrate specificity of T4Rnl2 lies somewhere between T4Dnl and T4Rnl1. It is capable of ligating nicks between double stranded substrates where both strands either side of the nick are DNA – like T4Dnl; but T4Rnl2 can also do this with double stranded RNA substrates and substrates with combinations of DNA and RNA – like T4Rnl1 (Bullard and Bowater, 2006). When the substrates are combinations of DNA and RNA, T4Rnl2 joins RNA 3'-OH/DNA 5'-PO_4 (Bullard and Bowater, 2006; Chauleau and Shuman, 2013). As a molecular tool T4Rnl2 is often used in NGS adapter ligation, and a truncated version of T4Rnl2, amino acids 1-249, has been shown to ligate 5' adenylated adapters to 3' ends of RNA more effectively than T4Rnl1. Despite this, adapter ligation remains a problematic step in next generation sequencing of micro RNAs (Viollet *et al.*, 2011) and attempts to rectify this have included lengthening the adapters (Sorefan *et al.*, 2012) but other than truncating T4Rnl2 no attempts to modify the ligases have been made.

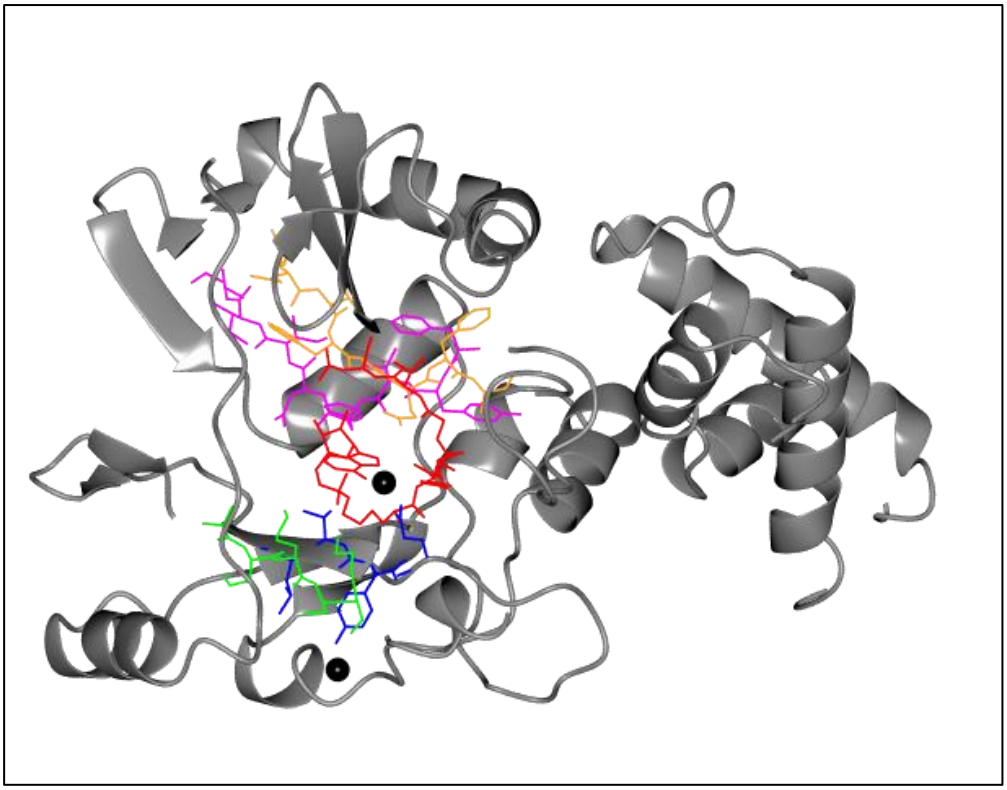


Figure 1.7. Crystallographic structures of T4Rnl2. T4Rnl2 with the adenylation domain motifs highlighted, motif I (red), motif III (orange), motif IIIa (pink), motif IV (blue) and motif V (green) The magnesium ions (black spheres) are located in the nucleic acid binding pocket. PDB file 2HVQ illustrated using CCP4MG software.

1.5.4. Problems with the T4 ligases

As stated above, the T4 ligases are valued molecular biology tools and structure bias with these ligases has been noted and become a well discussed issue. This is especially prevalent when ligating adapters for NGS (Zhuang *et al.*, 2012). The enzymes of choice for this ligation are usually either a combination of T4Rnl1 and T4Rnl2 or just T4Rnl1 used twice (Sorefan *et al.*, 2012). Substrate specificity experiments for both T4Rnl1 and T4Rnl2 have shown that both are capable of ligating a wider range of substrates than T4Dnl, and that the complementary strand could be either DNA or RNA for both of these ligases (Bullard and Bowater, 2006). Quantitative analysis of miRNAs, using pre-adenylated DNA adapters, showed that the T4 RNA ligases are not biased to sequence, but instead the secondary structure of the 3' end of DNA/RNA (Zhuang *et al.*, 2012). A number of studies have looked at this bias further and have primarily focussed on altering the adapters used for ligation in NGS. Using a pool of adapters, instead of a single sequence, improved ligation efficiency and reduced bias (Xu *et al.*, 2015). Another

major concern with the use of these ligases in NGS is the possibility of circularisation of the 3' adapter and concatamerisation of the input RNA, although this can be limited if there is a surplus of miRNA adapter and other factors, like temperature and time, remain constant (Hafner *et al.*, 2011; Sorefan *et al.*, 2012). For NGS, ligases have already undergone modification to recognise only adenylated RNA, which has helped in developing the necessary cloning techniques (Sorefan *et al.*, 2012). Some bias is expected when carrying out these sequencing techniques because of the nature of enzymatic reactions (Hafner *et al.*, 2011).

1.6 Research Aims

There has been a notable effort to reduce the bias seen in NGS adapter ligation, although it could be useful to tailor the bias in such a way to favour specific results. It may be possible with the use of molecular and synthetic biology techniques to create novel NA ligases with altered conserved motifs and additional domains that could have specific functions. By substituting specific amino acids in motif I, swapping conserved domains and creating chimeric ligases it may be possible to create NA ligases with improved phosphodiester bond formation with different or new substrates. It may also be possible to address differences in substrate specificity by mutating specific amino acids in motif I since the lysine residue in motif I is responsible for interactions with ATP and other residues in motif I interact directly with the DNA/RNA substrate (Nandakumar, Shuman and Lima, 2006).

The aims of this research were to modify the T4 RNA ligases resulting in novel ligases with novel or different substrate specificity and ligation activity. A total of 22 new versions of the T4 NA ligases are detailed in this body of work. The specific modifications to the genes or plasmids are detailed in individual chapters as a number of different techniques were used to create different variants. These new ligases include specific mutations to motif I in each of the T4 ligases, domain swapping/addition to the T4 RNA ligases and ligase-fusions with adenylate kinase, polynucleotide kinase and *Mycobacterium smegmatis* LigD; further background information about each of these enzymes can be found in the relevant chapter.

Chapter 2
Materials and Methods

Chapter 2: Materials and Methods

This chapter details the different experimental procedures used during this project. The bacterial strains and plasmids used, cloning procedures, purification methodologies and the ligation assay used to determine the activity and specificity of the ligases are included in this chapter. Specific methods for mutations and activity are included within the relevant results chapter.

2.1 Bacterial cell preparation and growth

2.1.1 Bacterial Strains

All bacterial strains used (Table 2.1) were grown either in Luria-Bertani (LB) liquid media or on inverted LB solid agar plates (Chapter 2.5.2) at 37 °C, unless stated otherwise. The LB liquid media contained 100 µg/ml Ampicillin, to determine positive transformant selection for protein purification procedures. Chloramphenicol was also added to final concentrations of 50 µg/ml.

2.1.2 Competent cell preparation

Both α-select and BL21 pLysS *E. coli* competent cells were made by streaking a stock of bacteria onto an LB agar plate (a selective plate containing chloramphenicol for BL21 pLysS cells), this was grown overnight at 37 °C. A 5 ml selective LB media was inoculated with a single colony from this plate and grown overnight at 37 °C with shaking at 200 rpm. Then a second set of two 50 ml LB selective media were sub-cultured using 500 µl of the overnight culture and grown at 37 °C with shaking at 200 rpm until the OD_{600} reached 0.5. The cells were pelleted by centrifugation at 4500 rpm using the Beckman Coulter Allegra 25R bench top centrifuge for 5 minutes at 4 °C, the supernatant was discarded. All steps were then performed at 4 °C, 40 ml of ice-cold TFB1 (Table 2.5) was used to resuspend the cell pellets. These were incubated on ice for 5 minutes then centrifuged at 4500 rpm for 5 minutes. The supernatant was discarded and the cells resuspended in 2 ml of ice-cold TFB2 (Table 2.5) and incubated on ice for 45 minutes. The cells were then quick frozen with liquid nitrogen and stored at -80 °C in 200 µl aliquots. Both Novagen and Bioline *E. coli* competent cells were used as per the

instructions provided by the manufacturer, with three different *E. coli* strains being used (Table 2.1). Once transformed the cells were plated out onto LB solid agar plates containing 100 µg/ml Ampicillin and 50 µg/ml Chloramphenicol (for DE3 strains) and incubated at 37 °C overnight.

Table 2.1. Details of the 3 different E. coli strains used throughout this project and where they were purchased from.

***E. coli* Strain**	**Brand**	**Genotype**
Rosetta (DE3) pLysS	Novagen	*F– ompT hsdS$_B$* (r_B^- m_B^-) *gal dcm* pRARE (Cam^R)
BL21 (DE3) pLysS	Invitrogen	*F- ompT hsdS$_B$* (r_B^- m_B^-) *gal dcm* (DE3) pLysS (Cam^R)
Alpha select bronze	Bioline	*F- deo$_R$ endA1 recA1 relA1 gyrA1 hsdR17(r_k^-, m_k^+) supE44 thi-1 phoA Δ(lacZYA-argF)U169 80lacZΔM15 λ-*

2.2 DNA procedures

2.2.1 The pET vector system

The pET system uses bacteriophage T7 machinery to control gene transcription (Studier and Moffatt, 1986). The pET system used for DNA and protein expression for all proteins was pET-16b; pET-16b has Ampicillin antibiotic resistance, His-Tag addition at the N-terminus, uses the T7 promotor and terminator to elicit gene control and can be induced using IPTG. All T4 nucleic acid (NA) ligases and variants of these ligases were modified and expressed using the pET16b vector, unless stated otherwise. The T4 NA ligases were already available on the pet16b vector (Figure 2.1 A, B and C) (Bullard and Bowater, 2006).

2.2.2 General Molecular Biology

Experimental protocols for gene modification and plasmid construction are given in specific chapters. Each set of plasmids containing the modified ligase genes was prepared differently and requires a separate methods section specific to the chapter to describe and explain the protocols required.

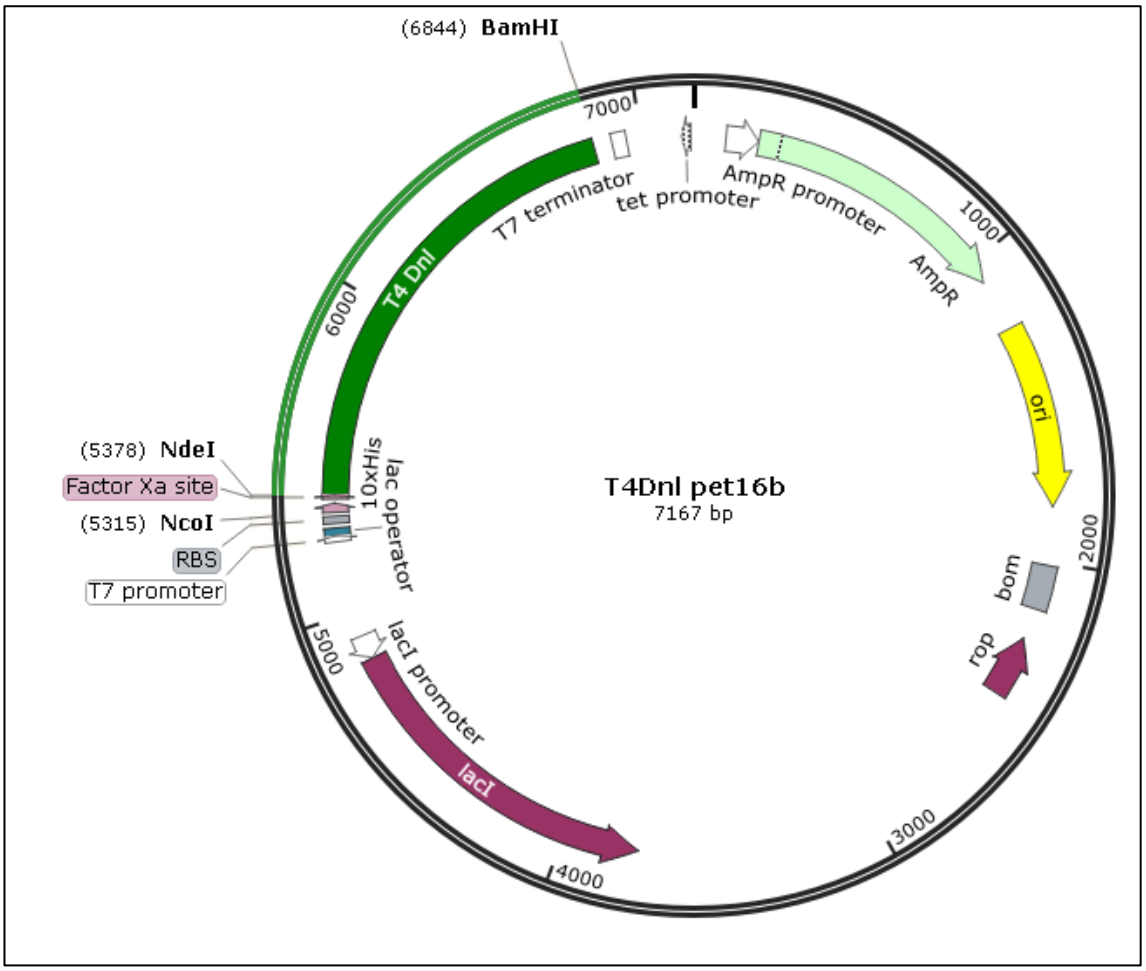


Figure 2.1. The construct already available to express T4DnI (Bullard and Bowater, 2006); on the pET16b plasmid denoted pRB255. The T4DnI gene (dark green) is between the *NdeI* and *BamHI* restriction sites. The plasmid map was generated using SnapGene software.

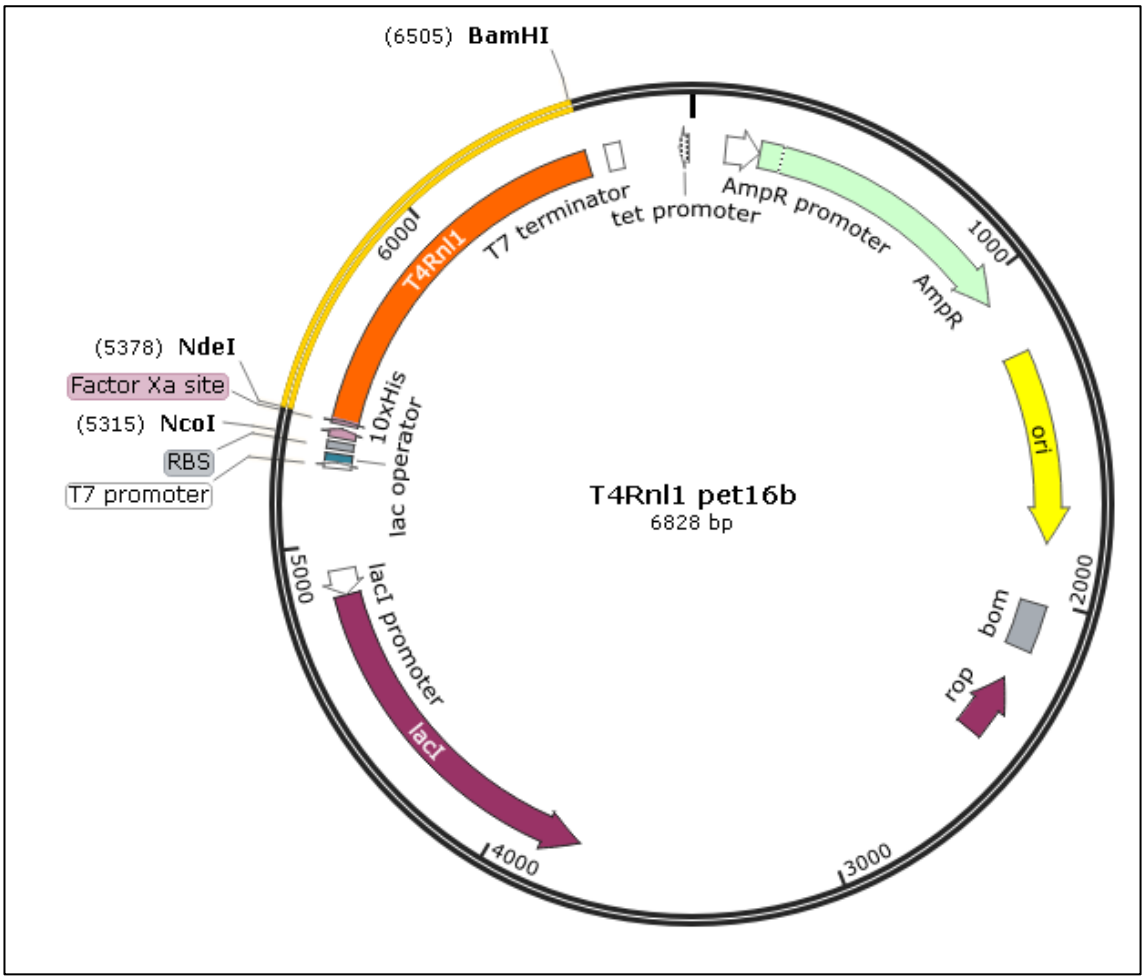


Figure 2.2. The construct already available to express T4Rnl1 (Bullard and Bowater, 2006); on the pET16b plasmid denoted pRB256. The T4Rnl1 gene (orange) is between the *NdeI* and *BamHI* recognition sites. The plasmid map was generated using SnapGene software.

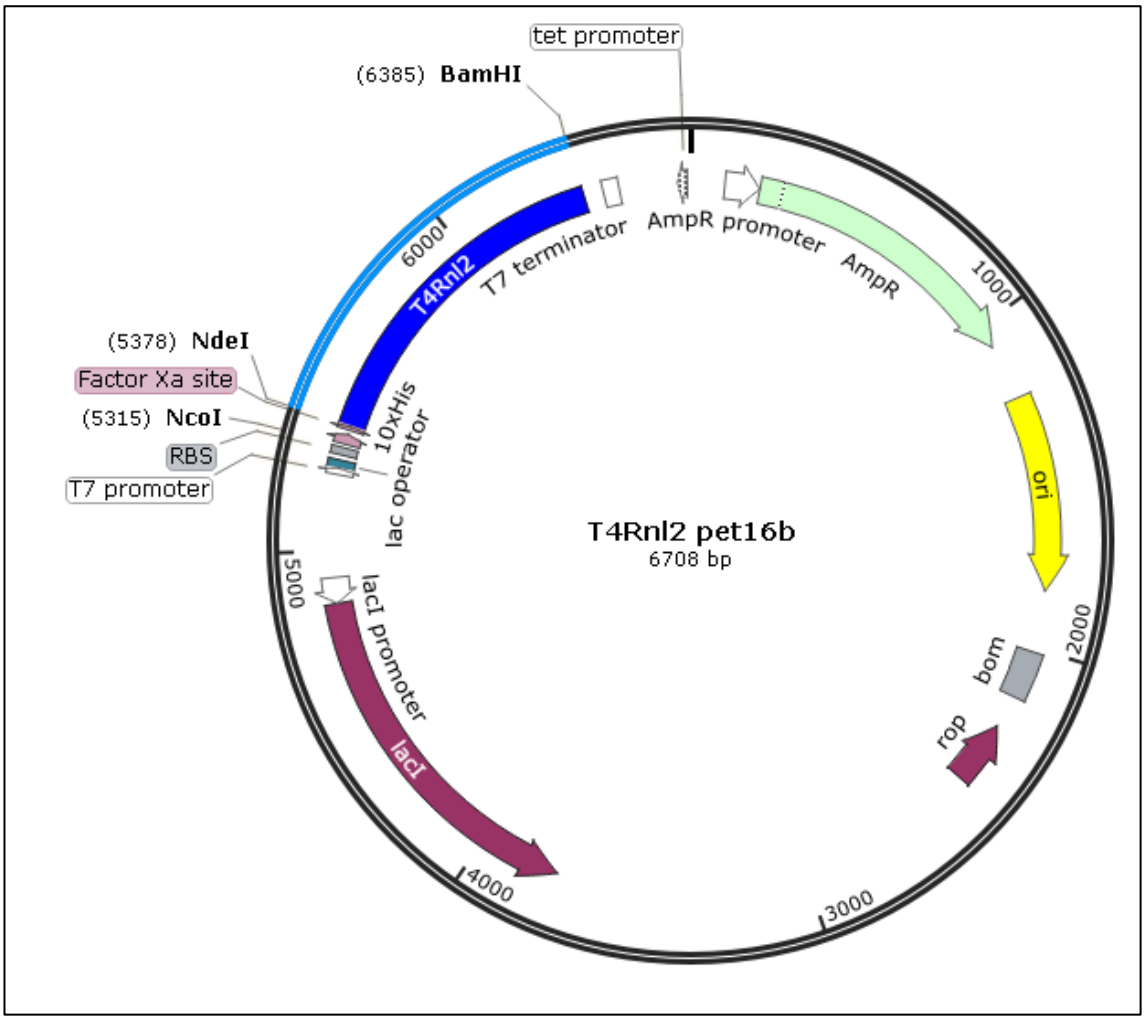


Figure 2.3. The construct already available to express T4Rnl2 (Bullard and Bowater, 2006); on the pET16b plasmid denoted pRB257. The T4Rnl2 gene (blue) is between the *NdeI* and *BamHI* recognition sites. The plasmid map was generated using SnapGene software.

2.2.3 The plasmids used and generated during the project.

Table 2.2. Description of the plasmids used and generated, with details of modification made to the plasmid. All of the plasmids generated in this body of work were confirmed by sequencing.

Plasmid Name	Description	Provider
pRB255	pET16b + T4Dnl	D. Bullard
pRB256	pET16b + T4Rnl1	D. Bullard
pRB257	pET16b + T4Rnl2	D. Bullard
pRB551	pUC57 + *E. coli* Adenylate Kinase (*Adk*)	Genscript
pRB552	pET16b vector + T4Dnl (pRB255) + *E. coli Adk* at the N- terminus,	This project
pRB553	pET16b vector + T4Rnl1 (pRB256) + *E. coli Adk* at the N- terminus	This project
pRB554	pET16b vector + T4Rnl2 (pRB257) + *E. coli Adk* at the N- terminus	This project
pRB555	pET16b vector + T4Rnl1 (pRB256) with motif I from T4Rnl2 (KEDGSL to KIHGTN)	This project
pRB556	pET16b vector + T4Rnl2 (pRB257) with motif I from T4Rnl1 (KIHGTN to KEDGSL)	This project
pRB557	pET16b vector + T4Rnl1 (pRB256) with OB domain from T4Dnl replacing the C-terminus (residues 242-374 from T4Rnl1 replaced with OB domain from T4Dnl residues 368 -487).	This project
pRB558	pET16b vector + T4Rnl2 (pRB257) with OB domain from T4Dnl replacing the C-terminus (residues 227-334 from T4Rnl2 replaced with OB domain from T4Dnl residues 368 -487).	This project
pRB559	pUC57 + T4 Polynucleotide Kinase (*Pnk)*	Genscript
pRB560	pET16b vector + T4Dnl (pRB255) + T4 *Pnk* at the N- terminus of the ligase	This project
pRB561	pET16b vector + T4Rnl1 (pRB256) + T4 *Pnk* at the N- terminus of the ligase	This project

pRB562	pET16b vector + T4Rnl2 (pRB257) + T4 *Pnk* at the N- terminus of the ligase	This project
pRB563	pET16b + T4Rnl1 (pRB256) D101H mutation	This project
pRB564	pET16b + T4Rn2 (pRB257) H37D mutation	This project
pRB565	pET16b + T4Dnl (pRB255) R164L mutation	This project
pRB566	pET16b + T4Dnl (pRB255) R164N mutation	This project
pRB567	pET16b + T4Rnl1 (pRB256) L104R mutation	This project
pRB568	pET16b + T4Rnl1 (pRB256) L104N mutation	This project
pRB569	pET16b + T4Rnl2 (pRB257) N40R mutation	This project
pRB570	pET16b + T4Rnl2 (pRB257) N40L mutation	This project
pRB571	pUC57 + *Mycobacterium smegmatis* LigD	Genscript
pRB572	pET16b + MSLigD (minus ligase) + T4Dnl (MSLigD with T4Dnl as ligase part on pRB255)	This project
pRB573	pET16b + MSLigD (minus ligase) + T4Rnl1 (MSLigD with T4Rnl1 as ligase part on pRB256)	This project
pRB574	pET16b + MSLigD (minus ligase) + T4Rnl2 (MSLigD with T4Rnl2 as ligase part on pRB257)	This project
pRB575	pET16b + full length MSLigD	This project
pRB576	pET16b + *Pnk* gene	This project

2.2.2 Plasmid extraction

Plasmids were extracted from 5 ml overnight liquid LB cultures with the appropriate selective antibiotic, these were inoculated from positive transformants. The Qiagen QiaPrep Spin Miniprep kit Quick Stick Protocol was used as per the manufacturer instructions. A Thermo Scientific Nanodrop One was used to determine the plasmid concentration.

2.2.4 Agarose gel electrophoresis

The agarose gels used for DNA and restriction enzyme digest analysis were 1 % TAE agarose gels run in 1x TAE buffer (Table 2.5). The gels were made by heating 1 g of Melford molecular biology grade agarose powder in 100 ml 1x TAE buffer to 100 ^{o}C in a conical flask with a sponge stopper. To visualise the gels 5 µl of 10 mg/ml ethidium

bromide was added to the TAE/agarose mixture and the gel poured into a casting tray with a comb insert in place. Once set the gel was positioned into a gel tank containing 1x TAE buffer, the DNA samples mixed with 5x loading dye were loaded into wells. Hyper Ladder 1kb (Bioline) was used as a ladder. The gels were run at 100 V, 95 mA for between 2 hours and 2.5 hours. The gels were visualised and imaged under UV light using the BioRad ChemiDoc XRS System.

2.2.5 Agarose gel purification

The required DNA fragments were cut out of the agarose gel and the DNA extracted using the Bioline Isolate II PCR & Gel Kit. The manufacturer's protocol was followed, the amount of DNA in each sample was quantified using the Thermo Scientific Nanodrop One.

2.2.6 Restriction enzyme digestion to analyse positive transformants

On all plasmids the new recombinant or mutated ligase genes were between the *NdeI* (upstream of the N- terminus) and *BamHI* (downstream of the C-terminus) cut sites. The enzymes required for restriction digestion were purchased from ThermoFisher Scientific. A stock solution containing 6.25 µl of *BamHI*, 6.25 µl of *NdeI* and 12.5 µl of FD Buffer (ThermoFisher Scientific) was made, 3 µl of this was added to 12 µl of up to 1 µg purified plasmid DNA and incubated at 37 °C for 2.5 hours. The DNA fragments from digestion were analysed using agarose gel electrophoresis where gels contained ethidium bromide and were visualised using UV.

2.2.7 Bacterial Transformation

Plasmids were transformed as follows: 50 µl of chemically competent *E. coli* cells were incubated with 1 ng of plasmid DNA on ice for 30 minutes. The cells were heat shocked at 42 °C for 30 seconds and immediately returned to ice for a further 5 minutes. To the cell suspension, 950 µl SOC media (Novagen) was added and the cells incubated at 37 °C, 300 rpm for 60 minutes to allow for recovery. From the recovered cells, 100 µl was spread onto selective LB plates and inverted for incubation at 37 °C overnight. When

using the Novagen/Bioline chemically competent cells manufacturer's instructions were followed.

2.2.8 DNA Sequencing

Plasmids were sequenced by Eurofins Genomics (i54 Business Park, Wolverhampton, UK) and required 100 ng/µl for accurate sequencing results, 15 µl was required for each sequencing reaction. The primers used were the T7 Promotor Forward primer and the T7 Terminator Reverse primer, these were supplied by Eurofins Genomics.

2.2.9 Oligonucleotides

Oligonucleotides were ordered with specific sequences from Eurogentec (The Square, Southampton, UK). Nuclease free water was used to resuspend the oligonucleotides to a final concentration of 100 µM and these were then stored at -20 °C.

Table 2.3. The oligonucleotide sequences from Eurogentec. These sequences in different combinations make the standard 8 substrates and the hairpin substrate required for the ligation assays; '+Flu' – addition of fluorescein or '+P' – addition of phosphate, denotes a specific modification at the 5' position. The 18 b and 15 b oligonucleotides were used to make a hairpin substrate which mimics the structures ligated *in vivo* by T4Rnl1.

Backbone	Oligo Name	Sequence
DNA	LigSub1 (8b) +Flu	5′ -GGC-CAG-TG- 3′
	LigSub2 (12b) + P	5′ -AAT-TCG-AGC-TCG- 3′
	LigSub1+2comp (20b)	5′ -CGA-GCT-CGA-ATT-CAC-TGG-CC- 3′
	DNAtRNALysSub2 (18b) + P	5′ -TTT-AAT-CAA-TTG-CGA-CCC- 3′
RNA	RNALigSub1 (8b) + Flu	5′ -GGC-CAG-UG- 3′
	RNALigSub2 (12b) + P	5′ -AAU-UCG-AGC-UCG- 3′
	RNALigSub1+2comp (20b)	5′ -CGA-GCU-CGA-AUU-CAC-UGC-CC- 3′
	RNAtRNALysSub1 (15b) + Flu	5′ -GGG-UCG-CAG-UUG-ACU- 3′

2.2.9.1 Oligonucleotide substrate preparation

A number of double stranded substrates were made (Table 2.3 and Table 2.4). The oligonucleotides were either 8 bases fluorescently labelled at the 5' end, 12 bases with a phosphate attached at the 5' part of the nick or 20 bases used as the complementary

strand. These 3 oligonucleotides were added together in a tube, at the ratio of 1:2:3, with 1.5 nmoles of LigSub1 (the fluorescently labelled strand) 3 nmoles of LigSub2 (strand with phosphate) and 4.5 nmoles of LigSub1+2 (complementary strand). This resulted in a final concentration of 15 µM LigSub1 in a final volume of 100 µl (Figure 2.4, Table 2.4), this concentration was the rate determining factor in the total amount of substrate that could be ligated. To the tube 10 µl of 10x TE Buffer (Table 2.5) was added. The tubes were heated to 95 °C for 5 minutes and left to cool overnight covered with aluminium foil to prevent inactivation of the fluorescein.

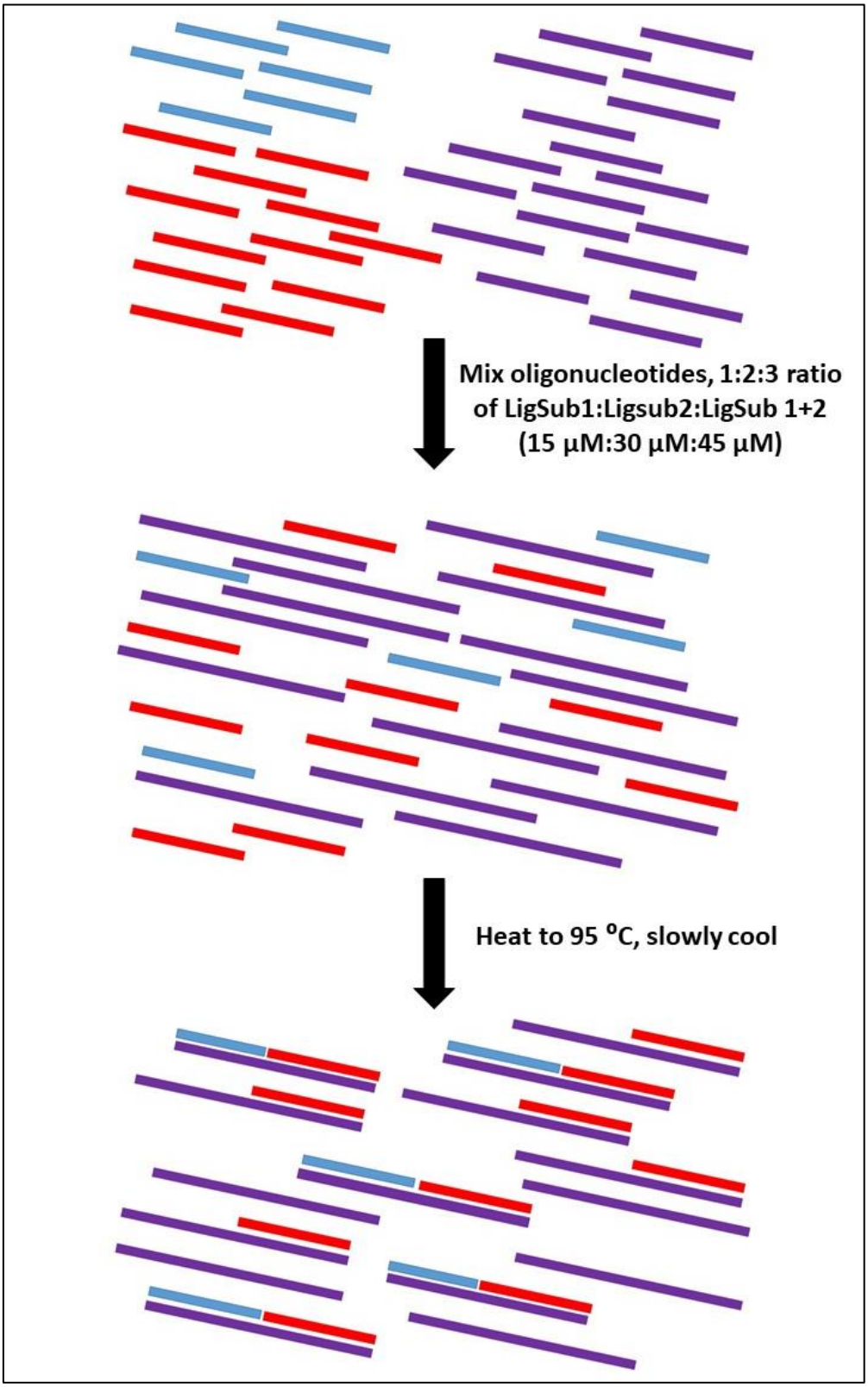
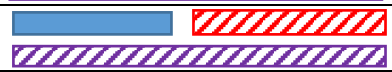


Figure 2.4. Ligation substrate formation at a ratio of 1:2:3 of the different oligonucleotides. The limiting oligonucleotide is LigSub1 (blue strand) which has fluorescein attached at the 5' position.

Table 2.4. The different DNA and RNA oligonucleotide components for the 8 standard double stranded substrates used in the end point ligation assay. The sequences for LigSub1 8 b (blue), LigSub2 12 b (red) and LigSub1+2 20 b (purple) were the same for each substrate (Table 2.3). The ligases were incubated with these 8 different substrates, buffer and ATP to ascertain whether they could ligate the nick between LigSub1 and LigSub2.

Name	Substrate (filled = DNA, hatched = RNA)
Substrate 1	
Substrate 2	
Substrate 3	
Substrate 4	
Substrate 5	
Substrate 6	
Substrate 7	
Substrate 8	

2.3 Protein Procedures

2.3.1 Protein expression in *E. coli*

For protein expression the plasmid DNA was transformed into BL21 pLysS (DE3) cells. Selective LB media was inoculated with 1/100th volume of overnight liquid culture, this was incubated at 37 °C, 180 rpm until an OD_{600} of 0.5 was obtained.

2.3.2. IPTG induction

The cultures containing the pET-16b vector were then induced with 0.5 mM IPTG and incubated at 30 °C for 4 hours. The cells were pelleted by centrifugation at 6000 rpm for 30 minutes in the Avanti J-26XP centrifuge with the 8.1000 rotor and the supernatant discarded. The whole cell pellets were stored at -20 °C until required.

2.3.3. Synthesis of Recombinant Proteins

The pellets were recovered and resuspended in 20 ml 20 mM Tris 100 mM NaCl (pH 7.5). The resuspended cells were lysed by sonication, 10 times for 10 seconds with a 30

second recovery period on ice. The soluble and insoluble fractions were separated by centrifugation at 40,000 rpm for 40 minutes at 4°C using the Beckman Optima XL-100K Ultracentrifuge with the 70Ti rotor. The supernatant was collected for protein purification; the pellet was discarded.

2.3.4. His-tag affinity chromatography for protein purification

A 5 ml HiTrap™ Chelating HP column (GE Healthcare Life Sciences, Amersham Place, Little Chlafont, UK) was used to purify the His-tagged proteins. The column was equilibrated using 25 ml water to remove the storage 20 % ethanol from the column, 50 ml 1x binding buffer, 50 ml 1x charge buffer and 25 ml 1x binding buffer (see Table 2.5 for all buffer components). The soluble fraction from the cell lysis had 100 µl removed (labelled as S1) and the rest of the sample was filtered through a 0.45 µm filter and loaded onto the column, the flow through from adding the soluble fraction was collected (labelled S2). For all further steps a 0.2 µm filter was used. A further 25 ml of 1x binding buffer was run through the column and the eluate collected (S3). The column was washed with 75 ml of 1x Wash Buffer and the eluate collected (S4). Finally 1x Elution Buffer was run through the column and the first 2 x 2 ml eluates collected (F1 an F2) and then 1 ml fractions collected (F3 – F10). This process was carried out at 4 °C. The column was then stripped using 50 ml 1x strip buffer and sealed for storage in 20 % ethanol (as per the manufacturer's instructions) at 4 °C. The fractions collected were analysed by SDS-PAGE and the fractions containing purified protein underwent buffer exchange.

2.3.5. Buffer exchange

The fractions containing protein were run through a disposable PD-10 column (GE Healthcare) to remove the imidazole and other salts. The PD-10 column was used as per the manufacturer's instructions for the gravity method. The proteins were eluted into 3.5 ml 20 mM Tris 100 mM NaCl (pH7.5) and flash frozen using liquid nitrogen before being stored at -80 ^{o}C.

2.3.6. Sodium dodecyl sulphate-Polyacrylamide Gel Electrophoresis

Sodium dodecyl sulphate (SDS) gels were made as per the Sambrook and Russel (2001) protocol. The gels were made to a 10 % polyacrylamide concentration. All purification fractions underwent SDS-PAGE to assess which fraction contained the purified ligase proteins. To visualise the proteins, the gels were stained using Expedeon InstantBlue for a minimum of 30 minutes.

2.3.7. Western Blot Analysis

The iBlotii system from Invitrogen was used, as per the manufacturer's instructions with the mini PVDF mini transfer stacks. After transfer for 7 minutes, the membrane was blocked using 1 % milk blocking solution (0.5 g milk powder in 50 ml TBS) for 1 hour whilst shaking at 4 °C. The membrane-milk solution had 2 µl of antibody added, the antibody used was monoclonal anti-poly histidine peroxidase antibody (from mouse). This was incubated overnight on the shaker at 4 °C. The following day, the membrane was washed using three washes with TBS-T and then two with TBS (Table 2.5). The membrane was covered with Thermoscientific Super Signal West Pico Chemillumiescent substrate, made up by combining 4 ml of stable peroxide solution and 4 ml of luminol/enhancer solution. The membrane was soaked for four minutes and imaged using the SynGene G:Box imager and GeneSys software.

2.3.8. Calculating protein concentration

The protein concentrations were calculated using the Protein Assay Microassay procedure from Bio-Rad and the absorbance measured at 595 nm using the Jenway 7315 Spectrophotometer. The absorbance of bovine serum albumin (BSA) samples with a known protein concentration (0, 1.25, 2.5, 5 and 10 µg/ml) were measured and plotted to create a standard curve. The absorbance of the purified protein samples were measured at different dilutions – 1/10, 1/100 and 1/1000 and the protein concentration obtained from the standard curve. The protein concentration was multiplied by the dilution factor to calculate the final protein concentration.

2.3.9. Protein storage

After glycerol was added to a final concentration of 20 % (v/v), the proteins were stored at -80 °C. To prevent the effects of freeze/thawing on the protein, the total amount was split into 500 µl aliquots, which were then further separated into 50 µl aliquots at first use.

2.4 Ligation Assay Analysis

Ligase activity assays were carried out to check the activity of all proteins. The assay assesses whether the ligase is able to create a bond between an 8-mer and 12-mer oligonucleotide (Figure 2.5). Formation of the 20-mer oligonucleotide was detected using FAM detection of the fluorescein attached to LigSub1. As such it was necessary to keep the substrates covered and away from light to prevent fluorescein inactivation.

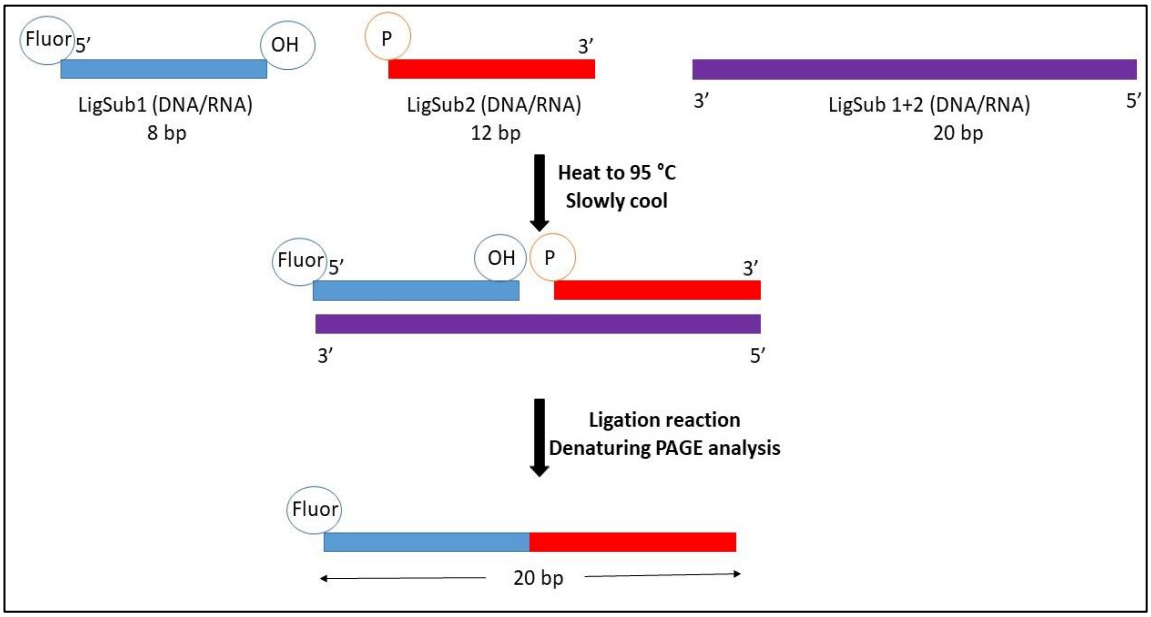


Figure 2.5. Schematic of the ligation reaction in the ligation assays. Detection of successful ligation requires the fluorescein at the 5' end of LigSub1. When ligation is successful the fluorescently tagged molecule increases from 8 bases to 20 bases in length.

2.4.1 End point ligation assay

For the end point assay, 8 different double stranded substrates made with different combinations of DNA and RNA were used to check ligation activity. For each substrate

reaction 1 µl 10x ligation buffer, 45 pmol of substrate and 70 pmol ligase were combined. Sterile nuclease free water was added to make the total volume 10 µl. The tubes were incubated at 37 °C for one hour while shaking at 250 rpm. At the end of the incubation 10 µl of 1x stop solution (Table 2.5) was added. The samples were heated to 100 °C for 5 minutes at the end of the incubation period and analysed using urea-PAGE denaturation, a 15 % urea-polyacrylamide gel was used (see Section 2.4.3).

2.4.2 Time course ligation assay

The time course ligation assay was performed using double stranded substrate 7. All ligases were able to ligate this substrate meaning substrate components were eliminated as a factor for differences in ligation. The assay was prepared as follows: 540 pmol of double stranded substrate, 6 µl 10x ligation buffer, 2 pmol T4Dnl or T4Rnl2 (or mutated versions of these ligases) or 600 pmol T4Rnl1 (or mutated versions of this ligase) or specified amounts. The final volume was made up to 60 µl using nuclease free sterile water. The reaction was incubated at 37 °C for one hour, with 5 µl aliquots removed and added to 5 µl 1x stop solution at the time points 0, 1, 2, 3, 4, 5, 10, 15, 20, 15 and 30 minutes. The different time point samples were heated to 100 ^{o}C for 5 minutes and analysed using urea-PAGE denaturation gels.

2.4.3 Urea-polyacrylamide gel electrophoresis analysis

The ligation reaction samples were analysed using urea-PAGE, the gels were made to 15 % polyacrylamide. Preparation for a single gel was as follows: 2.5 ml 30 % bis-acrylamide was combined with 0.5 ml 10x TBE and 2.4 g urea, this was shaken overnight until all urea has dissolved. Distilled water was added to make the total volume up to 5 ml. Finally 12.5 µl of 25 % ammonium persulphate and 7.5 µl of TEMED were added and the gels cast and left to set. The urea-PAGE gels were run in Bio-Rad Mini-PROTEAN© Tetra Cell Systems. The gels were heated for 30 min at 400 V using 0.5x TBE Buffer, the wells were cleaned with 0.5x TBE buffer to remove any excess urea before samples were loaded. The gels were run at 400 V until the dye front ran to the bottom of the gel, approximately 20 minutes. The gels were then imaged using a Typhoon FLA 9500 (GE Healthcare Life Sciences, Little Chalfont, Buckinghamshire, UK), with the FAM,

fluorophore detection setting. The images were then exported into ImageJ software, each lane was selected and the gel bands quantified using the measure tool. This provides 2 peaks, one for the ligated band and one for the non-ligated band. Each peak is selected with the wand tool to give a value for the contrast of each band compared to the background. These values were used to calculate the percentage of ligation which was converted into pmol – 540 pmol of substrate was used in total.

2.5 Media and Buffers

2.5.1 Media for bacterial growth

Luria-Bertani Media, pH 7: 1 % (w/v) tryptone, 0.5 % (w/v) yeast extract, 1 % (w/v) NaCl in 1L of distilled water. The pH was altered to 7 using 10M NaOH.

Luria-Bertani Agar: as above with 1.5 % (w/v) agar.

SOC Media: as provided by Novagen.

2.5.2 Media additions

The following stock solutions were prepared, filter sterilised using a 0.2 µm filter and kept at -20 °C.

Ampicillin, 100 mg/ml stock in sterile distilled water. Added to the media at a final concentration of 100 µg/ml.

Chloramphenicol, 34 mg/ml stock in ethanol. Added to the media at a final concentration of 50 µg/ml.

Isopropylthio-β-D-galactoside (IPTG), 100 mM stock solution. Added to the media at a final concentration of 0.5 mM.

Kanamycin, 100 mg/ml stock solution in sterile distilled water. Added to the media at a final concentration of 100 µg/ml.

L-Arabinose, 1M stock solution. Added to the media at a final concentration of 5 mM.

2.5.3 Buffers

Table 2.5. A list of buffers and their components used throughout this body of work.

Buffer	Chemical Components
10x SDS Running Buffer	248 mM Tris, 1.92 M glycine, 1 % w/v SDS Dilute to 1X for running SDS-PAGE gels
10x Tris-borate-EDTA (TBE)	0.9 M Boric Acid, 0.9 M Tris, 20 mM EDTA
10x Tris-EDTA (TE)	0.89 M Tris, 20 mM EDTA
1x Ligation Stop Solution	20 ml deionised formamide, 400 µl 0.5 M EDTA, 40 µl 1 M NaOH, 20 mg bromophenol blue
1x Tris-acetate-EDTA (TAE)	40 mM Tris, 20 mM Acetic Acid, 1 mM EDTA
4x SDS-PAGE Loading dye	200 mM Tris-HCl pH6.8, 8 % (w/v) SDS, 0.4 % (w/v) bromophenol blue, 40 % (w/v) glycerol, 400 mM β-mercaptoethanol
6x DNA loading dye	10 mM Tris-HCl pH 7.6, 0.03 % bromophenol blue, 0.03 % xylene cyanol FF, 60 % glycerol, 60 mM EDTA
4x Strip buffer	2 M NaCl, 80 mM Tris-HCl pH 7.9, 400 mM EDTA
8x Binding Buffer	4 M NaCl, 160 mM Tris-HCl pH 7.9, 40 mM Imidazole
8x Charge buffer	400 mM $NiSO_4$
8x Elution Buffer	2 M NaCl, 80mM Tri-HCl pH 7.9, 4 M Imidazole
8x Wash Buffer	4 M NaCl, 160 mM Tris-HCl pH 7.9, 480 mM Imidazole
1x Tris-buffered saline (TBS)	50 mM Tris pH 7.5, 150 mM NaCl
1x Tris-buffered saline + tween (TBS-T)	50 mM Tris pH 7.5, 150 mM NaCl, 0.1 % Tween 20
10x Ligation buffer	500 mM Tris-HCl, 100 mM MgCl, 100 mM DTT Dilute to 1x for ligation reaction
10x Ligation buffer + ADP	500 mM Tris-HCl, 100 mM MgCl, 100 mM DTT, 10 mM ADP Dilute to 1x for ligation reaction
10x Ligation buffer + ATP	500 mM Tris-HCl, 100 mM MgCl, 100 mM DTT, 10 mM ATP Dilute to 1x for ligation reaction
Ammonium molybdate-sulphuric acid solution	12.14 M ammonium molybdate tetrahydrate, 1 M sulphuric acid
Ferrous sulphate solution	388 mM Ferrous (II) sulphate, 2 drops 98 % H_2SO_4

Chapter 3
Domain Swaps and Mutational Analysis of Motif I in the T4 NA Ligases

Chapter 3: Domain Swaps and Mutational Analysis of Motif I in the T4 NA Ligases

3.1 Introduction

3.1.1 Conserved motifs in nucleic acid ligases

The conserved polypeptide motifs (I, IIIa, IIIb, IV, V and VI) that characterise the nucleotidyl transferase family are highly conserved, however only some of the amino acids in these motifs are conserved across the entire family (Figure 3.1) (Shuman and Lima, 2004). Before the discovery of T4Rnl2, motif I was denoted as KX(D/N)GXX (Shuman and Schwer, 1995; Timson, Singleton and Wigley, 2000), with the lysine residue identified as being responsible for creating the Lys-AMP intermediate necessary for ligation (Shuman and Schwer, 1995). Motif I in Rnl2-like ligases does not follow this KX(D/N)GXX model (motif I in T4Rnl2 specifically is KIHGTN) and motif I is now described as KXXGXX (Popow, Schleiffer and Martinez, 2012).

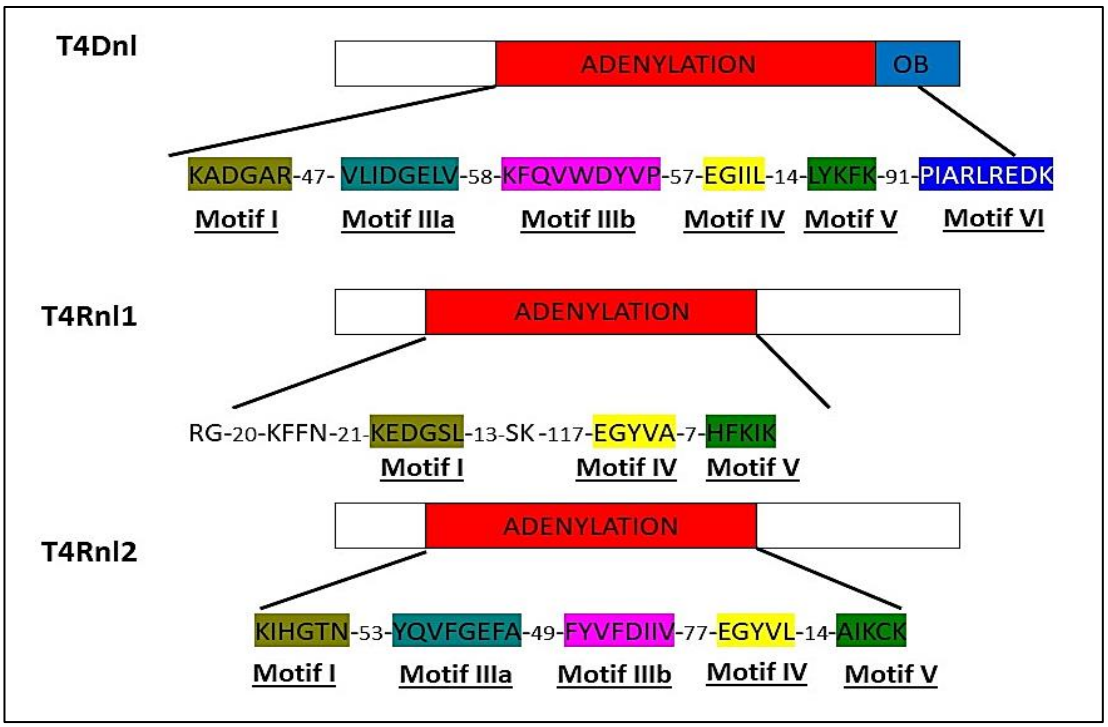


Figure 3.1. The T4 NA ligases and their conserved polypeptide motif amino acid sequences. The conserved amino acid residues are shown in each motif for T4Dnl, T4Rnl1 and T4Rnl2. Amino acids are designated following standard single letter abbreviations the numbers between motifs indicate the number of amino acids between the motifs.

These motifs are essential for the ligation reaction and line the nucleotide binding pocket. Motifs IIIa and IV in T4Rnl2 provide a hydrophobic environment which traps the nucleotide leaving the α-phosphate visible for interaction with the lysine in motif I (Shuman and Lima, 2004). The literature for T4Rnl1 often does not include a motif III or IIIa. One study stated motif IIIa, TYLDGDEI, was present when aligned against DNA ligases (Wang *et al.*, 2003), however this motif in T4Rnl1 is not in close proximity to the binding pocket (Figure 3.2) and most literature states only motifs I, IV and V are present in T4Rnl1.

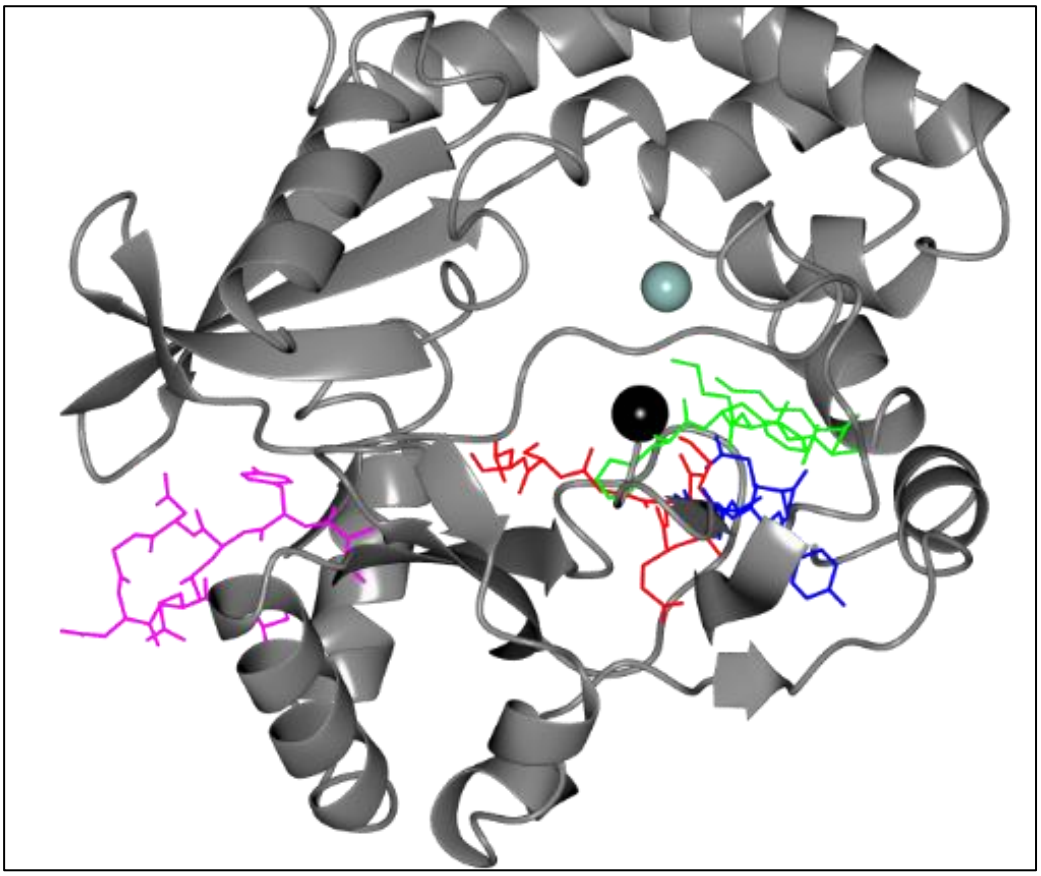


Figure 3.2. The nucleotide binding pocket in T4Rnl1. The conserved motifs are highlighted as follows: motif I (red), proposed motif IIIa (pink) (Shuman and Schwer, 1995), motif IV (blue), and motif V (green) with magnesium (black sphere) and calcium (teal sphere); PDB file 2C5U illustrated using the software CCP4MG. Mg^{2+} ions are essential for ligation of nucleic acids and ensure correct orientation of the phosphate, while high concentrations of calcium are essential for crystallisation of T4Rnl1 and do not directly interact with the ligase (El Omari *et al.*, 2006).

The conserved motifs that make up the nucleotidyl transferase family can be open to some interpretation. It is very clear that motifs I, III, IIIa, IV, V and VI (OB domain) are highly conserved amongst members of the family – DNA ligases, RNA ligases and mRNA capping enzymes. It is important to consider each of these subsets individually and assess residue conservation amongst Dnl-like proteins, Rnl1-like proteins and Rnl2-like proteins (Figures 3.3, 3.4 and 3.5).

3.1.2 Previous mutational analysis of motif I in the T4 ligases

Both nucleotidyl transferase domains found in T4Rnl1 and T4Rnl2 are located at the N-terminus and contain α-helices that flank antiparallel β-sheets (El Omari *et al.*, 2006). The conserved motifs in T4Rnl1 are motif I, IV and V; there is a notable absence of motifs III and IIIa. The crystallographic structure of T4Rnl1 has been solved and the amino acid residues involved in nucleotide binding and phosphodiester bond formation have been identified (El Omari *et al.*, 2006). The ligation reaction requires enzyme-adenylate and RNA-adenylate intermediates and the motifs present in T4Rnl1 are involved in the formation of these intermediates (Wang *et al.*, 2003). Not only are the Lys99 and Gly102 residues in motif I conserved, also conserved are Glu227 and Gly228 in motif IV and Lys240 and Lys242 in motif V that are vital for T4Rnl1 function. Glu227 interacts with calcium in the solved crystallographic structure, this calcium comes from the high concentrations needed in the crystallisation buffer and the site that would usually be occupied by Mg^{2+} (El Omari *et al.*, 2006; Wang, Schwer and Shuman, 2006). The stacking action of this residue and Tyr 246 increases the Van der Waals contact and contribute to the catalysis of steps 2 and 3 in the ligation reaction (Figure 1.5) (Wang, Schwer and Shuman, 2006). Lys240 and Lys242 are needed for step 1 of the ligation reaction (Figure 1.5) and interact with the phosphate of ATP (Wang, Schwer and Shuman, 2006). There is homology between the side chains of the residues conserved in T4Rnl1 and amino acids required for DNA ligases and mRNA capping enzymes but this is low (Shuman and Schwer, 1995).

The Asn40 and Glu34 residues essential for T4Rnl2 function are not conserved in T4Rnl1. There are five C-terminal residues that are crucial for T4Rnl2 activity: Lys35, Glu99, Phe119, Glu204 and Lys 227; these are unique to T4Rnl2 and are necessary for the C-

terminal interaction with RNA during the ligation reaction (Ho *et al.*, 2004). The N-terminal of T4Rnl2 contains seven residues crucial for function: Tyr5, Arg33, Lys54, Gln106, Asp135, Arg155 and Ser170 (Nandakumar *et al.*, 2004). The crystal structure of T4Rnl2 bound with AMP identified Tyr5 as the residue that makes contact with the adenosine ligand and forms a number of bonds with other residues in the motifs of T4Rnl2 (Nandakumar *et al.*, 2004). Ser170 and Glu34 interactions help stabilise conformation in motif I allowing adenylate binding and reaction catalysis (Nandakumar *et al.*, 2004). Arg33, Asp135 and Arg155 form a network of interactions that stabilise the nucleotidyl transferase binding pocket (Nandakumar *et al.*, 2004). Lys54 and Gln106 do not interact with the AMP phosphate but instead form part of T4Rnl2's RNA binding surface (Nandakumar *et al.*, 2004).

When considering the structural similarities between the two T4 RNA ligases it is notable that only T4Rnl2 has all of the conserved motifs of nucleotidyl transferases and could be described as being more similar to T4Dnl. For all 3 of the T4 NA ligases where they have the same motifs, there are some amino acids that are always conserved therefore it is likely that these conserved motifs are indicative of a shared nucleotidyl transfer mechanism (Shuman & Lima, 2004).

3.1.3 RNA ligase alignments and suggested mutations to motif I

The motif I residues are well conserved among Dnl-like, Rnl1-like and Rnl2-like proteins. When comparing motif I in both T4Rnl1 and T4Rnl2, the histidine and aspartate residues are always conserved in these two different groups of proteins. Comparing histidine to aspartate, it is noted that the two amino acids have some similarities and differences: histidine is often a polar amino acid, it can become positively charged at pH 5 (Rötzschke *et al.*, 2002) and has been described as an amino acid that does not substitute well with other amino acids (Betts and Russell, 2003). By contrast, aspartate is polar and charged usually preferring to be on the surface of proteins (Betts and Russell, 2003). These differences in properties may have an effect on the DNA or RNA substrates that the T4 RNA ligases can work with. Previously, when studying T4Rnl2, the Histidine (H) at position 37 was changed to both alanine (A) and aspartate (D) creating new variants of

the ligase T4Rnl2 H37A and H37D (Ho and Shuman, 2002). Of these two new ligases, only H37D was still able to function as a ligase and circularise a single stranded RNA product (Ho and Shuman, 2002). It was also concluded from this study that the histidine is not important for the first step of adenylation in ligation, but rather for the second step where the phosphodiester bond is formed (Ho and Shuman, 2002; Yin, Ho and Shuman, 2003).

These observations raised the question of whether this change from histidine to aspartate would affect the type of double stranded substrate the T4Rnl2 could repair and whether the same would be true for T4Rnl1 when changing the aspartate at position 101 to histidine. When considering the amino acids in motif I in the 6th position in the motif – arginine for T4Dnl, leucine for T4Rnl1 and asparagine for T4Rnl2, it was again noted that these residues are always conserved (Figures 3.3, 3.4 and 3.5). The residues in this position in the T4 DNA and RNA ligases interact directly with the ribose sugar (Shuman & Lima, 2004). Therefore, it was hypothesised that substituting these residues with other amino acids (Table 3.1) could result in a change in interaction with the ribose sugar, ultimately changing the substrate profiles of the T4 NA ligases.

```
T4Dnl       MILKILNEIASIGSTKQKQAILEKNKDNELLKRVYRLTYSRGLQYYIKKWPKPGI-ATQS     59
T4Rnl1      -MQELFNNLMELCKDSQRKFFYS--DDVSASGRTYRIFSY-NYA-SYSDWLLPD------     49
T4Rnl2      ----MFKKYSSLENH----------YNSKFIEKLYSLG------LTGGEWVAREKIHGTN     40
                                                                 Motif I

T4Dnl       FGMLTLTDMLDFIEFTLATRKLTGN-AAIEELTGYIT--DGKKDDVEVLRRVMMRDLECG    116
T4Rnl1      ----------------------------------------------------ALECR     54
T4Rnl2      FSLIIERDKVTC-------AKRTGPILPAEDFFGYEIILKNYADSIKAVQDIM----ETS     89


T4Dnl       ASVSIANKVWPGLIPEQPQMLASSYDEKGINKNIKFP--AFAQLKADGARCFAEVRGDEL    174
T4Rnl1      GIMFEMDGEKPVRIASRPMEKFFNLNENPFTMNIDLNDVDYILTKEDGSLVSTYLDGDEI    114
T4Rnl2      --------------AVVSYQVFGEFAGPGIQKNVDYGDKDFYVFDIIVT-----------    124
                          Motif IIIa       Motif IIIb/Motif I

T4Dnl       DDVRLLSRAGNEYLGLDLLKEELIKMTAEARQIHPEGVLIDGELV-YHEQVK-KEPEGLD    232
T4Rnl1      L----FKSKGSI-------K--------SEQALMANGILMNINHHRLRDRLKELAEDGFT    155
T4Rnl2      ------TESG----------------------------------------------------    128
                                              Motif IIIa

T4Dnl       FLFDAY-PENSKAKEFAEVAESRTASNGIANKSLKGTISEKEAQCMKFQVWDYVPLVEIY    291
T4Rnl1      ANFEFVAPTNRIVLAYQEMK------------IIL--LNVRE-----NETGEYISYDDIY    196
T4Rnl2      ----------------------------------------------DVTYVDDY    136
                                                        Motif IIIb

T4Dnl       SLPAFR----LKYDVRFSKL------EQMTSGYDKVILI-------------ENQVVNNL    328
T4Rnl1      KDATLRPYLVERYEIDSPKWIEEAKNAENIEGYVAVMKDGSHFKIKSDWYVSLHSTKSSL    256
T4Rnl2      MMESFC----NTFKFKMAPLLGRGKFEEL-------IK-----------------LPNDL    168
                                    Motif IV    Motif V

T4Dnl       DEAKVIYKKYIDQGLEGIILKNIDGLWENARSKNLYKFKEVIDVDLKIVGIYPHRKDPTK    388
T4Rnl1      DNPEKLFKTIIDGASDDLKA-----MY----ADDEYSYRKIEAFETTYLK--------YL    299
T4Rnl2      DSVVQDYNFTVDHA--GL-VDANKCVWNAEAKGEVFTAEG-----YVLKPCYPS----WL    216
                                         Motif V    Motif IV

T4Dnl       AGGFILE-----SECGKIK---------VNAGSGLKDKAGVKSHELDRTRI---------    425
T4Rnl1      DRALFLVLDCHNKHCGKDRKTYAMEAQGVAKGAGMDHLFGIIMS----------------    343
T4Rnl2      RNGNRVPIKCKNSKFSEKKKSDKPIKAKVELSEADNKLVGILACYVTLNRVNNVISKIGE    276
                 Motif V

T4Dnl       ------MENQNYYIGKILECEC-NGWLKSDGRTDYVKLFLPIAIRLREDKTKANTFEDVF    478
T4Rnl1      -------LYQGYDSQEKVMCEI-EQNFL-----KNYKKFIPEGY----------------    374
T4Rnl2      IGPKDFGKVMGLTVQDILEETSREGITLTQ--ADNPSLIKKELVKMVQDVLRPA-WIELV    333
                                                  Motif VI

T4Dnl       GDFHEVTGL                                                       487
T4Rnl1      ---------                                                       374
T4Rnl2      -------S-                                                       334
```

Figure 3.3. Clustal Omega alignment of the T4 ligases to each other. The three T4 ligases T4Dnl, T4Rnl1 and T4Rnl2 do not align well to each other, this is highlighted by a lack of alignment with the conserved polypeptide motifs - motif I (dark yellow), IIIa (light blue), IIIb (purple), IV (yellow), V (green) and VI (blue). Motif IIIb (pink) of T4Rnl2 aligns to motifs I (dark yellow) of both T4Dnl and T4Rnl1 and motif IV (yellow) of T4Rnl2 doesn't align to the two other ligase sequences.

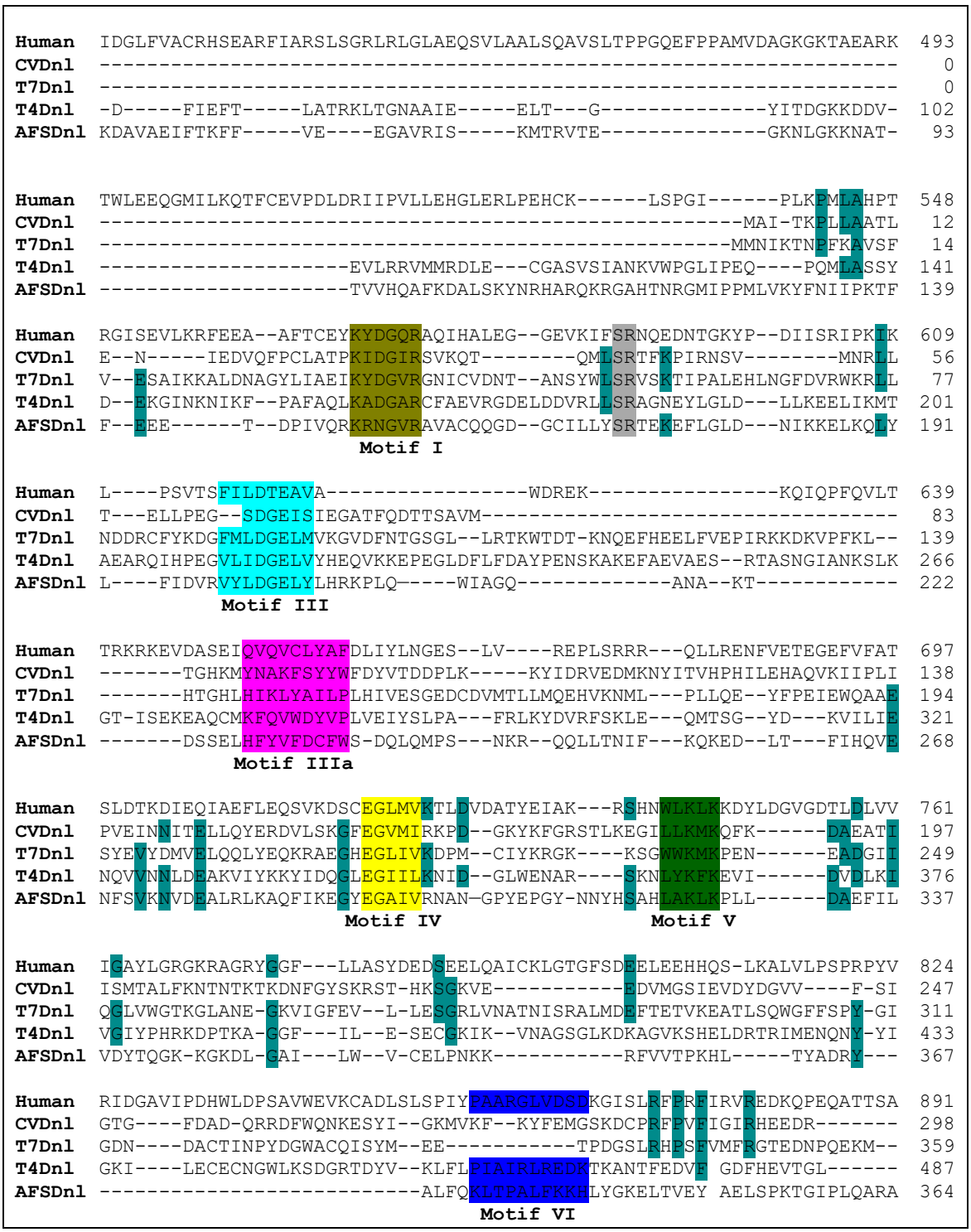


Figure 3.4. Sequence alignment for T4Dnl and T4Dnl-like proteins. Alignment of *E. coli* bacteriophage T4 DNA ligase (T4Dnl), Human DNA ligase I residues 268 – 419 (Human), *E. coli* bacteriophage T7 DNA ligase (T7Dnl), African Swine fever DNA ligase (AFSDnl) and *Chlorella virus* ATP-dependent Dnl (CVDnl). The conserved polypeptide motif I (dark yellow), IIIa (light blue), IIIb (purple), IV (yellow), V (green) and VI (blue). The 100 % conserved residues (grey) and highly conserved resides (teal) are shown. Alignments were performed using Clustal Omega, https://www.ebi.ac.uk/Tools/msa/clustalo/, amino acids are designated following standard single letter abbreviations.

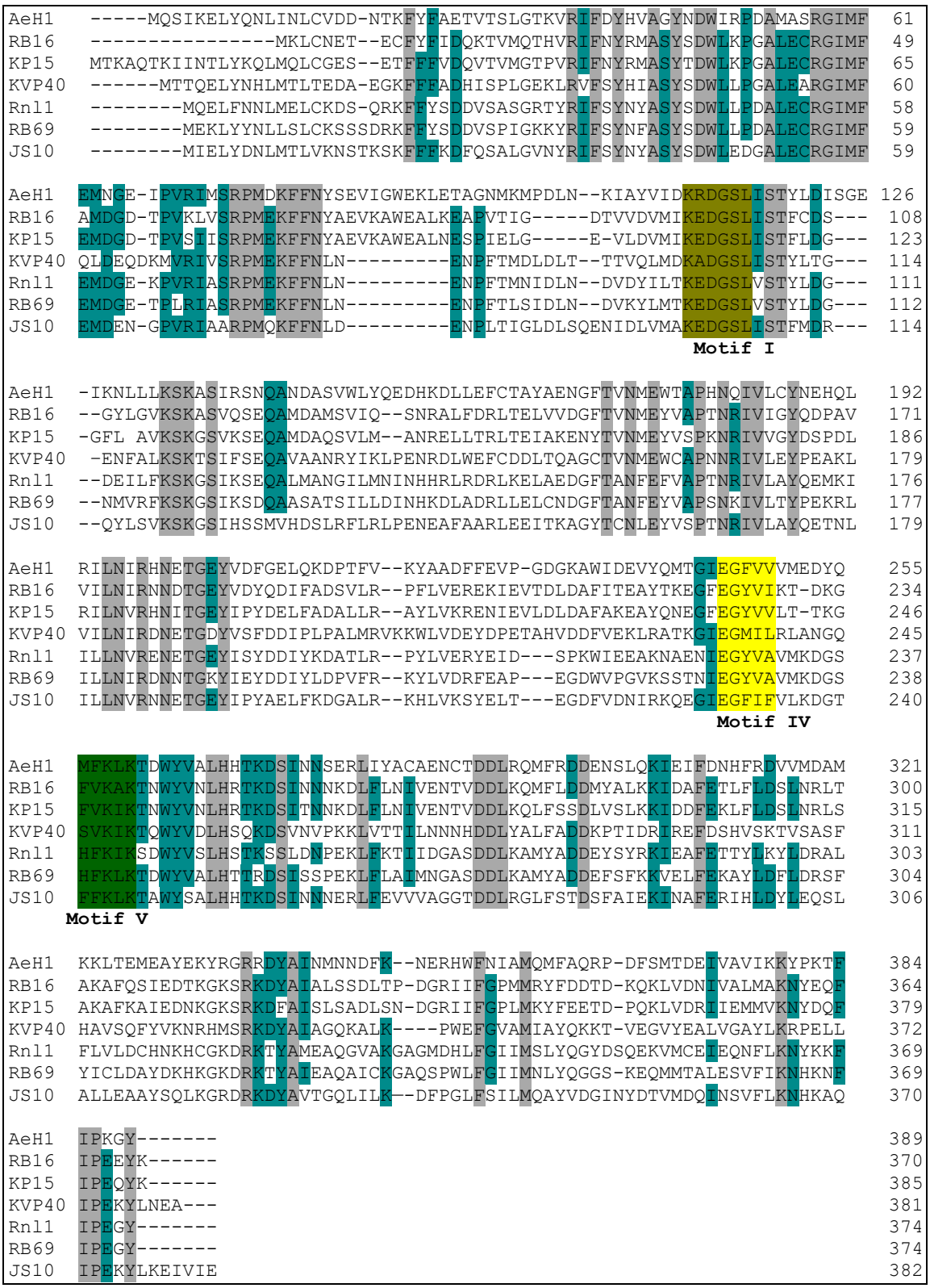


Figure 3.5. Sequence alignment for T4Rnl1 and T4Rnl1-like proteins. Alignment of *Aeromonas virus* (AeH1) RnlA, *E.coli bacteriophage* 16 (RB16) RnlA, *Klebsiella bacteriophage* 15 (KP15) RNA ligase I, *Vibrio bacteriophage* 40(KVP40) RnlA, *E. coli bacteriophage* T4 RNA ligase I (Rnl1), *E. coli bacteriophage* 69 (RB69) RnlA and *E. coli bacteriophage* JS10 (JS10) RnlA. The alignment shows conserved polypeptide motif sequences for motif I (dark yellow), motif IV (yellow) and motif V (green), 100 % conserved residues (grey) and highly conserved residues (teal). Alignments performed using Clustal Omega, https://www.ebi.ac.uk/Tools/msa/clustalo/, amino acids are designated following standard single letter abbreviations.

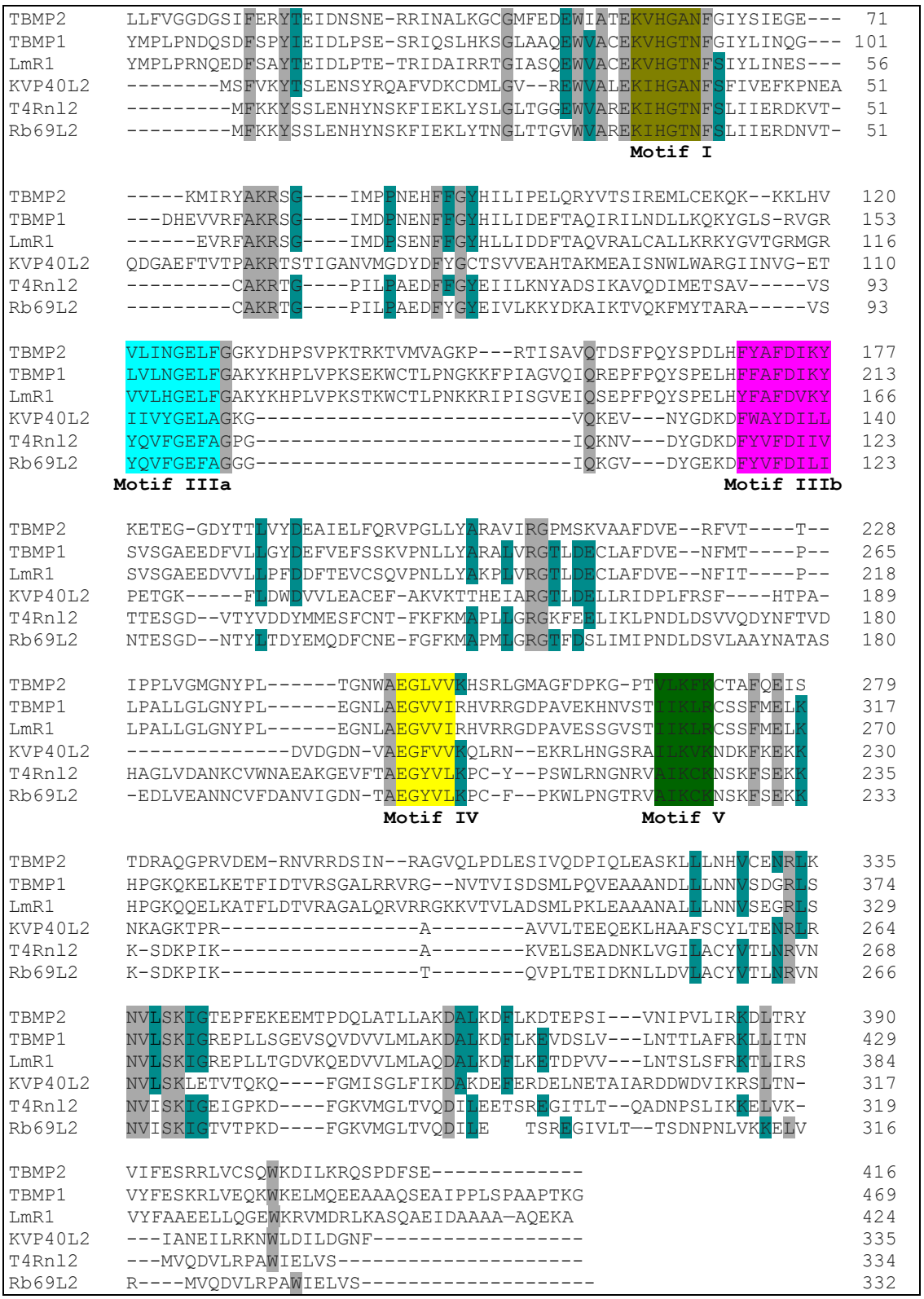


Figure 3.6. Sequence alignment of T4Rnl2 and T4Rnl2-like proteins. Alignment of *Trypanosoma brucei* RNA editing ligases 1 and 2 (TBMP1 and TBMP2), *Leishmania major* RNA editing ligase 1 (LmR1), *Vibrio bacteriophage* 40 RNA ligase 2 (KVP40L2), *E.coli bacteriophage* T4 RNA Ligase 2 (T4Rnl2), *E. coli bacteriophage* 69 RNA ligase 2 (RB69L2)**.** The alignment shows the conserved motifs I (dark yellow), IIIa (light blue), IIIb (purple), IV (yellow), V (green) and VI (blue). The 100 % conserved residues (grey) and highly conserved residues (teal). Alignments performed using Clustal Omega, https://www.ebi.ac.uk/Tools/msa/clustalo/ amino acids are designated following standard single letter abbreviations.

3.1.4. Golden Gate Cloning and the OB domain swaps

A noticeable difference between the T4 RNA ligases and T4Dnl is the OB domain, which contains motif VI, as this domain is missing from both T4Rnl1 and T4Rnl2. The motif in the OB domain is not involved in nick sealing but does contribute to self-adenylation (Timson, Singleton and Wigley, 2000). As the OB domain is always missing from Rnl1- and Rnl2-like proteins, it was decided that addition of this to T4Rnl1 and T4Rnl2 may yield ligases with improved adenylation capacity, which could improve ligation yield. To move the OB domain, a technique called Golden Gate Cloning (see Section 1.2.3) was used to create the recombinant plasmids.

3.2 Changes to the T4 nucleic acid ligases

Motif I has been well studied and its involvement in phosphodiester bond formation well tested. It was, therefore, sensible to begin alterations to the T4 ligases within this motif. The T4 NA ligases were aligned to each other and Dnl-, Rnl1- or Rnl2-like proteins (Figures 3.3, 3.4 3.5 and 3.6) using the Clustal Omega software (available at http://www.ebi.ac.uk/Tools/msa/clustalo/). The different ligase sequences were taken from www.uniprot.org and the ligases were selected randomly while trying to ensure a good spread of different ligases from different organisms. After performing amino acid alignments of the T4 RNA ligases it was proposed that the aspartate (D101) and histidine (H37) residues in motif I for T4Rnl1 and T4Rnl2, respectively, are always conserved (Figures 3.4 and 3.5) and might be responsible for the differences seen in substrate specificity of the two T4 RNA ligases. An equivalent residue in motif I for DNA ligases is not as highly conserved (Figure 3.3).

3.2.1 Site Directed Mutagenesis of pRB255, pRB256 and pRB257

The DNA sequence for motif I in T4Dnl is 5’ AAAGCTGATGGAGCTCGG 3’, the motif I for T4Rnl1 is 5’ AAAGAAGACGGGTCTTTG 3’ and the motif I for T4Rnl2 is 5’ AAGATTCACGGCACAAAT 3’. For each of the different amino acid substitutions and to swap motif I (Table 3.1), site directed mutagenesis PCR was carried out (Table 3.2). The primers were designed to introduce amino acid changes to motif I, from the original amino acid to the amino acid seen in another ligase (see Table 3.1). The site directed mutagenesis was carried out using the New England Biolabs (NEB) Q5-site directed

mutagenesis kit. The NEB protocol provided was followed, for PCR the annealing temperatures see Table 3.2. The plasmids used for mutagenesis were pRB255, pRB256 and pRB257 (T4Dnl, T4Rnl1 and T4Rnl2, respectively) (Bullard and Bowater, 2006).

Table 3.1. The altered motif I sequences for the different T4 NA ligases. The part of the sequence that will be changed is highlighted in red.

New Ligase Name	T4 Ligase before mutation	New Motif I sequence 5' – 3'
D101H	T4Rnl1	AAAGAA**C**ACGGGTCTTTG
H37D	TRnl2	AAGATT**G**ACGGCACAAAT
R164L	T4Dnl	AAAGCTGATGGAGCT**TTG**
R164N	T4Dnl	AAAGCTGATGGAGCT**AAT**
L104R	T4Rnl1	AAAGAAGACGGGTCT**CGG**
L104N	T4Rnl1	AAAGAAGACGGGTCT**AAT**
N40R	T4Rnl2	AAGATTCACGGCACA**CGG**
N40L	T4Rnl2	AAGATTCACGGCACA**TTG**
T4Rnl1M1	T4Rnl1	**CGTGAAAAGATTCACGGCACAAATTTC**
T4Rnl2M1	T4Rnl2	**CTAACAAAAGAAGACGGGTCTTTGGTA**

Table 3.2. The primers required for amino acid substitutions to motif I sequences. The primer sequences were generated using the NEB Base Changer tool, available at http://nebasechanger.neb.com/.

Primer Name	Primer Sequence	Annealing Temperature
D101HF	AAC-AAA-AGA-A**c**A-CGG-GTC-TTT-G	57 °C
D101HR	AGA-ATA-TAA-TCA-ACA-TCG-TTT-AAA-TC	
H37DF	TGA-AAA-GAT-T**g**A-CGG-CAC-AAA-TTT-CTC-ATT-G	66 °C
H37DR	CGA-GCT-ACC-CAC-TCC-CCA	
R164LF	TGA-TGG-AGC-T**tt-g**TG-TTT-TGC-TG	57 °C
R164LR	GCT-TTT-AAC-TGA-GCA-AAG	
R164NF	TGA-TGG-AGC-T**aa-t**TG-TTT-TGC-TGA-AGT-TAT-TAG	57 °C
R164NR	GCT-TTT-AAC-TGA-GCA-AAG	
L104RF	AGA-CGG-GTC-T**cg-g**GT-ATC-AAC-TTA-TTT-AG	60 °C
L104RR	TCT-TTT-GTT-AGA-ATA-TAA-TCA-ACA-TCG	
L104NF	AGA-CGG-GTC-T**aa-t**GT-ATC-AAC-TTA-TTT-AG	58 °C
L104NR	TCT-TTT-GTT-AGA-ATA-TAA-TCA-ACA-TC	
N40RF	TCA-CGG-CAC-A**cg-g**TT-CTC-ATT-GAT-TAT-TGA-G	56 °C
N40RR	ATC-TTT-TCA-CGA-GCT-ACC	
N40LF	TCA-CGG-CAC-A**tt-g**TT-CTC-ATT-GAT-TAT-TG	56 °C
N40LR	ATC-TTT-TCA-CGA-GCT-ACC	
T4Rnl1M1F	CGG-CAC-AAA-TTT-CTC-AAC-TTA-TTT-AGA-CGG-TGA-TG	59 °C
T4Rnl1M1R	TGA-ATC-TTT-TCA-CGA-ATA-TAA-TCA-ACA-TCG-TTT-AAA-TCG	
T4Rnl2M1F	CGG-GTC-TTT-GGT-ATC-ATT-GAT-TAT-TGA-GCG-TGA-TAA-AGT-GAC	67 °C
T4Rnl2M1R	TCT-TCT-TTT-GTT-AGA-GCT-ACC-CAC-TCC-CCA-CC	

The PCR product was treated with a kinase, ligase, *DpnI* (KLD) mix and then transformed into Bioline α-select bronze cells as per the manufacturer's instructions. The cell suspension was spread onto LB-agar plates containing 100 µg/ml Ampicillin and incubated at 37 °C overnight. On the following day colonies were selected for overnight culture in 10 ml LB selective media.

3.2.2. Golden Gate Cloning to add the OB domain to the T4 RNA ligases

Addition of the OB fold from T4Dnl was carried out using the GGC enzyme *BsaI*, the *BsaI* recognition site was added into the gene sequences at the beginning of the T4 RNA ligases, at the position where the OB fold would be inserted and at the end of the OB fold allowed for insertion of the OB domain (Figure 3.7). The primers were designed to include a 4 nucleotide overhang sequence for the fragments to enable ligation of fragments in the correct order (Table 3.3) (Appendix 1). The vector used was the MEGGA plasmid (Marcus Edwards Golden Gate Assembly) pBad vector induced using arabinose, with a C–terminal His tag, obtained from Dr Marcus Edwards. This vector was selected as it already contained the *BsaI* cut sites required to insert the new recombinant genes. Primers were designed to amplify regions of the T4 ligases at the gene level with the addition of the B*saI* cut site.

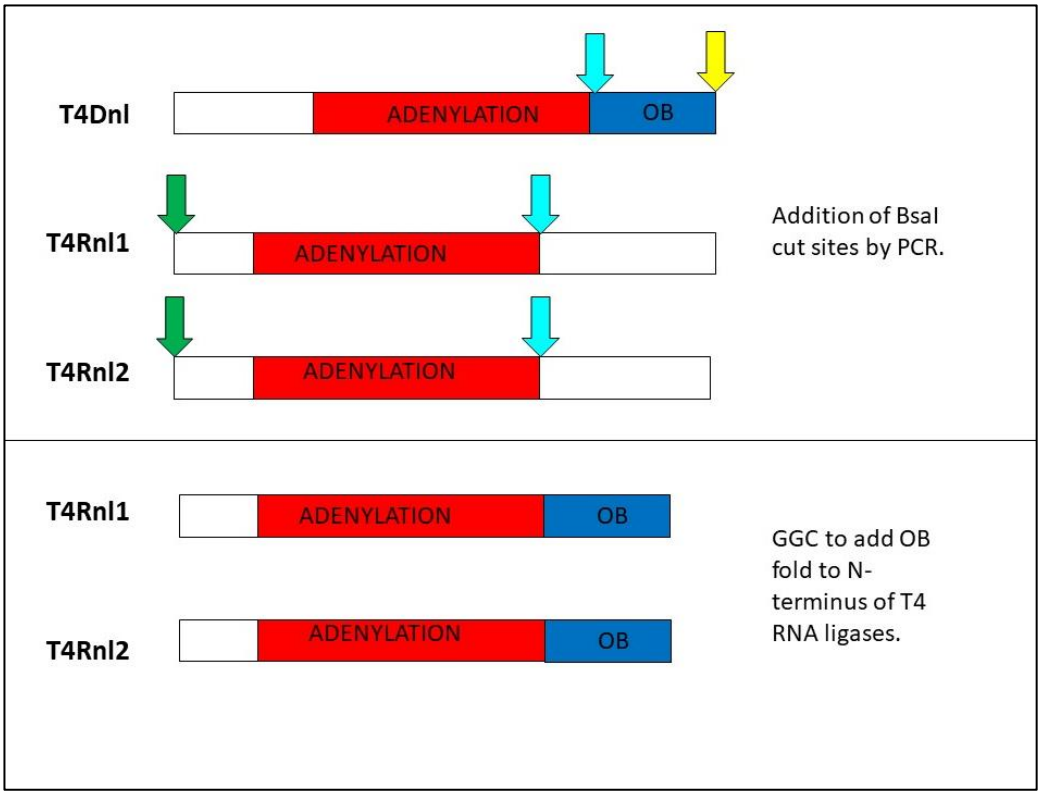


Figure 3.7. Addition of the OB domain to the T4 RNA ligases. Sequences designed so that the C-terminal sequence from T4Dnl replaced the C-terminal sequences for both T4Rnl1 and T4Rnl2. The arrows represent addition of a *BsaI* cut site, the arrow colours correspond to the overhang sequences that ensured the correct order of the fragments.

Table 3.3. The primers designed for addition of the BsaI cut sites.

Primer	Name	Sequence (5'-3')
T4 Dnl OB fold Forward	T4DnlBsaIOBF	TGG-TCT-CTG-AAG-TAA-TTG-ATG -TTG-ATT-TAA-AAA-TTG-TAG-GAA-TTT-ATC-C
T4 Dnl OB fold Reverse	T4DnlBsaIOBR	TGG-TCT-CTG-ATG-GGA-TCC-TCA-TAG-ACC-AGT-TAC-CTC-AT
T4 Rnl1 Forward	T4Rnl1BsaIF	TGG-TCT-CTA-CCC-ATG-CAA-GAA-CTT-TTT-AAC-AAT-TTA-ATG-GAA-CTA-TG
T4Rnl1 Reverse + OB	T4Rnl1BsaIOBR	TGG-TCT-CTC-TTC-CTT-AAT-TTT-AAA-ATG-AGA-ACC-ATC-TTT-CAT-CAC-AG
T4 Rnl2 Forward	T4Rnl2BsaIF	TGG-TCT-CTA-CCC-ATG-TTT-AAA-AAG-TAT-AGC-AGT-CTT-GAA-AAT-CAT-TAC
T4 Rnl2 Reverse + OB	T4Rnl2BsaIOBR	TGG-TCT-CTC-TTC-CTT-GCA-TTT-AAT-CGC-TAC-ACG-ATT-TCC
MEGGA_HIS_F	MEGGA backbone His tag Forward	CAT-CAT-CAC-CAT-CAC-CAT-TGA-GTT-TAA-ACG-GTT-TC
MEGGA_RBS_R	MEGGA backbone Ribosome binding site Reverse	GGG-TAT-GTA-TAT-CTC-CTT-CTT-AAA-GTT-AAA-CAA-AAT-TAT-TTC

The following reactions were set up to add the *BsaI* cut site (Appendix 1) and amplify regions of the T4 RNA ligases, the MEGGA backbone and to also amplify only the OB domain from T4Dnl. The following were combined: 25 µl NEB Q5 Hot start High Fidelity 2x Master Mix, 2.5 µl 10 µM Forward Primer, 2.5 µl 10 µM Reverse Primer, 1 µl of 1-25 ng DNA (T4 RNA ligase or T4 Dnl for the OB Fold) and 18 µl water. For the primers and PCR conditions see Tables 3.3 and 3.4.

Table 3.4. The PCR conditions for GGC fragment amplification.

Reaction Step	Temperature (°C)	Time (s)	
Initial Denaturation	98	30	
Denaturation	98	10	Repeat for 30 cycles.
Annealing	58 – RNA ligases and OB fold, 62 for MEGGA backbone	30	
Extension	72	300	
Final Extension	72	300	

After the PCR, *DpnI* digestion was carried out to remove the original template, as per the ThermoFisher *DpnI* manual. The reaction was scaled up 4x to incorporate higher concentrations of DNA fragments for the GGC reaction. PCR clean-up was carried out using the Bioline Isolate II PCR and Gel Kit as per the instructions provided. The DNA was eluted using molecular grade nuclease free sterile water instead of the elution buffer provided. The DNA concentration of the samples was determined using Nanodrop Quantification. For the GGC reaction ratios of 1:3:3 for the MEGGA vector : RNA ligase : OB domain were added to 1.5 µl T4 DNA ligase buffer, 1.5 µl cut smart buffer, 1 µl *BsaI* and 1 µl ThermoFisher T4DNA ligase and underwent the cycling conditions in Table 3.5. The product from the Golden Gate reaction was transformed into Bioline α-select bronze cells (see Section 2.2.6).

Table 3.5. GGC cycling conditions.

Reaction Step	Temperature (°C)	Time (min)	
BsaI Digestion	37	3	Repeat for 25 cycles.
T4 Ligation of Strands	16	4	
BsaI Deactivation	50	5	
T4 DNA Ligase Deactivation	80	5	

The assembled T4Rnl1OB and T4Rnl2OB genes were then transferred into the pET16b vector system by addition of the *NdeI* and *BamHI* cut sites (N- and C- terminus respectively) using PCR and the Q5 Site Directed Mutagenesis Kit (New England Biolabs) as per the manufacturer's instructions. For the primers and annealing temperatures see Table 3.6. The PCR fragments were then digested using *NdeI* and *BamHI* and ligated into the pET16b vector (see 2.2.5). The ligation product was transformed into α-select bronze cells (Bioline) (see Section 2.2.6).

Table 3.6. Primers and Annealing temperatures to add the T4 RNA ligases + OB domains to the pET16 vector system. The bold and underlined bases indicate the introduction of *NdeI* (CATATG) and *BamHI* (GGATCC) cut sites.

Primer Name	Primer Sequence 5'-3'	Annealing Temperature
557+NdeI F	tcgt**CATATG**caagaacttttaacaatttaatggaacta tgtaaggatt	65 °C
557+BamHI R	ccga**GGATCC**tcatagaccagttacctcatgaaaat	
558+NdeI F	tcgt**CATATG**tttaaaaagtatagcagtcttgaaaatcat tacaactctaaattta	65 °C
558+BamHI R	ccga**GGATCC**tcatagaccagttacctcatgaaaat	

3.2.3. Plasmid Isolation and Mutation Confirmation

The plasmids were isolated using the Qiagen QiaPrep Spin Miniprep kit (see Section 2.2.2) as per the manufacturer's instructions. Samples were sent to Source Bioscience (Nottingham Business Park, Nottingham, NG86PX), for Sanger sequencing, the results were analysed using the BioEdit software. Once the amino acid substitution was confirmed the new plasmids were transformed into competent BL21 pLysS cells (see Section 2.2.6) for protein expression and purification. The motif swap plasmids pRB555, pRB556, pRB557 and pRB558 were also confirmed using colony PCR.

3.2.4 Colony PCR

Colony polymerase chain reaction uses a single colony from an organism to amplify DNA fragments. The technique uses a single colony and does not require the isolation of pure DNA (Cao *et al.*, 2009). The bacterial DNA was prepared by isolating one bacterial colony from the selective LB agar plate. This colony was resuspended in 10µl of sterile water and then heated for 5 minutes at 100 °C. To 12.5 µl 2x My Taq Red Mix (Bioline) the following was added: 1 µl Forward Primer (10 µM), 1 µl Reverse Primer (10 µM), 1 µl bacterial DNA from heated solution and 9.5 µl water. The cycling conditions in Table 3.7 and the primers from Table 3.8 were used. After PCR the samples were kept at 4 °C, until analysis was carried out using 1 % agarose gels.

Table 3.7. The Cycling Conditions for the colony PCR to determine successful mutants.

Reaction Step	Temperature (°C)	Time (s)	
Initial Denaturation	95	60	
Denaturation	95	30	Repeat for 35 cycles.
Annealing	Dependent on primers used	30	
Extension	72	90	
Final Extension	72	300	

Table 3.8. The different primers to confirm motif I and the OB domain swaps.

Mutant Check	Primer	Sequence (5'-3')
Motif I switch	256 Motif I	CTA-ACA-AAA-GAA-GAC-GGG-TCT-TTG-G
	257 Motif I	CGT-GAA-AAG-ATT-CAC-GGC-ACA-AAT-TTC
	T7 Terminator	GCT-AGT-TAT-TGC-TCA-GCG-G
OB fold addition	T4 Rnl1 Forward	TGG-TCT-CTA-CCC-ATG-CAA-GAA-CTT-TTT-AAC-AAT-TTA-ATG-GAA-CTA-TG
	T4 Rnl2 Forward	TGG-TCT-CTA-CCC-ATG-TTT-AAA-AAG-TAT-AGC-AGT-CTT-GAA-AAT-CAT-TAC
	T4 Dnl OB fold Reverse	TGG-TCT-CTG-ATG-GGA-TCC-TCA-TAG-ACC-AGT-TAC-CTC-AT

3.2.5. Protein Synthesis and Purification

To synthesise the new ligase enzymes the plasmids were transformed (see Section 2.2.6) into BL21 DE3 *E. coli* cells (Table 2.1). Proteins were expressed using either 0.5 mM IPTG or 0.5 mM arabinose (T4Rnl1OB and T4Rnl1OB) and purified using a Hi-Trap Chelating column (see Sections 2.3.1 – 2.3.6). For proteins that were insoluble under these conditions, different concentrations of IPTG or arabinose (0 mM – 10 mM) were used at different incubation temperatures (10 °C – 30 °C). The amount of protein was quantified using the Bradford assay (see Section 2.3.8).

3.2.6. Ligation Assays

The end point ligation assay (see Section 2.4.1) was carried out to assess whether the new soluble ligases could still effectively ligate breaks between the DNA/RNA substrates. If the ligases could ligate substrate 7, the initial rate was also determined

using the time course ligation assay (see Section 2.4.2). The ligation assays were analysed using urea-PAGE (see Section 2.4.3).

3.3 Analysis of the modified T4 NA ligases

3.3.1 Sequence Analysis of New Gene Constructs

To confirm the gene sequences contained the new modified motif I sequences or contained motif swaps (Figures 3.9 and 3.10), the plasmids were sent for Sanger sequencing at Eurofins (i54 Business Park, Valiant Way, Wolverhampton, WV9 5BG, UK).

3.3.1.1 Colony PCR for Preliminary Sequence Analysis

Colony PCR was carried out to check for possible positive transformants for motif swaps before plasmids were purified and sent for sequencing. Colonies that showed a positive result from the colony PCR (Figure 3.8) underwent plasmid extraction (see Section 2.2.2) and these were then sent for Sanger sequencing. The gene sequences for all mutated proteins were successfully prepared (Figures 3.9 and 3.10).

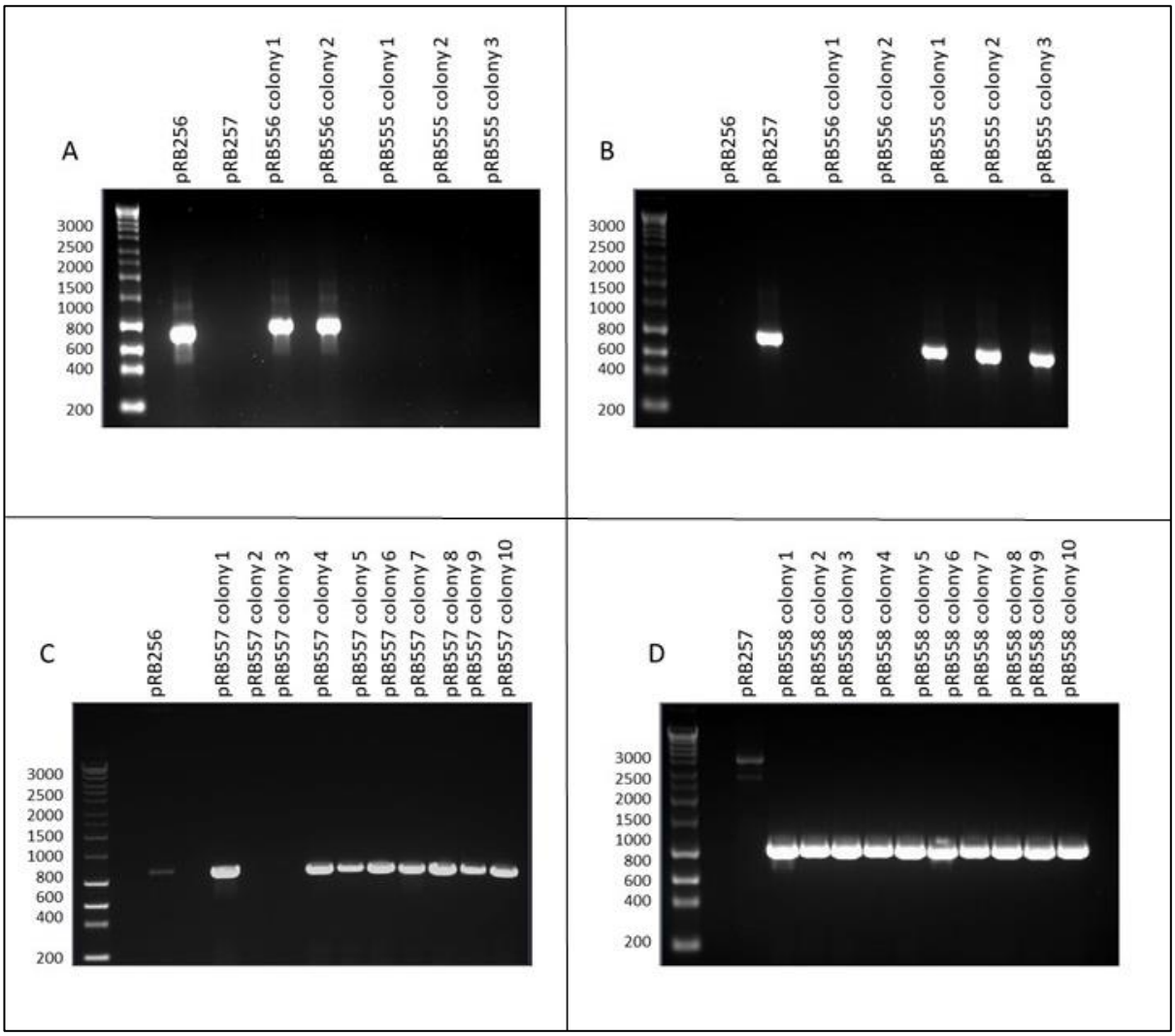


Figure 3.8. Colony PCR analysis for motif swap analysis. Identification of positive clones, a band is seen where the colony contains both the ligase N- terminus and the OB domains, analysed using 1 % TAE agarose gels; A) pRB556 positive transformants for motif I from T4Rnl1 (pRB256) and negative controls with no band; B) pRB555 positive transformants for motif I from T4Rnl2 (pRB257) and negative controls with no band; C) pRB557 transformants containing both T4Rnl1 N- terminus (pRB256) and the OB domain from T4Dnl, a light band is seen with pRB256 as this contains only the N- terminus for T4Rnl1; D) pRB558 transformants containing both T4Rnl2 N- terminus (pRB257) and the OB domain from T4Dnl, a light band is seen for pRB257 as it contains only the N- terminus for T4Rnl2.

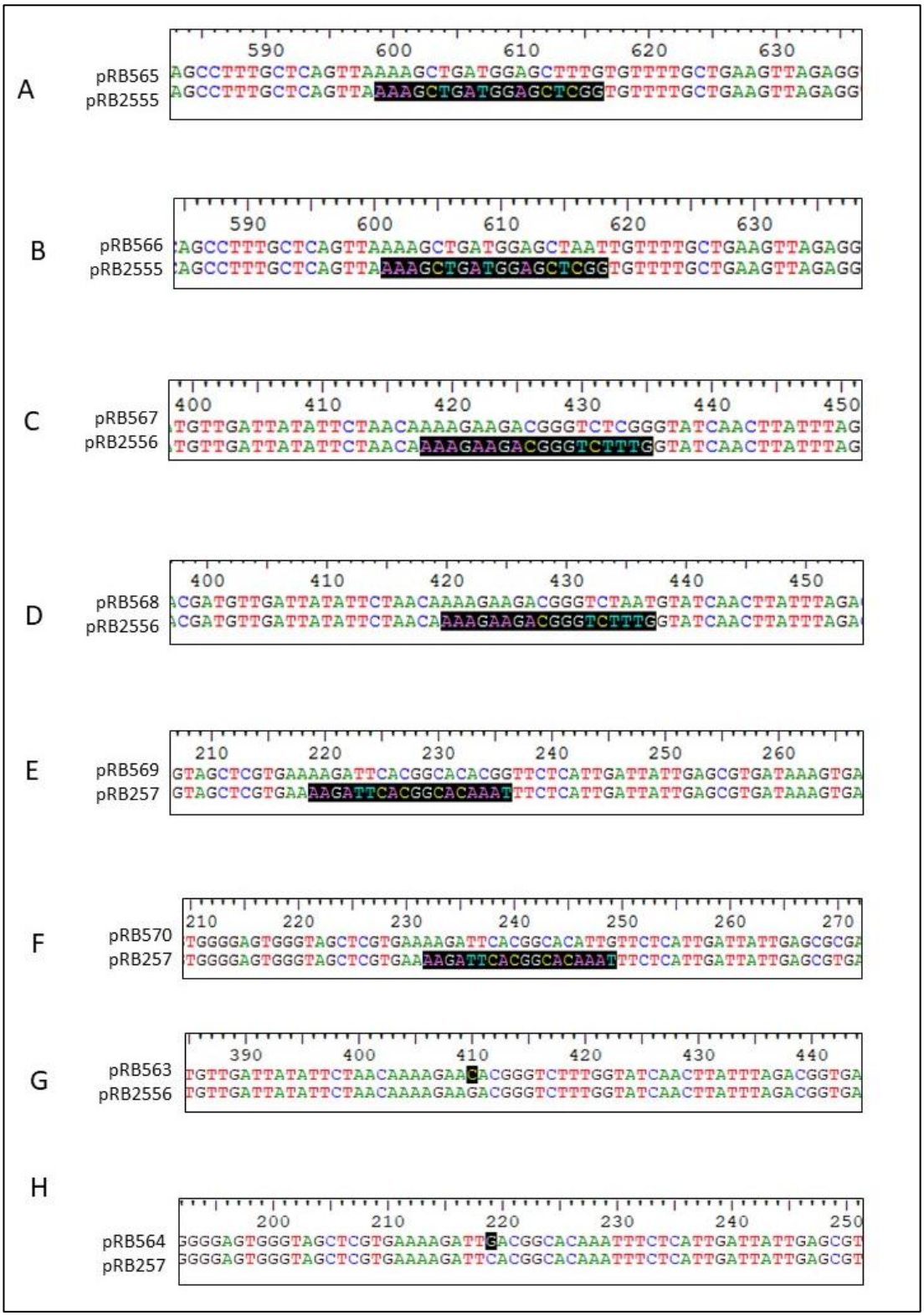


Figure 3.9. Sequence analysis for each of the mutations made to the T4 nucleic acid ligases. Plasmids were sequenced using the T7 forward and T7 terminator reverse primers by Eurofins Genomics. The sequences received from Eurofins were aligned to the original ligase gene sequences A) pRB565, B) pRB566, C) pRB567, D) pRB568, E) pRB569, F) pRB570, G) pRB563 and H) pRB564. The shaded regions show where sequence mutations were made.

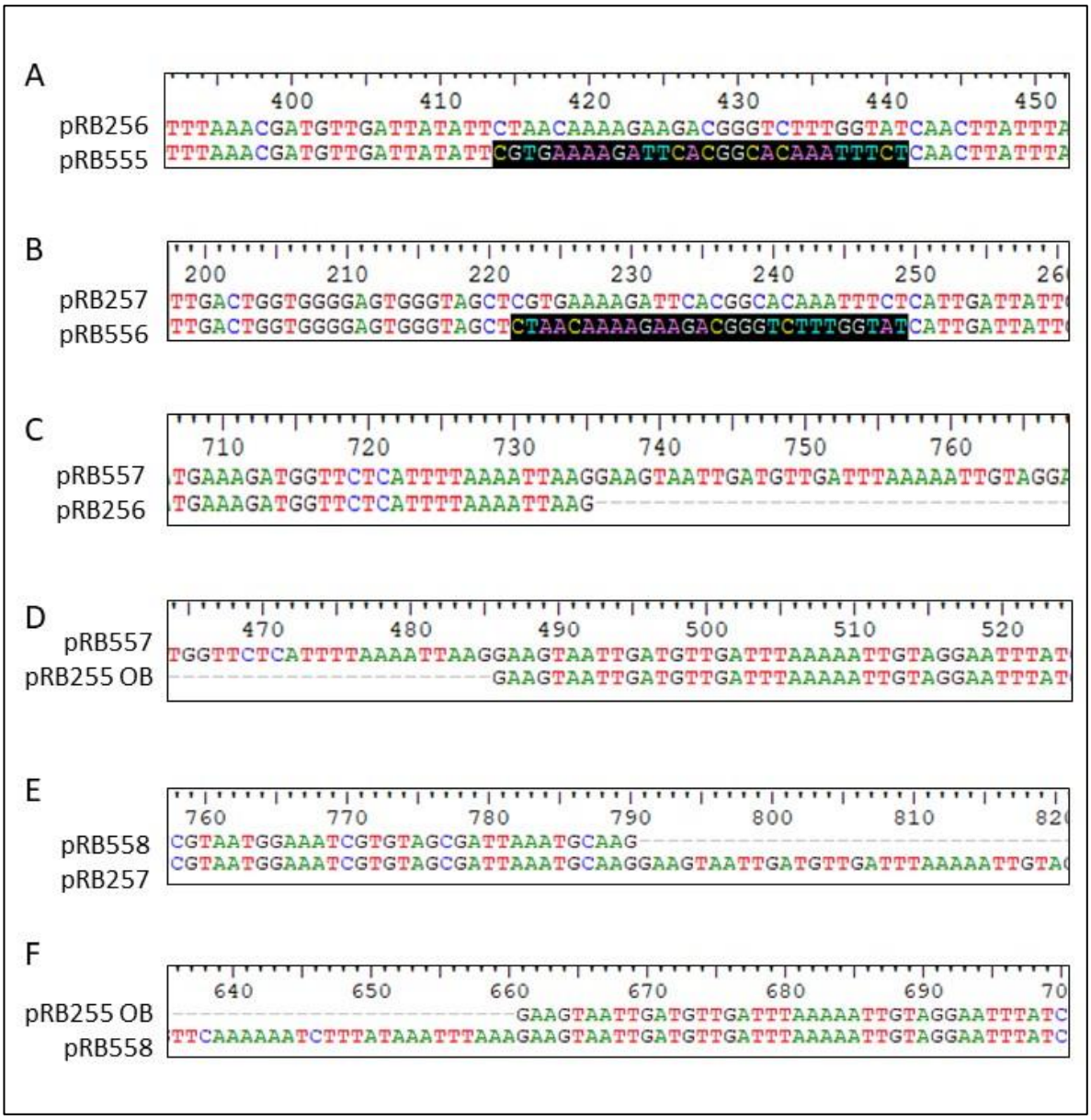


Figure 3.10. Sequence analysis for the motif switch plasmids. The sequences from Eurofins were aligned to the original sequences for the RNA ligases and the T4Dnl OB domain A) pRB555 and pRB256 aligned, motif I (highlighted) has been successfully changed; B) pRB556 and pRB257 aligned, motif I (highlighted) has been successfully changed; pRB557 aligned to pRB256 N- terminus (D) and pRB255 OB domain (E); pRB558 aligned to pRB257 N- terminus (E) and pRB255 OB domain (F).

3.3.2 Protein Solubility and Purification

Using the new DNAs that contain the mutated gene sequence, numerous attempts were made to purify soluble T4Rnl2M1, but the protein remained insoluble under different conditions for protein synthesis. As such this protein was not purified and further

analysis was not conducted. The motif swap proteins T4Rnl1OB and T4Rnl2OB were synthesised from both the MEGGA vector and also the pET16b system, it was unclear whether T4Rnl2OB was insoluble in the pET16b system, the soluble fraction was purified and the fractions analysed using 10 % SDS PAGE (data not shown) but no purified protein was seen. The mutated forms of T4Rnl2 remained insoluble from both systems (Figures 3.11 and 3.12) and further analysis was not conducted. T4Rnl1M1 and T4Rnl1OB were soluble and were purified (Figure 3.13) but did not have ligation activity (data not shown). The ligases with amino acid substitutions to motif I were soluble and were purified (Figure 3.14) for end point and time course ligation analysis (Figures 3.15 – 3.18). The time course analysis was carried out for both 30 and 60 minute with aliquots removed at different time points (Figure 3.18) to determine the reaction end point.

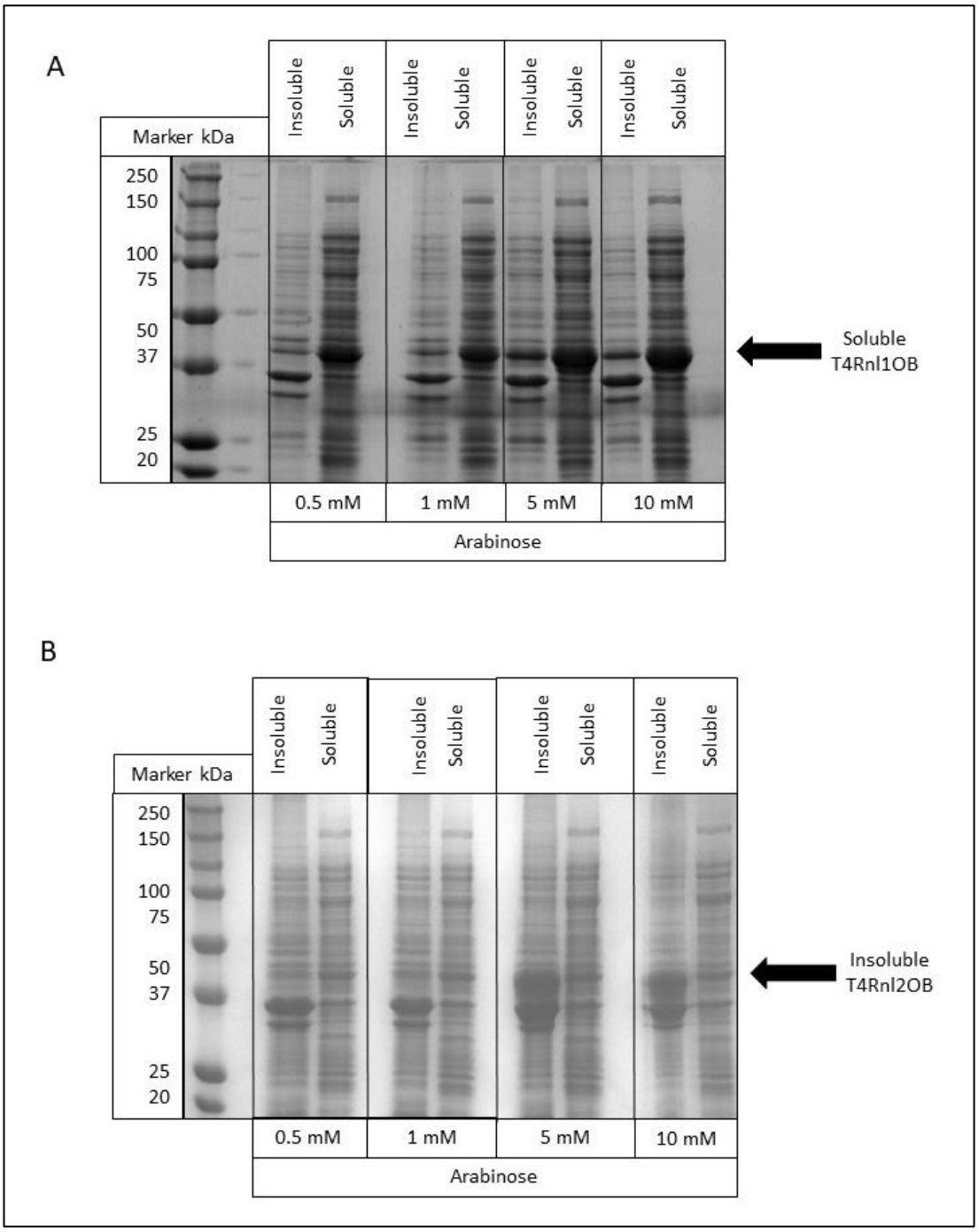


Figure 3.11. Solubility analysis from the MEGGA plasmid arabinose induction for T4Rnl1OB and T4Rnl2OB. Arabinose induction with the MEGGA plasmid for both proteins at 30 °C for 4 hours, analysed using 10 % SDS- A) T4Rnl1OB was soluble (predicted MW 41.7 kDa), B) T4Rnl2OB (predicted MW 39.3 kDa) was insoluble at all arabinose concentrations.

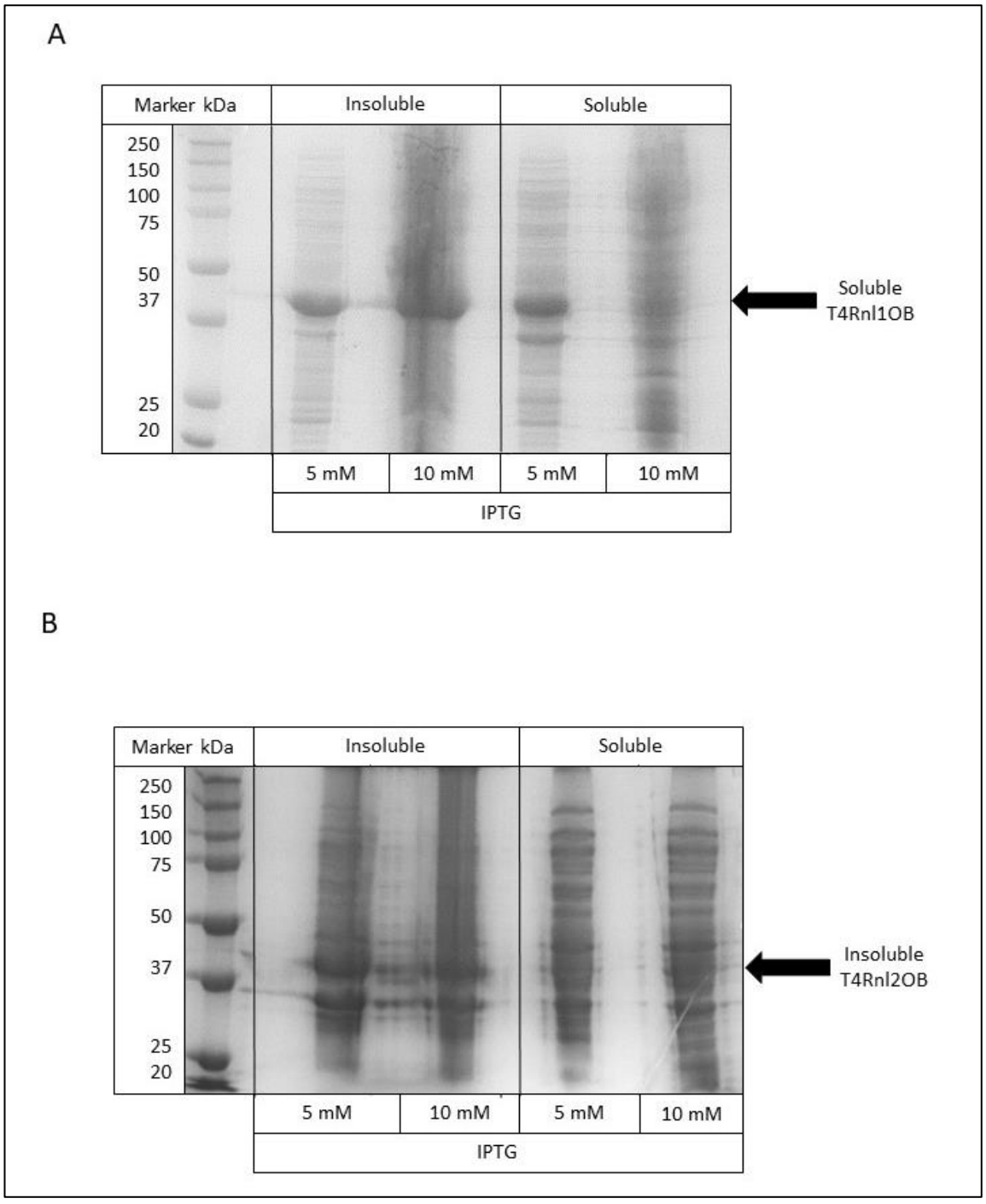


Figure 3.12. Solubility Analysis for T4Rnl1OB and T4Rnl2OB with the pET16b plasmid and IPTG induction. Induction with 0.5 mM IPTG at 30 °C for four hours, the soluble and insoluble fractions were analysed using 10 % SDS-PAGE; A) T4Rnl1OB was soluble while it remained difficult to determine whether B) T4Rnl2OB was soluble or insoluble, the soluble T4Rnl2OB fraction was purified (data not shown) but no protein was seen.

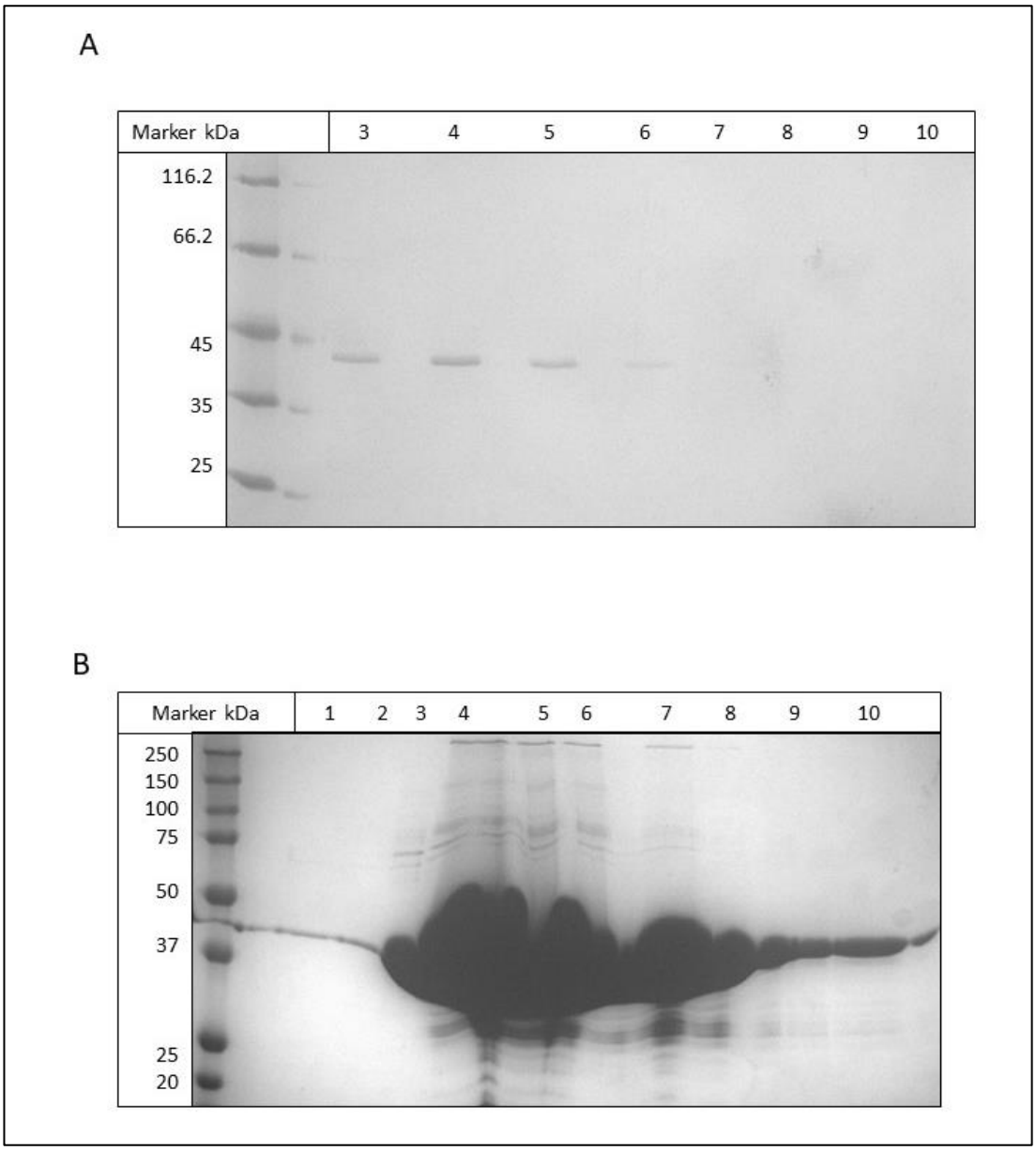


Figure 3.13. Purification Analysis of T4Rnl1M1 and T4Rnl1OB. Purification fractions for A) T4Rnl1M1 fractions 3-10 and B) T4Rnl1OB fractions 1-10, analysed using 10 % SDS PAGE. The purification fractions for T4Rnl1OB have high levels of pure protein when compared to T4Rnl1M1, as seen by the concentrated banding across fractions 3-10 (B).

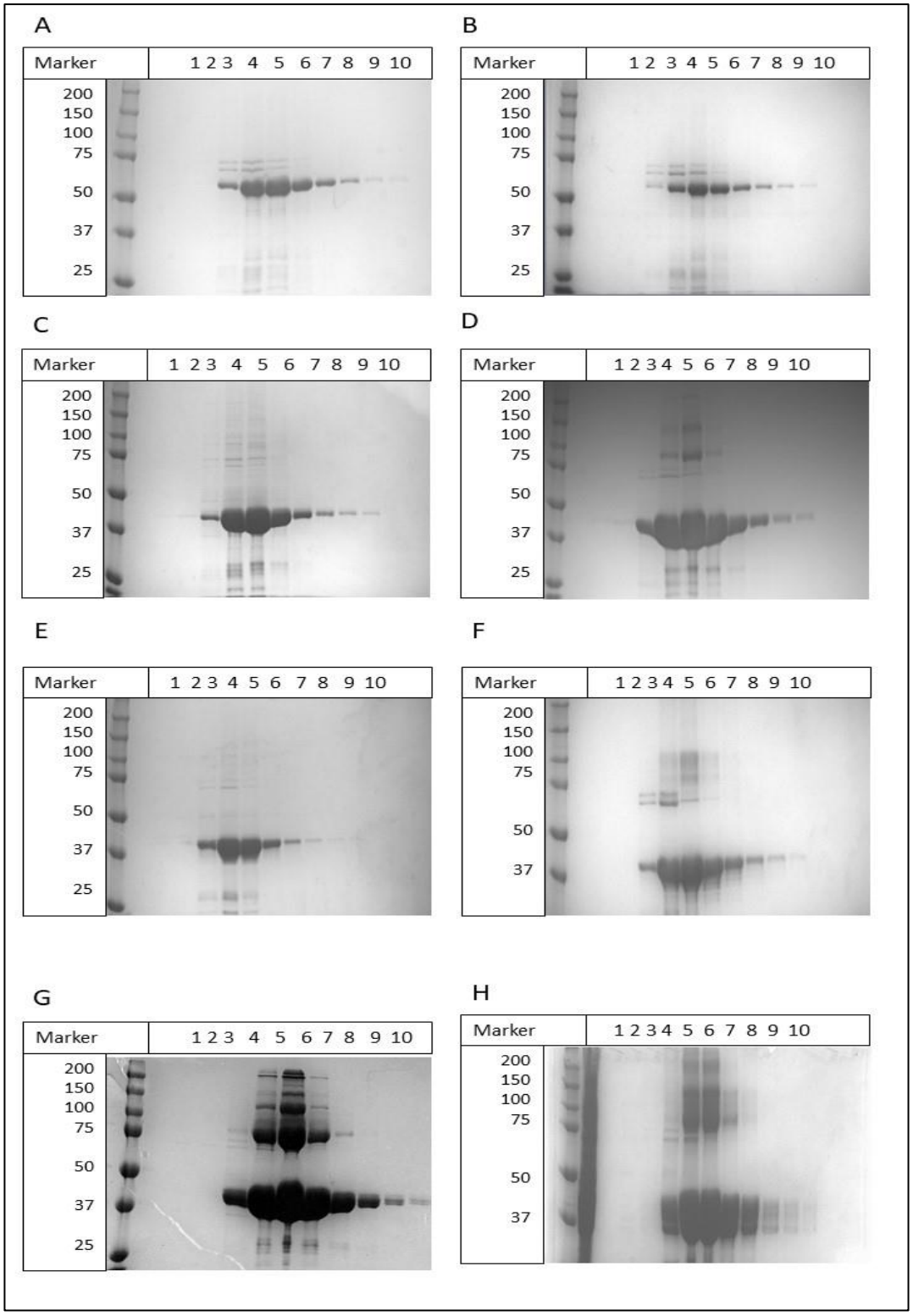


Figure 3.14. Purification fractions of the amino acid substitutions to motif I in the T4 NA ligases. Ni Hi-trap purification analysed using 10 % SDS PAGE, protein purification elution fractions 1- 10 for point mutations made to each of the T4 ligases. The expected molecular weight for each ligase is: A) R164L 55.2 kDa, B) R164N 55.2 kDa, C) L104R 43.5 kDa, D) L104N 43.5 kDa, E) N40R 37.6 kDa F) N40L 37.6 kDa, G) D101H and H) H37D 37.6 kDa. The expected sizes of each ligase are: T4Dnl 55 kDa, T4Rnl1 43 kDa and T4Rnl2 37.5 kDa (gels for wild type proteins are shown in Figure 4.9).

3.3.3. Ligation Activity Analysis

All proteins purified as soluble were checked for ligation activity using end point ligation (see Section 2.4.1) with double stranded substrates 1-8. In some instances the new ligase proteins showed no ligation activity (data not shown) and no further analysis was carried out on these ligases. Where proteins demonstrated ligation activity with substrate 7, time course ligation analysis (see Section 2.4.2) was also carried out.

3.3.3.1. End Point Ligation

End point ligation of all soluble proteins was attempted with double stranded substrates 1-8 and the hairpin substrate (see Section 2.4.1). The hairpin substrate mimics the tRNA anticodon loop that T4Rnl1 repairs during *E. coli* infection. The ligases with motif swaps, T4Rnl1M1 and T4Rnl1OB, showed no activity for the end point ligation assay (data not shown) and no further analysis was carried out with these proteins. The motif I substitution proteins demonstrated ligation activity (Figures 3.15 and 3.16), with the exception of L104R. The only ligase able to improve ligation yield was H37D with substrate 7 (Figure 3.16). Some of the new ligases had a severe reduction in total amount of ligated product: R164N with both substrates 1 and 7; D101H especially with substrate 2 and all T4Rnl2 mutants could no longer ligate substrate 1 (Figure 3.16). As substrate 7 was the substrate that all of the T4 ligases can ligate (Bullard and Bowater, 2006), this substrate was used to consider whether the rate of ligation for the mutated ligases could be calculated. When tested with the hairpin substrate L104N was the only protein to successfully ligate this substrate (Figure 3.17).

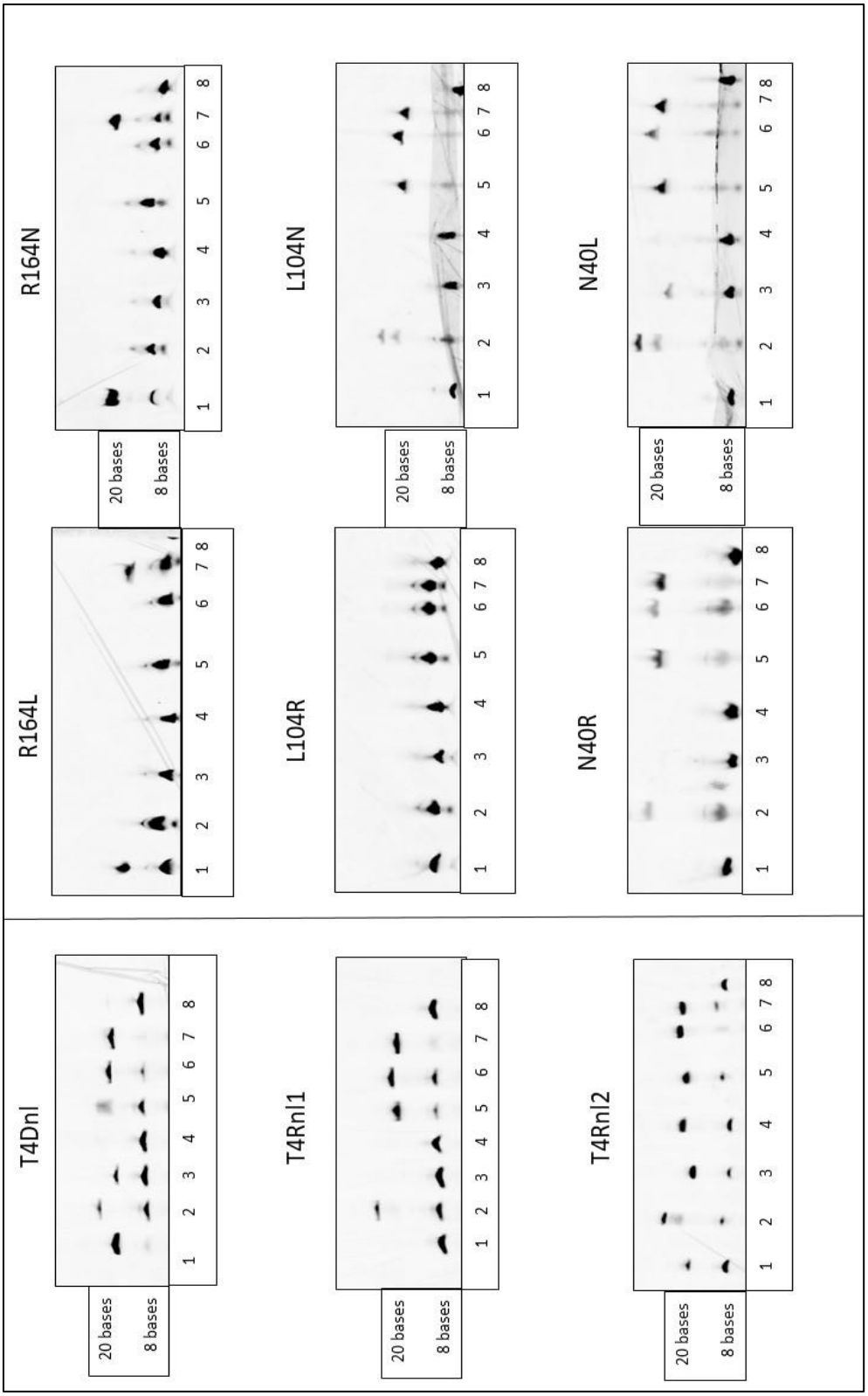


Figure 3.15. End point ligation analysis of ligases with substitutions to motif I. Urea 15 % polyacrylamide denaturing gels for end point ligation at 37 °C for 30 minutes A) T4Dnl R164L, B) T4Dnl R164N, C) T4Rnl1 L104R, D) T4Rnl1 L104N, E) T4Rnl2 N40R, F) T4Rnl2 N40L, G) T4Rnl1 D101H and H) T4Rnl2 H37D with 20 base oligonucleotide DNA/RNA substrates 1-8 (Table 2.4). In all cases ligation is indicated by the bands that migrate more slowly through the gels. The ligated products containing RNA run slightly slower than those containing just DNA, this is a known phenomenon but all substrates when ligated are 20 bases in size (Bullard and Bowater, 2006).

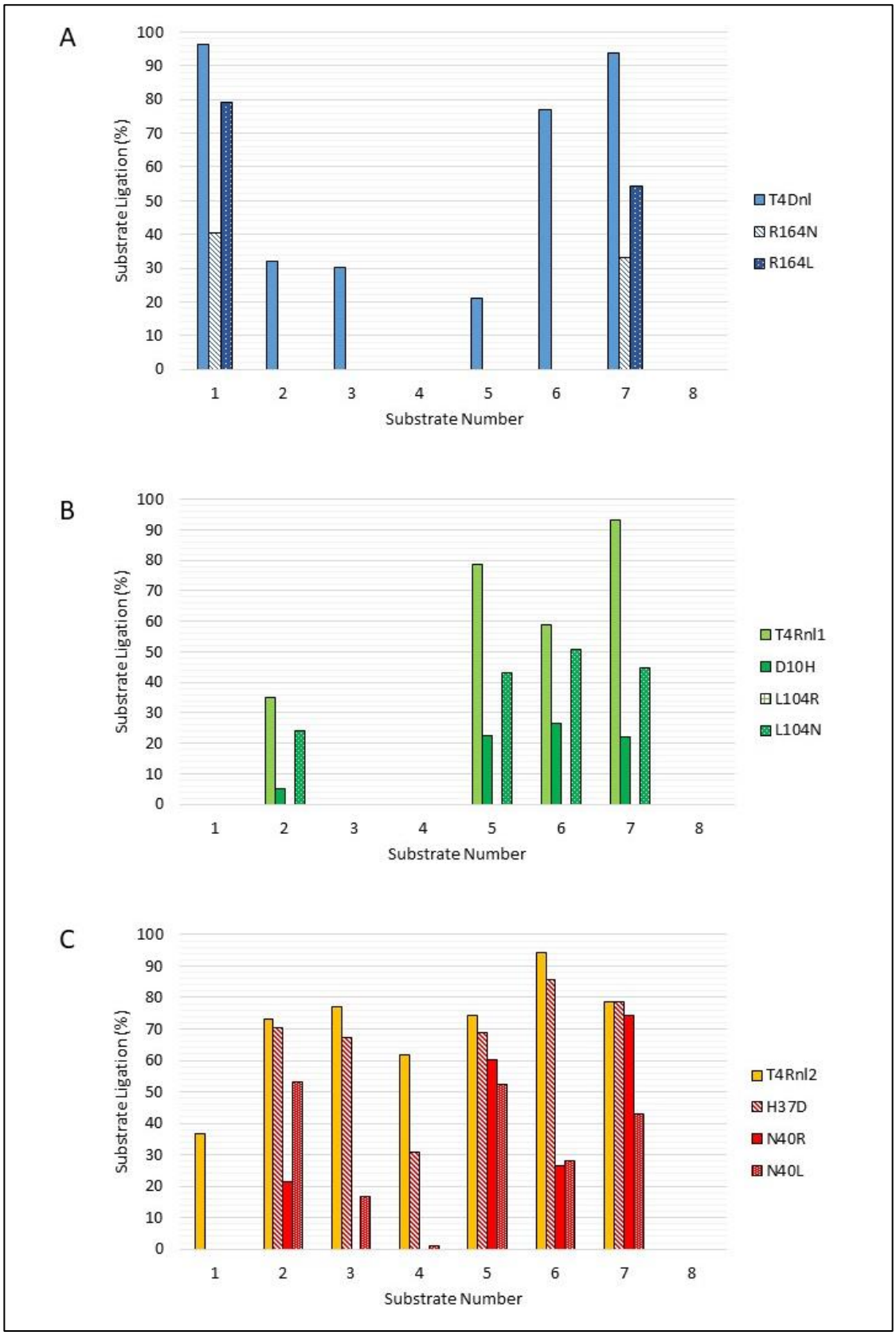


Figure 3.16. The percentage of ligation for each ligase with each of the double stranded 20 base substrates (1-8). Quantitative analysis of the ligation data from Figure 3.15; A) T4Dnl, R164N and R164L; B) T4Rnl1, L104N, L104R and D101H; and C) T4Rnl2, H37D, N40L and N40R. Quantification was carried out using the ImageJ software.

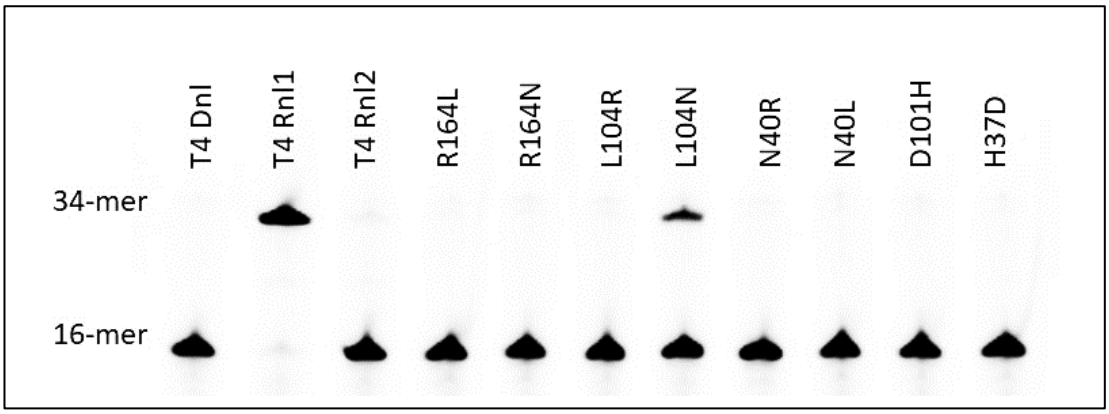


Figure 3.17. Hairpin End Point Ligation. Ligation of the hairpin substrate (45 pmol) with the T4 nucleic acid ligases (70 pmol) and the mutated versions of the ligases, incubated at 37 ^{o}C for one hour.

3.3.3.2 Time Course Ligation

Time course ligation analysis (See 2.4.2) was carried out with substrate 7 (Table 2.4) and analysed using Urea-PAGE analysis (2.4.3) for quantification using ImageJ software. The ligases tested for initial rate analysis (Figure 3.18) were H37D and N40L, as these ligases appeared able to ligate the same total amount of substrate 7 as T4Rnl2 in the end point analysis (Figure 3.16). The rates of ligation for H37D and N40L were compared to T4Rnl2 (Table 3.9) and N40L had 10-fold less activity with substrate 7.

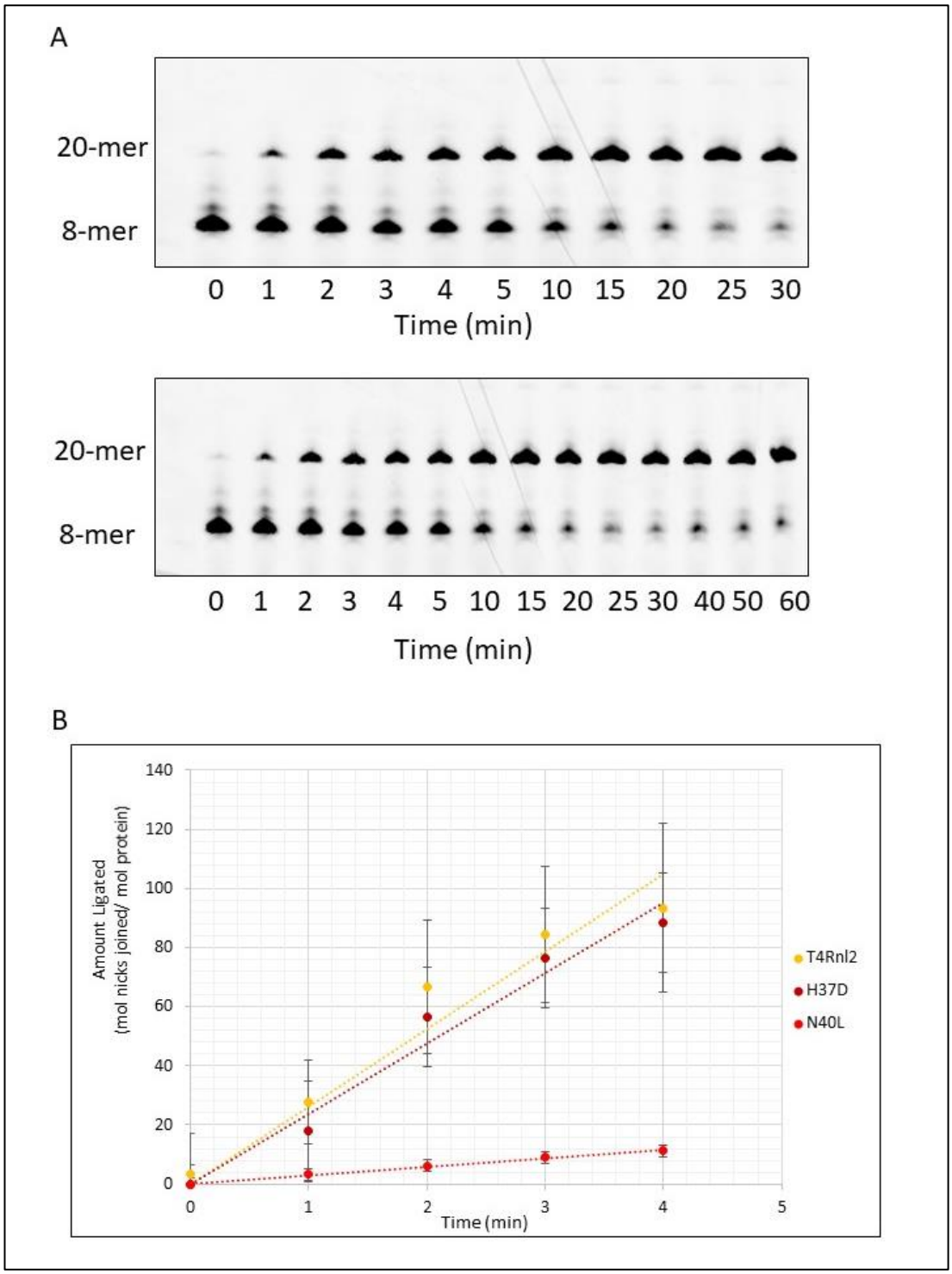


Figure 3.18. Initial rate analysis of T4Rnl2, H37D and N40L. Reactions over 30 and 60 minutes with substrate 7, at 37 °C and repeated in triplicate, with error bars showing the standard deviation for each time point. A) Examples of a time course experiments with T4Rnl2 using 15 % urea-PAGE for both 30 and 60 minutes; B) time course analysis, the gels are visualised using ImageJ software and the bands quantified to give the amount ligated as a rate.

Table 3.9. Rates of nick joining in double stranded substrate 7. Rates of ligation for T4Rnl2, H37D and N40L, experiments were repeated in triplicate and the standard deviation is shown.

Ligase	Rate of Ligation at 37 ^{o}C (mol nicks joined/mol protein/min)
T4Rnl2	26.28 (+/- 7.95)
H37D	23.8 (+/- 1.16)
N40L	2.94 (+/- 0.18)

3.4 Discussion

The results from this work yielded varied results, each motif I mutation gave different results and the motif swaps gave unexpected results for each RNA ligase. It is apparent that when working with the 3 different T4 NA ligases, a one-approach fits all does not work especially when trying to compare conserved sequences. The gene sequences for the T4 NA ligases are quite dissimilar to each other, there is some similarity but it is important to consider residues outside of the polypeptide motifs. The bioinformatics alignments for the T4 ligase sequences (Figures 3.3, 3.4 and 3.5) showed a number of highly conserved amino acids surrounding the motifs that are not mentioned in the literature. This raised the question about whether these contribute to the substrate specificity profiles for the different ligases and have involvement in the nucleotide binding pocket. A good example of this from the alignment data is motif I (Figures 3.3, and 3.5). Motif I is often described in literature as KxxGx, originally it was KxDGx but the discovery of T4Rnl2 meant that the 3rd position in this motif is not always aspartate. From the alignment data it would seem appropriate to make motif I a longer chain of amino acids and where this chain starts depends on the different ligase sub family. For example, in T4Rnl1 motif I has previously been identified as KEDGSL, but when aligned with other Rnl1-like proteins from other bacteriophages it appears that other residues downstream of these 6 are also always conserved. This was also seen in motif I of T4Rnl2 and Rnl2-like proteins, where the amino acids either side of motif I are always conserved. The motifs were originally identified by mutational analysis to specific amino acids and by modifications to the DNA/RNA substrates (Sriskanda and Shuman, 1998; Ho and Shuman, 2002; Wang, 2002; Nandakumar and Shuman, 2004; Nandakumar, Shuman and Lima, 2006). The interaction of the amino acids within the motifs, with DNA/RNA substrates was further investigated by crystallographic studies (Ho *et al.*,

2004; Nandakumar, Shuman and Lima, 2006; Pascal, 2008). It would be possible to study the motifs and their interactions further using computational software to improve predictions. Phylogenetic data of bacterial ATP-dependent DNA ligases identified a distinct group of ligases – LigE that were likely from bacteriophages originally (Williamson, Hjerde and Kahlke, 2016).

3.4.1. Motif I swaps – T4Rnl1M1 and T4Rnl2M1

These modified forms of ligases were incredibly difficult to purify and for T4Rnl2M1 it was not possible to purify soluble protein. The activity assay was carried out with T4Rnl1M1 but no activity was detected. This sparks further questions about which amino acids need to be conserved for ligation activity. The amino acids selected for the motif I swap included two amino acids upstream and one downstream of the 6 residues in motif I. In the case of T4Rnl1 it is likely that as these residues are also always conserved although not included in the published motif sequence, they are necessary for ligation activity and interaction with other residues in the nucleotide binding pocket.

3.4.2. OB fold domain additions – T4Rnl1OB and T4Rnl2OB

To decide how much of the OB fold from T4Dnl was used the theoretical structure for T4Dnl was created using Phyre2 software (Appendix 2). This was not a definitive structure and was mostly modelled on a human DNA ligase. The similarity between these two sequences was low as Human DNA ligase I is much bigger than T4Dnl, but 92 % of the residues from T4Dnl were modelled with 100 % confidence against Human DNA ligase I. The N- terminus of both RNA ligases was selected, up to the end of the adenylation domain. The part taken from T4Dnl starts at the end of its adenylation domain to the end of the C- terminus, a region that includes the OB fold but also most of the C- terminus from the T4 DNA ligase. Mutations to motif VI in the DNA ligase from *Chlorella virus* reduced ligation capacity and demonstrated that motif VI is needed for adenylation (Sriskanda and Shuman, 1998). The insolubility seen by T4Rnl2OB is likely caused by the addition of such a large domain, a more directed approach at adding only motif VI to the T4 RNA ligases is likely to be better although deciding exactly where to place this motif remains problematic. While this large addition did not affect the

solubility of T4Rnl1OB, it did result in a complete loss of ligation activity. The GGC technique made recombinant gene and plasmid construction straightforward and reduced the possibility of false cloning. The arabinose induction trial for T4Rnl1OB showed the protein was in the soluble fraction and the SDS-PAGE showed a clear band where the protein was but no ligation activity was observed.

3.4.3. Single amino acid substitutions

Specific amino acid mutations in motif I resulted in soluble protein for purification. All of these new ligases were active but did not demonstrate a different substrate profile or vary in the rates of activity. In all instances the changes to amino acids in motif I resulted in either a total loss of ligation activity or reduced activity. Modified versions of T4Dnl – R164N and R164L both resulted in reduced ligation activity, neither could ligate to the same total ligation amount as T4Dnl (Figure 3.16). For substitutions of T4Rnl1, both D101H and L104N remained functional; L104R was completely inactive with all versions of the double stranded substrates and could no longer ligate the hairpin substrate while D101H and L104N showed a decrease in total ligation (Figures 3.15 C and 3.16). The T4Rnl2 variants H37D and N40L were still able to ligate and showed a similar substrate profile as T4Rnl2, although all new ligases lost the capacity to ligate substrate 1 – double stranded DNA (Figure 3.15). When considering substrate 7 for rate analysis, N40R was not able to ligate this substrate as effectively in end point ligations and was not considered for rate analysis. The end point ligations used 45 pmol of substrate with 70 pmol of ligase, so it is possible that the amino acid mutagenesis to the T4 ligases has caused a shift in K_M and that the ligases are no longer saturated with substrate. It would therefore be useful to determine the V_{max} for each of the T4 ligases and also each of the variants of these ligases to see if V_{max} is reached. A high K_M would indicate a low substrate affinity and that a greater amount of substrate is required to reach V_{max}. The amount of substrate wasn't changed when testing these new ligases, if the new ligases could not ligate the amount the original T4 ligases ligated then no further analysis was carried out. A higher K_M for these new ligases may mean that 45 pmol of substrate was not enough to saturate and the ligase would not be able to reach $V_{max.}$

The rate of ligation for H37D is comparable to T4Rnl2 (Table 3.9, Figure 3.18) and the substitution to aspartate does not adversely affect ligation with double stranded substrate 7. The rate of ligation for N40L was reduced when compared to T4Rnl2 (Table 3.9) suggesting the substitution of asparagine to leucine impacts on either interaction of the enzyme with AMP or the substrate. Asparagine is polar, usually found on protein surfaces and can form hydrogen bonds. Leucine is hydrophobic and does not form hydrogen bonds (Betts and Russell, 2003) and, therefore, likely affects the protein structure for the nucleotide binding pocket in N40L. The composition of the nucleotide binding pocket of T4Rnl1 changes in steps 1 and 3 of the ligation reaction (Figure 1.5) (Wang, Schwer and Shuman, 2006). Residues in the motifs of both T4Rnl1 and T4Rnl2 are responsible for hydrogen bonding when repairing breaks in DNA and RNA (Nandakumar, Shuman and Lima, 2006; Wang, Schwer and Shuman, 2006).

3.5. Conclusions

To conclude, the T4 NA ligases have poor overall sequence similarity to each other and the bioinformatics alignment of sequences similar to the T4 RNA ligases showed a number of additional regions outside of the conserved motifs that are also conserved and may be important. A better approach to alter ligase activity and specificity would be to look at proteins from the T4Rnl1- and T4Rnl2-like families that ligate under different conditions or ligate different substrates and take a similar amino acid substitution approach to change the amino acids involved in particular activities. Specific residues in similar ligase proteins have been studied and the aspartate residue in motif I of T4Rnl1 is likely involved in AMP transfer to the nucleotide (Shuman and Lima, 2004). An attempt was made to mutate the aspartate in human DNA ligase 1 to glutamate, asparagine and glutamine resulting in loss of enzyme activity (Kodama, Barnes and Lindahl, 1991) and also the aspartate in DNA ligase from *Thermus thermophillus* is required for deadenylation (Luo and Barany, 1996).

It is difficult to decide modifications for T4Dnl as there is no structure available for this ligase. While a theoretical model was made (Appendix 2), it is modelled against human DNA ligase I; however the model aligns all of the T4Dnl sequence to part of human DNA

ligase – a much larger ligase. Without a crystallographic structure for this enzyme it is difficult to know how the motifs interact in the nucleotide binding pocket – even with solved structure for the T4 RNA ligases it is noted that these pockets are very different and have specific amino acid interactions for ligation activity. Thus, the T4 NA ligase genes are dissimilar and, as such, should be treated as separate groups of enzymes.

Chapter 4
Construction and Analysis of Adenylate Kinase-Ligases

Chapter 4: Construction and Analysis of Adenylate Kinase-Ligases

4.1 Introduction

4.1.1 Tagged and Fusion Proteins

The addition of amino acids and proteins to other proteins is a widely used tool for a number of processes, including purification and improving visibility and solubility of proteins. Common tags include Histidine for protein purification, Streptavidin/Biotin and alkaline phosphatase (Kimple and Sondek, 2004). A number of coloured tags have also been developed to improve visualisation of proteins; fluorescent tags, for example fluorescein and green fluorescent protein, allow visualisation of tagged proteins under specific wavelengths of light. While all of the tagging systems mentioned in section 1.3.1 offer a way to purify and visualise proteins, they do not offer a way to further enhance the function of the protein being purified. Often additional functions are not desirable as it would be difficult to determine what properties could be attributed to the protein of interest instead of the tag. However, where proteins are relatively well studied, the addition of a tag that provides a secondary function to the protein may be of interest; and as such these new fusions would result in multi-function enzymes.

4.1.2 Fusion proteins for improved activity

Fusion proteins to date have often focussed on taking domains from proteins and attaching these to other functional proteins. Recently advances have been made in combining binding sites from monoclonal antibodies to improve therapeutic efficacy (Holliger and Hudson, 2005). Another study used gold-binding peptide (GBP) fusions for detection of the Hepatitis B virus (Zheng *et al.*, 2010). When considering the T4 nucleic acid ligases, attempts have been made to improve the activity of T4Dnl by fusing it to a number of other proteins. Creating a fusion of T4Dnl with NF-κB p50 and T4Dnl bound to an artificial chimeric transcription factor improved ligation by 160 % (Wilson *et al.*, 2013). The addition of adenylate kinase to improve the solubility of T4Dnl has previously been studied (Liu *et al.*, 2015), an additional benefit of this protein-protein fusion was that the ligation buffer could contain ADP instead of ATP for phosphodiester bond formation (Liu *et al.*, 2015).

4.1.3 ATP degradation in buffers

The addition of adenylate kinase (AK) to the N-terminus of T4 DNA ligase improved solubility of the protein (Liu *et al.*, 2015). While solubility of T4Dnl is not primarily an issue, it is notably harder to get the same levels of soluble protein for the T4 RNA ligases. There was also the issue with different amounts of soluble protein after individual preparations for purifications (Nordlund *et al.*, 2008) and this can also be seen when purifying the T4 NA ligases. Addition of AK to help improve solubility was thought to be a way to ensure maximum amounts of soluble T4 nucleic acid ligases. The real interest in this chimeric construction was the possibility that the ligases would be able to use ADP as a co-factor as well as ATP, offering a novel function to the T4 RNA ligases.

ATP degradation in buffers is often used as a plausible reason for a decrease in ligase activity. This is usually remedied by making up new buffer with fresh ATP, while not overly time consuming this can become problematic when using large quantities of buffer or carrying out time dependent ligation. ATP breaks down to inorganic phosphate and ADP/AMP (Poldolsky and Morales, 1956; Hulett, 1970). ATP degradation at different temperatures and pH provide information on the best buffer storage conditions (Hulett, 1970) and while this helps to reduce issues with buffer, it could be better to provide an enzyme that could restore ADP to ATP for ligation.

It would be beneficial to both consumers and companies to produce a ligase that can work regardless of this degradation process, especially the T4 NA ligases which have varied rates of activity with different substrates (Bullard and Bowater, 2006). The adenylate kinase-T4Dnl ligase chimera has already been created (Liu *et al.*, 2015) and while the study demonstrates that both portions of the new enzyme work, nothing was reported on the rates and whether ADP works as well as ATP for phosphodiester bond formation. The AK-T4 RNA ligases as chimeric enzymes have not been made previous to this work and are novel variants of the enzymes.

4.1.4 Adenylate Kinase

Adenylate kinase (AK) is a nucleoside monophosphate (NMP) kinase which belong to a family of enzymes called phosphotransferases. In *E. coli*, AK is involved in the catalysis of AMP and ATP from ADP intermediates using magnesium as a cofactor (Glaser, Nulty and Vagelos, 1975; Saint Girons *et al.*, 1987; Müller *et al.*, 1996). The enzyme is comprised of 3 domains – the CORE domain, the LID domain and the AMP/NMP binding domain (Hayward, 2004) and has a mass of 23,585 Da (Figure 4.1) (Saint Girons *et al.*, 1987). The CORE domain is the part of the protein that remains fixed during phosphate movement, the LID domain covers the phosphates in the active site and the NMP binding domain is the binding site for AMP (Müller *et al.*, 1996). Adenylate kinase has the ability to transfer a phosphate group resulting in the reversible conversion of AMP + ATP to 2ADP. A structural conformational change occurs when adenylate kinase is bound to the substrate, from open to closed (Figure 4.2) (Müller *et al.*, 1996; Metzenberg, 2002; Snow, Qi and Hayward, 2007).

```
         10                  20                  30                  40                  50
MRIILLGAPG          AGKGTQAQFI          MEKYGIPQIS          TGDMLRAAVK          SGSELGKQAK
         60                  70                  80                  90                 100
DIMDAGKLVT          DELVIALVKE          RIAQEDCRNG          FLLDGFPRTI          PQADAMKEAG
        110                 120                 130                 140                 150
INVDYVLEFD          VPDELIVDRI          VGRRVHAPSG          RVYHVKFNPP          KVEGKDDVTG
        160                 170                 180                 190                 200
EELTTRKDDQ          EETVRKRLVE          YHQMTAPLIG          YYSKEAEAGN          TKYAKVDGTK
        210
PVAEVRADLE          KILG
```

Figure 4.1. The protein sequence for *E. coli* adenylate kinase (AK). AK has 3 domains: the CORE domain (green), the AMP binding domain (red) and the LID domain (blue). Amino acids are designated following the standard single letter abbreviations.

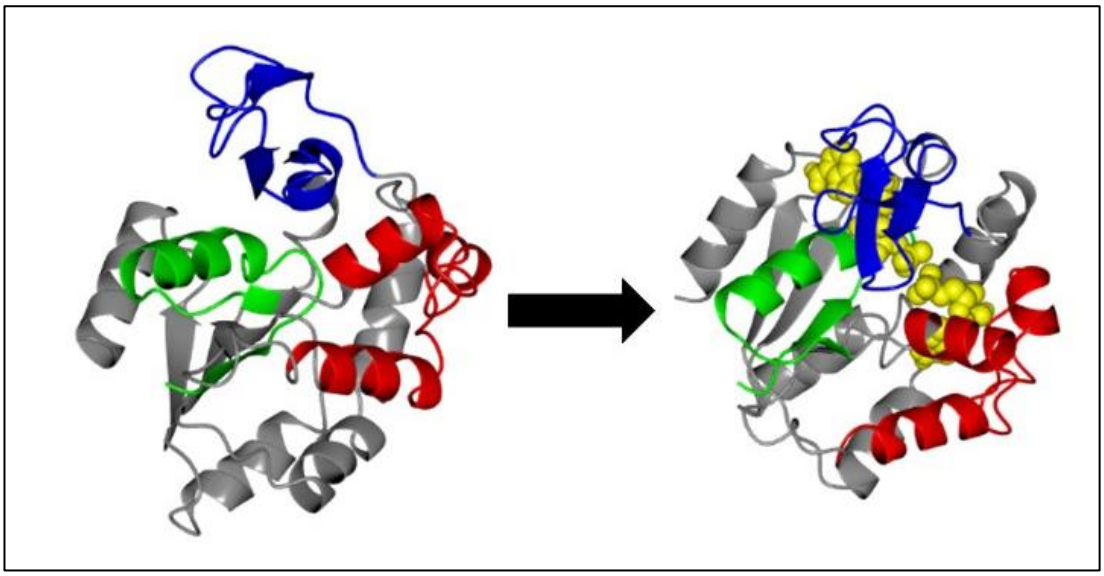


Figure 4.2. The crystallographic structure of AK. The crystal structure of Adenylate kinase with the CORE domain (green), AMP binding domain (red) and LID domain (blue) highlighted. Adenylate kinase has two different conformational states, when a NMP is bound, regions between the 3 domains bend; the open state (left) when unbound and the closed state (right) with AMP (yellow spheres) bound (Müller and Schulz, 1992). PDB files 1AKE and 1ANK illustrated using CCP4MG software.

When in the closed, bound conformation, adenylate kinase has 5 conserved arginine residues: Arg36, Arg88, Arg123 Arg156 and Arg167 that are involved in substrate binding (Yan, Shi and Tsai, 1990; Berry *et al.*, 2006). Arg36 is in the AMP binding domain, Arg123 is in the LID domain, the other conserved arginine residues sit between the three domains. Structural data suggested that Arg36, Arg88 and Arg156 are involved in AMP binding and Arg123 interacts with ADP in the nucleotide binding site (Berry *et al.*, 2006). When Arg138 was changed to lysine, the mutant adenylate kinase had a decrease in K_{cat} by a factor of 10^4, it is thought Arg138 is involved in transition state stabilisation (Yan, Shi and Tsai, 1990). The Asp89 residue holds the Mg^{2+} cofactor required by AK via two water molecules (Abele and Schulz, 1995; Schlauderer, Proba and Schulz, 1996).

The idea of using adenylate kinase as a solubility tag (Liu *et al.*, 2015) was repeated for the T4Dnl and the two T4 RNA ligases to determine if there was any effect on the protein solubility and substrate specificity of the ligases. When considering the sequence and structure of adenylate kinase it was necessary to note that the 3 domain sequences are at the N- terminus of the protein (Müller *et al.*, 1996), joining another enzyme at this

point would likely prevent LID domain movement rendering the adenylate kinase portion of the new chimeric enzyme inactive. The new AK-ligases were tested with 8 double stranded substrates and the hair-pin substrate to check if the addition of AK at the N- terminus of each of the T4 NA ligases affected the ligases substrate profile. The rates of the new ligases were also determined with both ATP and ADP in the buffer.

4.2 Construction of Adenylate kinase ligases

4.2.1 Recombinant Plasmid Construction

Three new plasmids were designed (pRB552, pRB553 and pRB554) with the adenylate kinase gene sequence at the N- terminus of each of the T4 ligases. These plasmids were made using restriction digestion and ligation. The adenylate kinase gene (*Adk*) was pre-ordered and synthesised in pUC57 (denoted pRB551) with the addition of *NcoI* and *NdeI* cut sites to allow for insertion into the plasmids containing the T4 nucleic acid ligase (pRB255, pRB256 and pRB257).

4.2.1.1 Restriction Digestion

To create the fragments required for the recombinant plasmids, the T4 ligase plasmids (pRB255, pRB256 and pRB257) and synthesised adenylate kinase (pRB551) underwent double digestion with *NcoI* and *NdeI* (ThermoFisher). The double digestion is specific to the formation of these recombinant plasmids and was set up as follows: 10 µl plasmid DNA (up to 1 µg), 10 µl sterile water, 1.25 µl *NcoI*, 1.25 µl *NdeI* and 2.5 µl ThermoFisher FD buffer. This was then incubated at 37 °C for up to 2.5 hours. To confirm digestion had been successful the samples were analysed on a 1 % agarose gel with 5 µl of 10 mg/ml ethidium bromide.

4.2.1.2. Gel Extraction

The double digested fragments containing the linearised vector and ligase for plasmids pRB255, pRB256 and pRB257; and the gene fragment for *Adk* underwent DNA extraction using the Bioline Isolate II PCR & Gel Kit. The manufacturer's protocol was followed, the amount of DNA in each sample was quantified using the Thermo Scientific Nanodrop one quantification system.

4.2.2. Ligating the Adenylate Kinase Gene to the N- terminal of the T4 ligases

To determine the ratio of vector to insert the following equation was used, this equation gives the ng of insert required for the specified ratio:

$$\left(ng\ vector\ x \frac{kb\ insert}{kb\ vector}\right)\ x \left(\frac{molar\ ratio\ of\ insert}{molar\ ratio\ of\ vector}\right)$$

This equation gives the amount of insert required in ng for a 1:1 ratio. The calculation was carried out for 1:1, 1:2, 1:3, 1:4 and 2:6. For the ligation reactions 50 ng of vector was used. The total sample volume for the ligation reaction was 8 µl, made up to this volume with nuclease free water. To this 1 µl of NEB T4 DNA ligase and 1 µl of 10x ligation buffer (Table 2.5.3) was added. The samples were incubated at 16 °C overnight. This was repeated for each ligation ratio for pRB255, pRB256 and pRB257 with adenylate kinase as the insert. The following day the total volume of ligation reaction was transformed into 50 µl Bioline α-select bronze cells, following the Bioline transformation procedure.

To confirm whether AK had been added to the N- terminal as a tag, the plasmid was extracted (See 2.2.2) and sent for Sanger sequencing at Eurofins. Once the sequence was confirmed (Figure 4.4), the new plasmid DNA was transformed into BL21 pLysS *E. coli* cells for protein expression and purification. The new plasmids were denoted pRB552, pRB553 and pRB554 (Figures 4.3, 4.4 and 4.5) for the T4 DNA ligase, RNA ligase I and RNA ligase II, with adenylate kinase as the N- terminal tag, respectively.

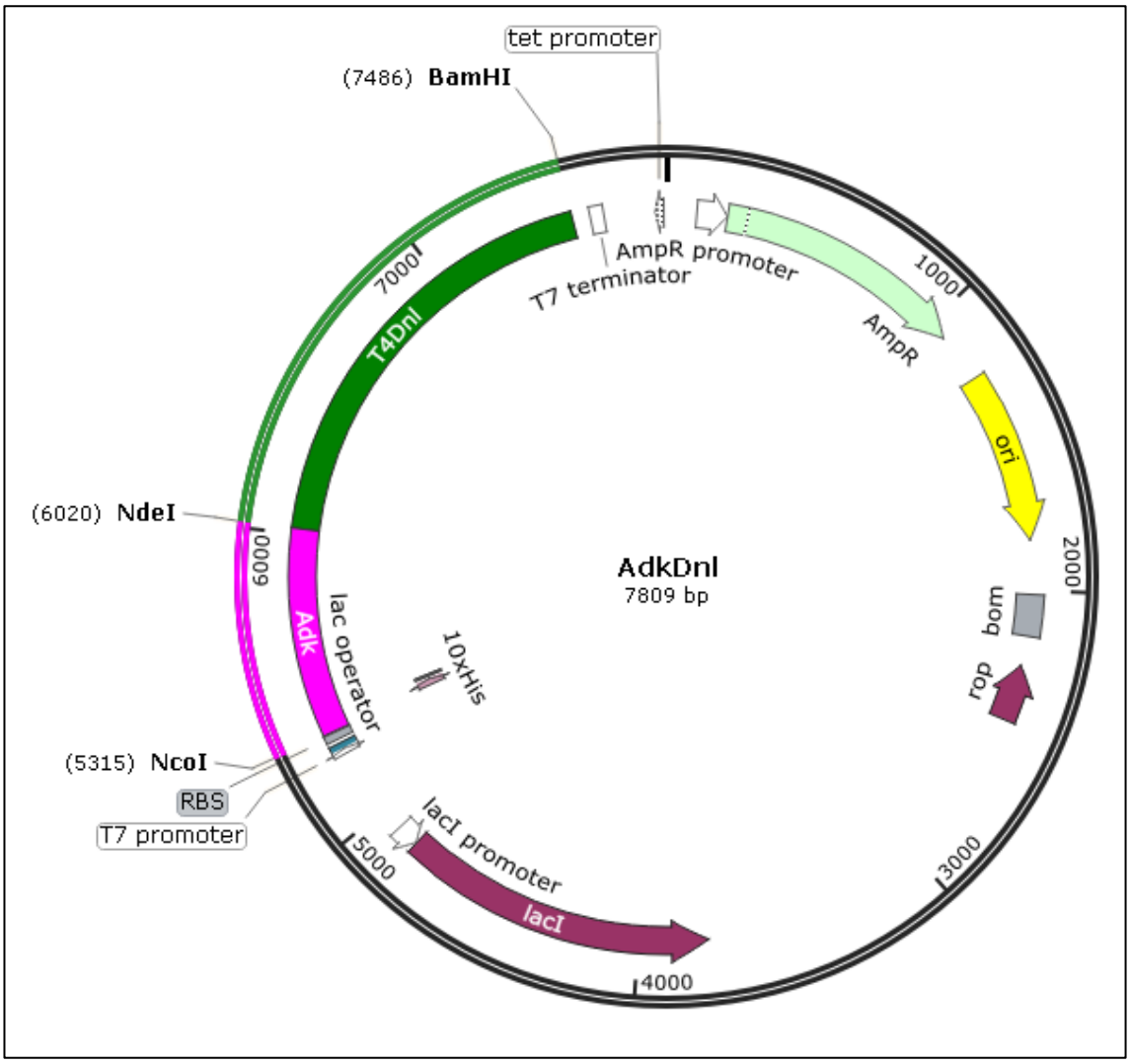


Figure 4.3. The vector map for the fusion of AK and T4Dnl to form AKDnl (pRB552). The AK (*Adk*) gene is between the *NcoI* and *NdeI* restriction sites and the T4Dnl gene between the *NdeI* and *BamHI* restriction sites. This maps was created using the SnapGene software.

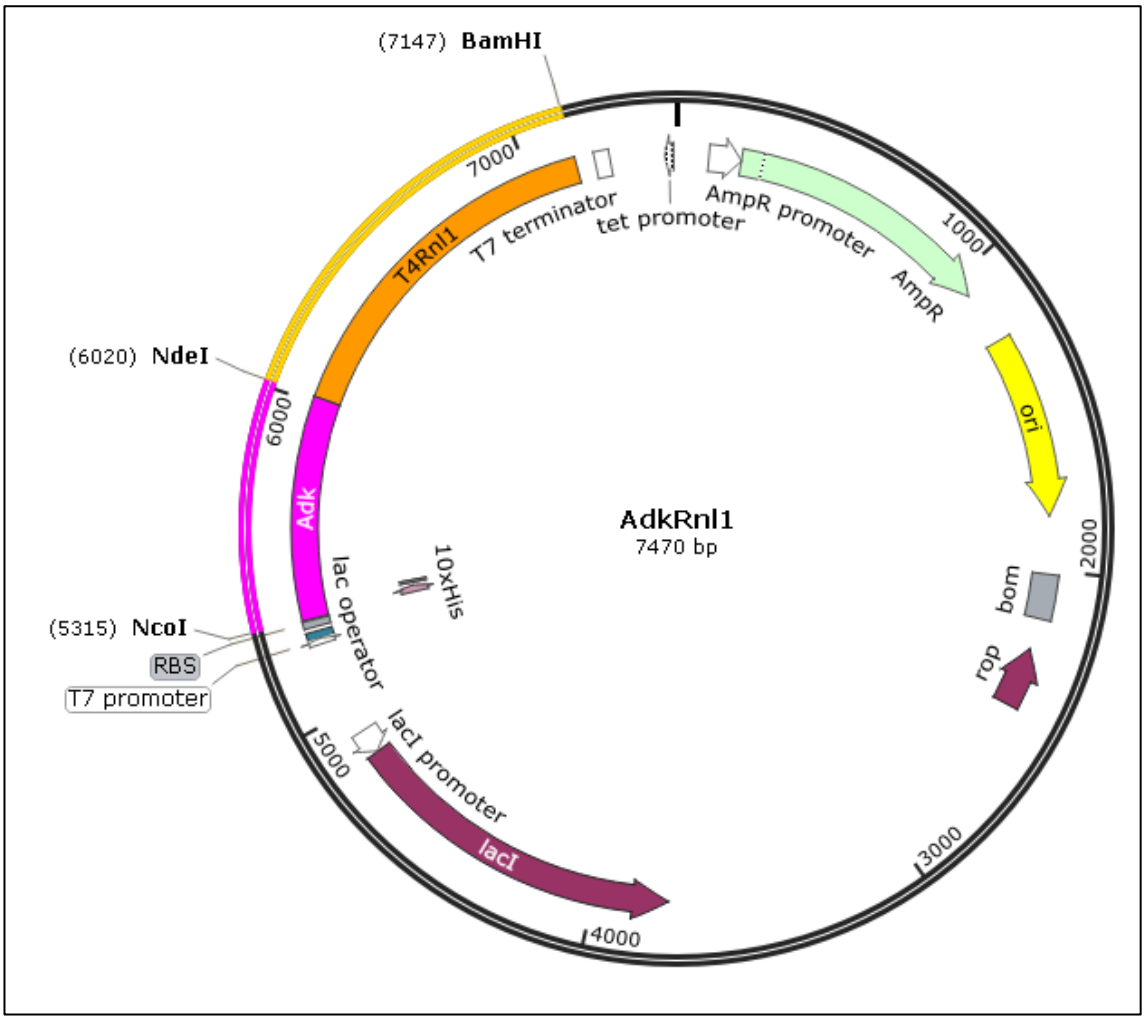


Figure 4.4. The vector map for the fusion of AK and T4Rnl1 to form AKRnl1 (pRB553).

The AK gene (*Adk*) was between the *NcoI* and *NdeI* restriction sites and the T4Rnl1 gene between the *NdeI* and *BamHI* restriction sites. This map was created using the SnapGene software.

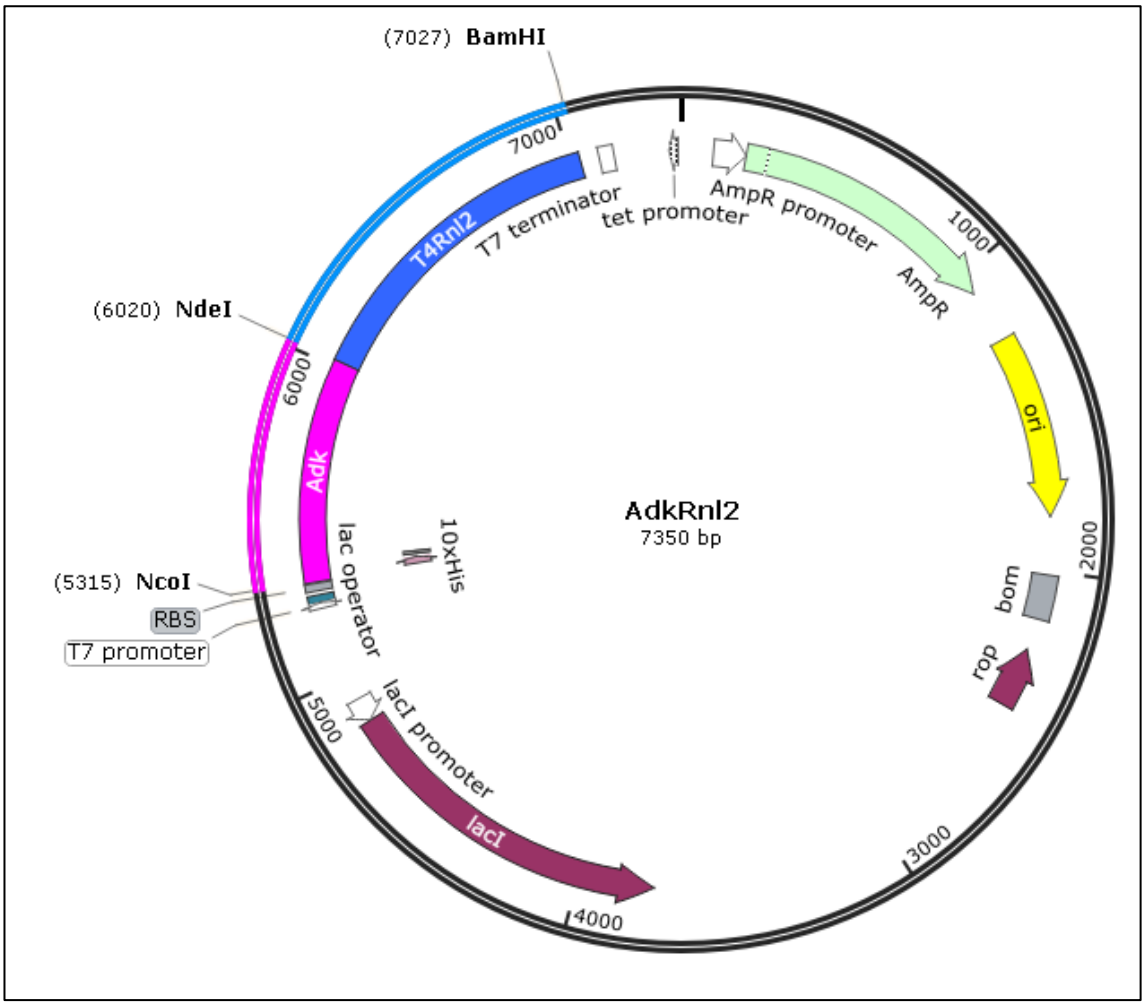


Figure 4.5. The vector map for the fusion of AK and T4Rnl2 to form AKRnl2 (pRB554). The AK gene (*Adk*) between the *NcoI* and *NdeI* restriction sites and the T4Rnl2 gene between the *NdeI* and *BamHI* restriction sites. This map was created using the SnapGene software.

4.2.3. Protein Characterisation

For protein induction and synthesis from 1 L of selective media see Chapter 2 – 2.3.1, 2.3.2 and 2.3.3, Western blot analysis was carried out on the new AK-ligases to confirm the 10x his tag was still present when the new, larger proteins were expressed (see 2.3.7). This tag is required for protein purification using the Ni-Hi Trap column. The proteins were purified using a Ni Hi-Trap column (See 2.3.4) and desalted using a pd10 column (see 2.3.5). SDS-PAGE was used to determine the correct fractions for desalting from protein purification, fractions F3-F5 were selected for all AK-ligases (Figure 4.7) and also to confirm the increase in size of the AK-ligases when compared to the original

ligases (Figure 4.6) (see 2.3.6). The concentration of the AK-ligases was determined using the Bradford micro assay (see 2.3.8).

4.2.4. Characterisation of biochemical activities

The activity of these new ligases was observed using the end point ligation assay and the rates calculated using the time course ligation assay (see methods 2.4.1 and 2.4.2). To confirm activity of the adenylate kinase part of the fusion was functional, the end point ligation assay was set up (see 2.4.1) using 3 different buffer conditions – buffer without cofactor, buffer with ADP and buffer containing ATP. The original T4 ligases are unable to ligate the break in the double stranded substrates when using ADP as the co factor in the buffer, if the adenylate kinase part of the new chimeric enzymes was functional then the conversion of ADP to ATP would allow the ligase half of the chimera to successfully ligate the breaks and form a phosphodiester bond. The time course reaction was also set up with ADP as the co-factor to determine the rate of ligation when relying on the AK part of the chimera to provide the ATP for phosphodiester bond formation.

4.3 Adenylate kinase Ligase Activities

4.3.1 Confirmation of AK-T4 Ligase DNA constructs

To confirm the *Adk* gene had been successfully incorporated into the plasmid, the new plasmids were sent for Sanger sequencing at Eurofins. The new plasmids and pRB255, pRB256 and pRB257 were also digested using the restriction enzymes *NcoI* and *BamHI*. The digested plasmids were analysed using agarose gels containing ethidium bromide. The new AK-ligase plasmids were then compared to the original size of the T4 ligase plasmids.

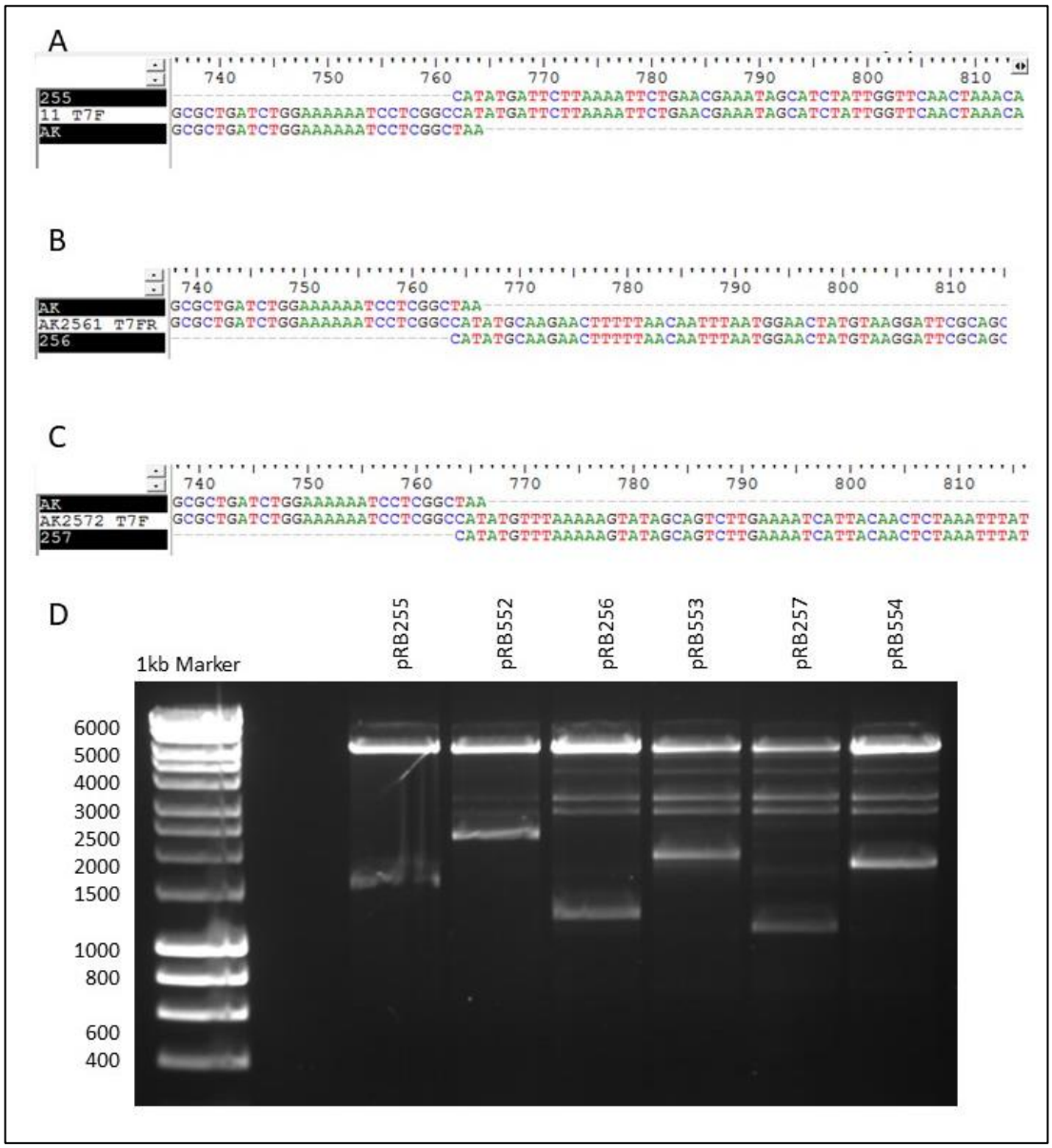


Figure 4.6. Sequence analysis of the new AK-ligases. A) AKDnl, B) AKRnl1 and C) AKRnl2 using the BioEdit software. Alignment shows the AK-ligase sequences, the original T4 ligase sequence (pRB255, pRB256 or pRB257) and the *Adk* gene sequence. The gene size increase was also confirmed using restriction enzyme digestion with both *NcoI* and *BamHI* (D). The expected sizes were: T4Dnl 1473 bp, AKDnl 2183 bp, T4Rnl1 1134 bp, AKRnl1 1844 bp, T4Rnl2 1014 bp and AKRnl2 1724 bp.

4.3.2 Protein purification

The induced AK-ligases were soluble (Figure 4.7) and the shift in ligase size was clear, the expected sizes were: AKDnl – 78.8 kDa, AKRnl1 – 67.1 kDa and AKRnl2 – 61.2 kDa (Figure 4.6). As the addition of AK has previously been reported to improve solubility of T4Dnl (Liu *et al.*, 2015) it was necessary to check whether the addition of AK affected

IPTG induction for protein purification (Figure 4.7). This may affect the amount of protein purified for the AK ligases. The concentration of protein from different purification preparations of the T4 NA ligases and the AK ligases (Table 4.1) showed some differences, however these are not consistent and are not reproducible.

Table 4.1. Differences in the concentration of purified protein for the T4 ligases and AK-ligases. Protein purified from 1 L of pelleted *E. coli* cells induced with 0.4 mM IPTG at 0.5 OD_{600}. Protein concentrations were calculated using the Bradford's Assay protocol (See 2.3.8).

Ligase	Protein Concentration (mg/ml)	
	Prep 1	Prep 2
T4 Dnl	6.69	2.08
AK-Dnl	2.94	6.95
T4 Rnl1	3.77	2.59
AK-Rnl1	2.04	4.97
T4 Rnl2	3.64	2.37
AK-Rnl2	9.2	6.46

There are noted differences between the 2 protein preparations for each ligase which is a phenomenon that has been reported before for other proteins (Nordlund *et al.*, 2008). The addition of AK to each of the ligases did not conclusively appear to impact the concentration of protein purified using the purification procedure detailed in this work.

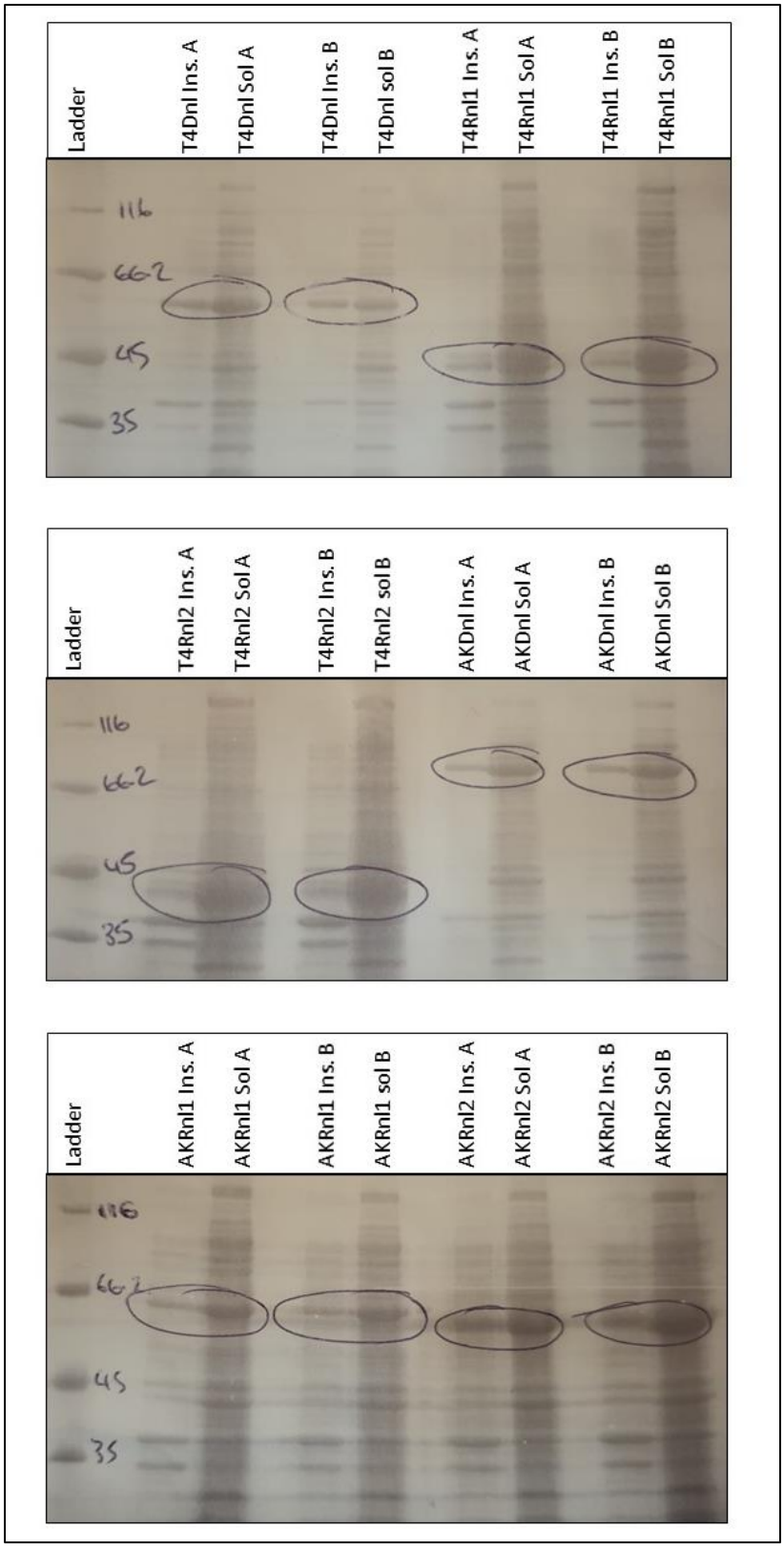


Figure 4.7. IPTG induction trials of the T4 and AK ligases. Each induction was carried out with either 0.5 mM IPTG (A) or 1 mM IPTG (B) as detailed in section 2.3.2. The cells were lysed via sonication and the soluble and insoluble fractions separated by centrifugation. The insoluble pellet was resuspended in 20 ml 20 mM Tris 100 mM NaCl (pH 7.5). The soluble and insoluble fractions were analysed using SDS-PAGE.

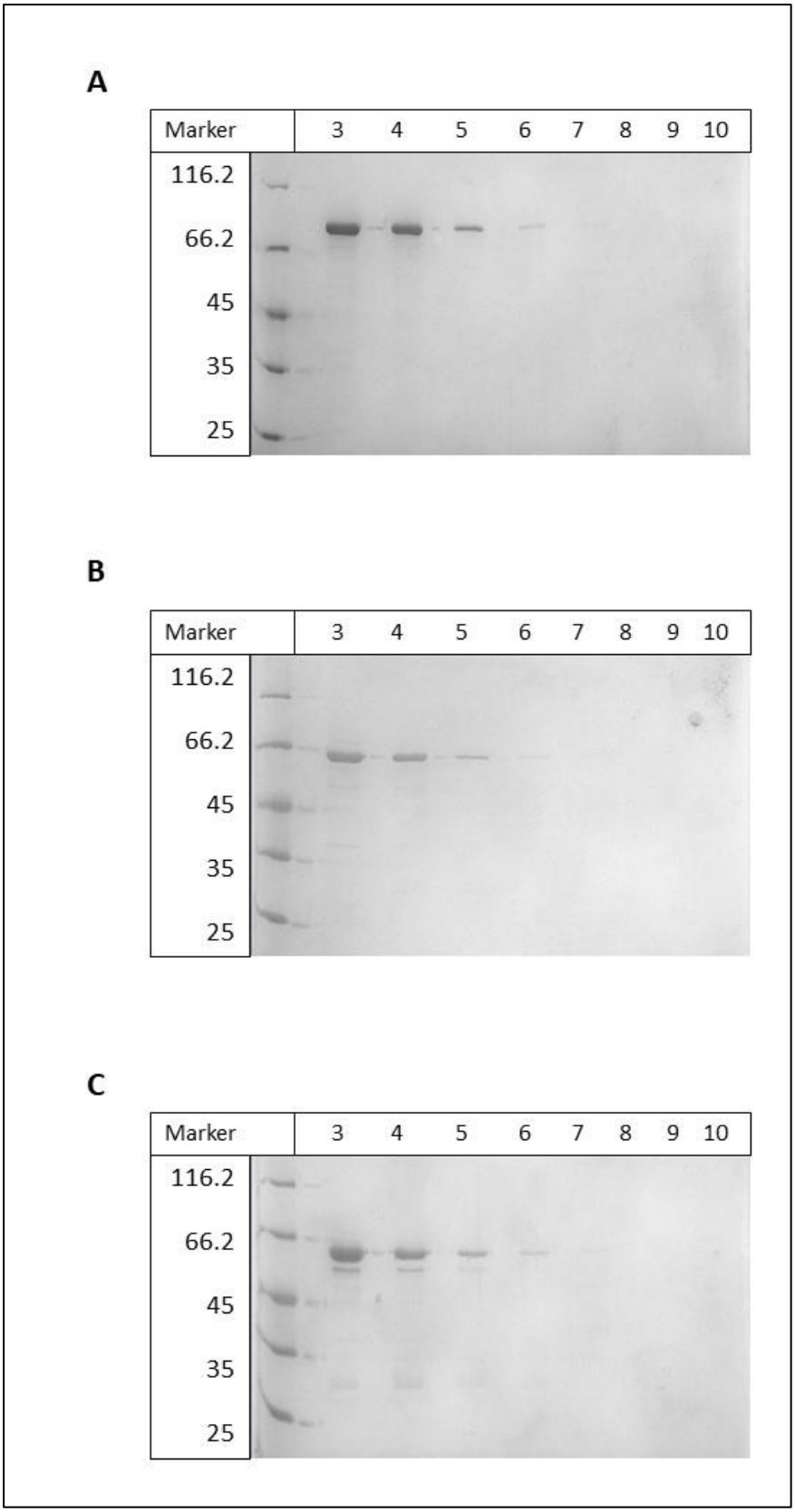


Figure 4.8. Protein purification analysis using 10 % polyacrylamide SDS-PAGE. Elution fraction 3-10 from the protein purifications (see section 2.3.4) were analysed by SDS-PAGE for A) AKDnl, B) AKRnl1 and C) AKRnl2 to determine whether protein was successfully purified. The expected sizes of each ligase are: AKDnl 78 kDa, AKRnl1 67 kDa, and AKRnl2 61 kDa.

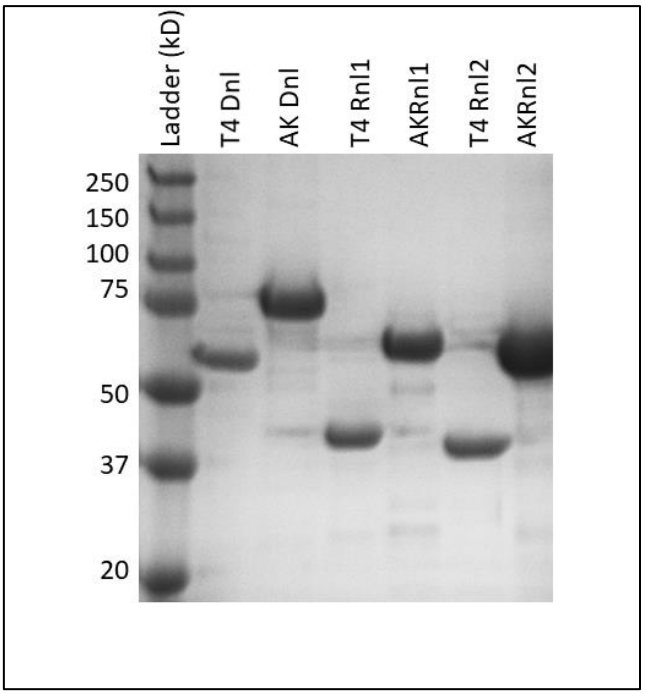


Figure 4.9. SDS-PAGE (10 % polyacrylamide) analysis of the T4 ligases and the modified AK-ligases. The addition of AK to the T4 nucleic acid ligases can be seen clearly by the size difference between the ligases. The expected sizes of each ligase are: T4Dnl 55 kDa, AKDnl 78 kDa, T4Rnl1 43 kDa, AKRnl1 67 kDa, T4Rnl2 37.5 kDa and AKRnl2 61 kDa.

4.3.3. Characterisation of the AK-ligases

4.3.3.1 End point ligation analysis

The three T4 ligases and the three AK-T4ligases were analysed using the end point ligation assay to determine whether the substrate profile for the ligases was altered. These profiles were also used to calculate the percentage ligation for each of the AK-ligases with the 8 different double stranded substrates (Figure 4.10).

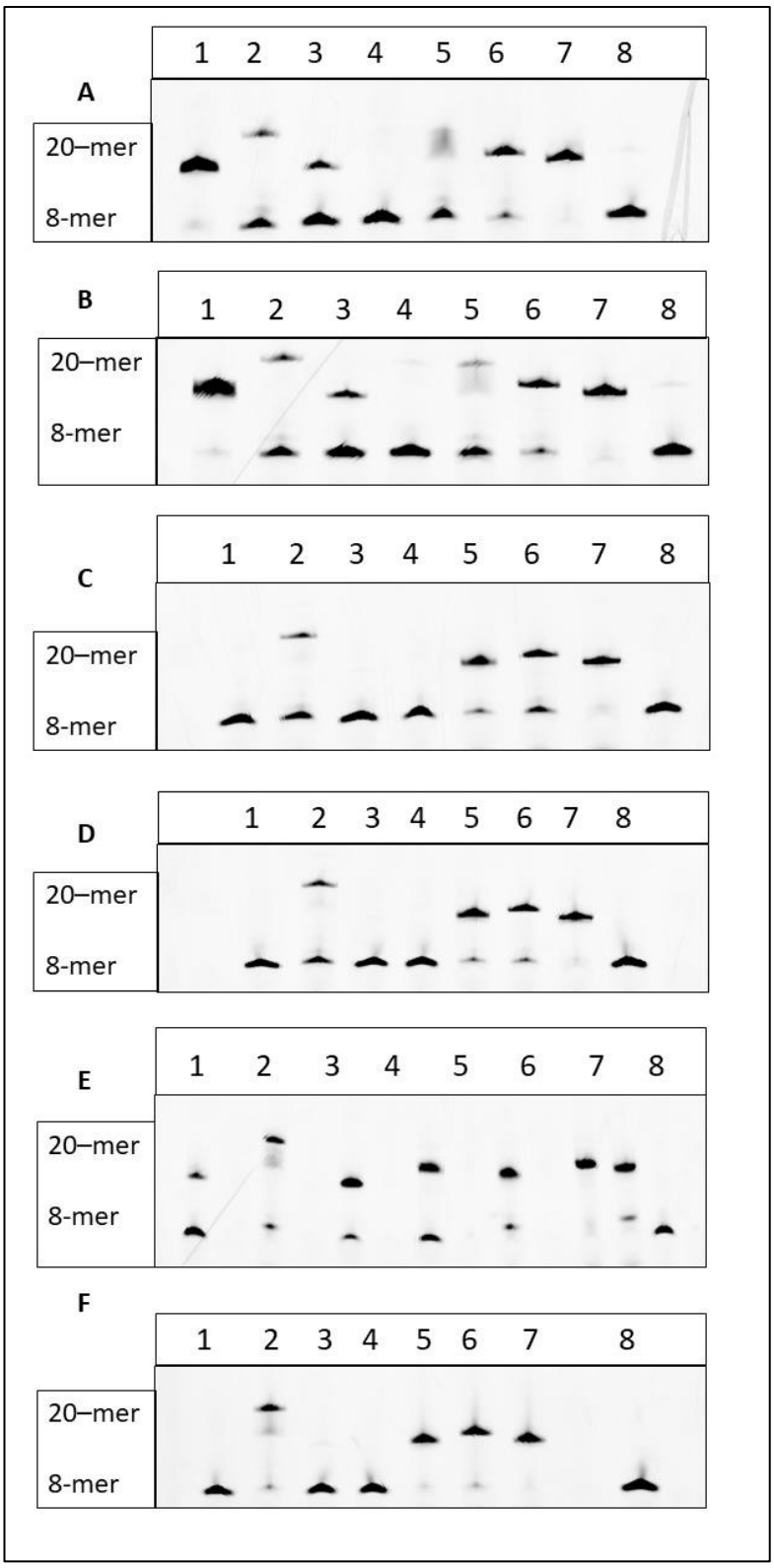


Figure 4.10. Urea-PAGE denaturation gels for the end point ligation reactions. These were used to determine substrate specificity and to calculate total ligation percentage of A) T4Dnl, B) AKDnl, C) T4Rnl1, D) AKRnl1 E) T4Rnl2 and F) AKRnl2 for each substrate. All of the ligated products are 20 bases in length, slight differences in their position in the gels was caused by the substrate composition (DNA/RNA).

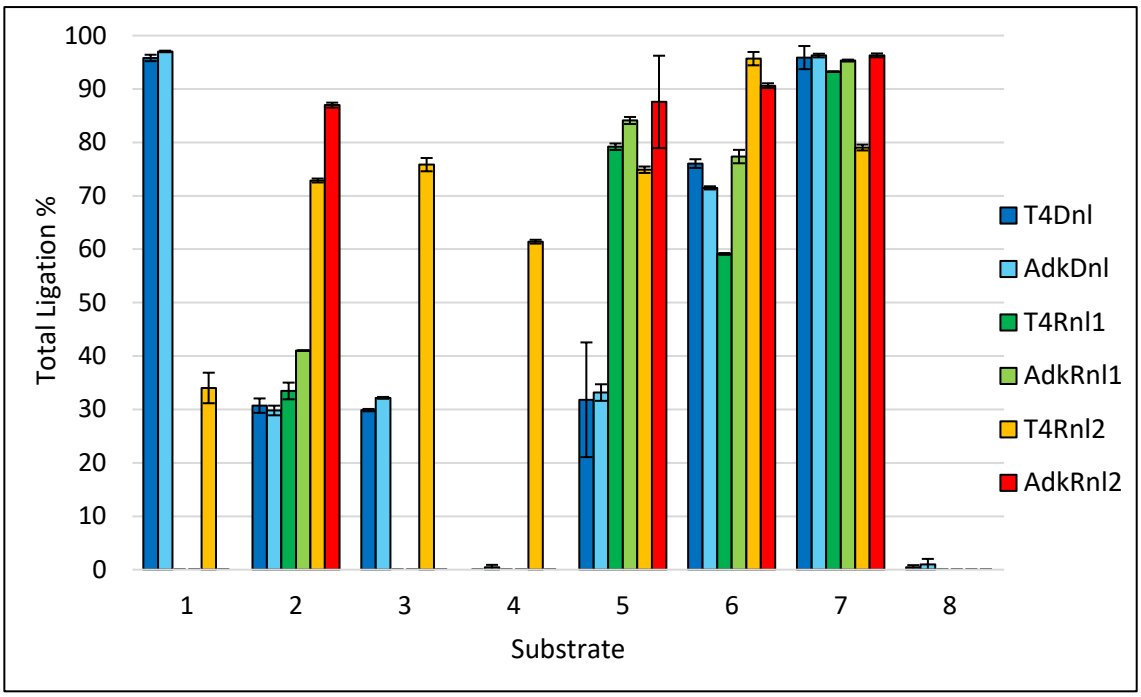


Figure 4.11. Total ligation percentage from end point analysis. Ligation with double stranded substrates 1-8, with 1 mM ATP at 37 °C for 60 minutes. In each reaction 70 pmol of ligase and 45 pmol of substrate was added. The reactions were carried out in triplicate and analysed using Urea-PAGE and the bands on the gels were quantified using ImageJ.

Addition of adenylate kinase at the N- terminal of the ligases does not appear to affect the total amount of ligated product when using a high amount of ligase compared to substrate (Figure 4.9).

4.3.3.2 Initial rate of reaction analysis.

When determining the rates of ligation, the amounts used were 2 pmol for T4Dnl and T4Rnl2 (and AKDnl and AKRnl2) and 600 pmol for T4Rnl1 (and AKRnl1) as per results by Bullard and Bowater (2006). The time course analysis was carried out as detailed in Chapter 2.4.2 for each of the enzymes and performed in triplicate to generate the error bars. The aliquots for each time point run on urea polyacrylamide gels (Figure 4.12) (See 2.4.3) and the gel image data quantified using ImageJ software (Figure 4.13).

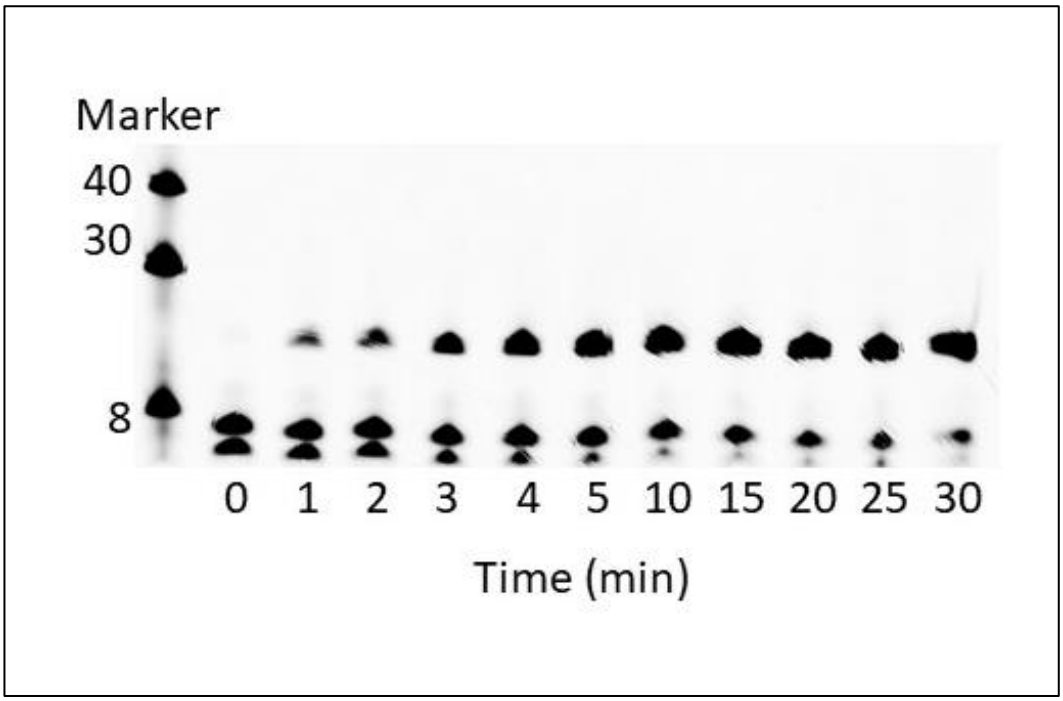


Figure 4.12. An example of time course ligation Urea-PAGE. Time course ligation for 2 pmol T4Dnl with 540 pmol substrate 7, incubated at 37 °C with ligation buffer containing ATP (Table 2.5) analysed using 15% acrylamide urea gels. The ligation reactions were run for 30 minutes as the reaction goes to completion at this point (Figure 3.18).

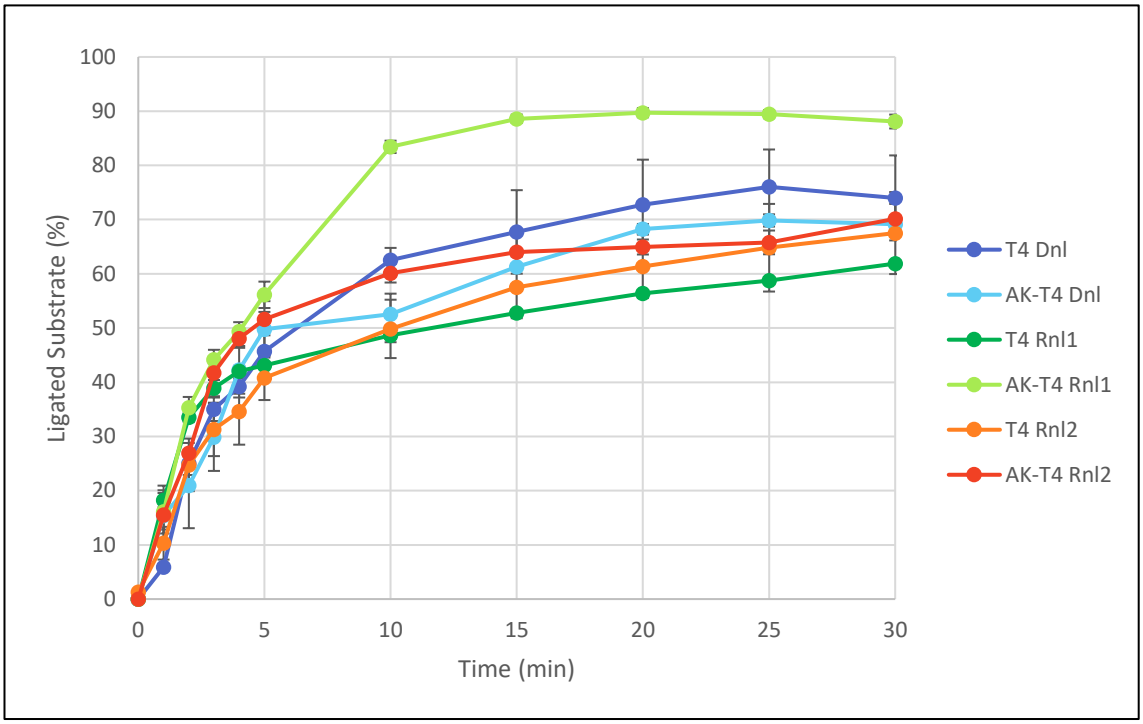


Figure 4.13. Time course ligation analysis for the T4 NA ligases and the AK-ligases. Carried out over 30 minutes at 37 ^{o}C, comparing each T4 nucleic acid ligase with the new AK-ligases; using 540 pmol of substrate 7 with 2 pmol of T4Dnl, AKDnl, T4Rnl2 and AKRnl2; and 600 pmol T4Rnl1 and AKRnl1. Each ligation was performed in triplicate to determine the standard deviation used create the error bars. The ligated and non-ligated bands from time course gels were quantified using ImageJ software to calculate the percentage of ligated substrate.

The time course ligation analysis was used to determine the rate of ligation for the initial 4 minutes of the incubated reaction. Different amounts of enzyme were used (2 pmol or 600 pmol) to obtain a similar ligation percentage over the initial 4 minutes of the reaction (Figure 4.12 and 4.13). The results are presented and discussed in Table 4.2.

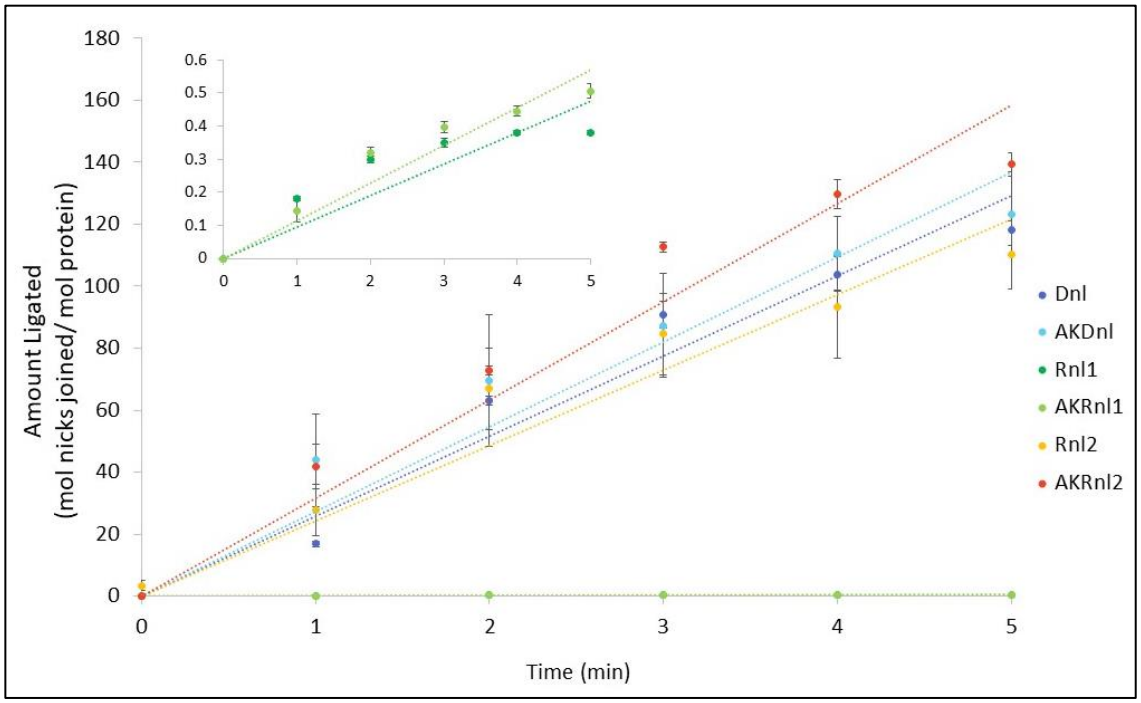


Figure 4.14. Initial ligation rate analysis of the T4 ligases and the AK-ligases. For the time course reactions 2 pmol of T4Dnl, AKDnl, T4Rnl2 and AKRnl2 was used; and 600 pmol of T4Rnl1 and AKRnl1 (inset results). The percentage of ligated product was calculated from the time course gels using ImageJ. From this it was possible to determine the amount of ligated substrate in pmol. This number was divided by the amount of substrate used in pmol to give the amount ligated (mol nicks joined/ mol protein). The amount ligated in the first 5 minutes of the time course reactions in much lower with Rnl1 and AKRnl1 (inset), Dnl, AKDnl, Rnl2 and AKRnl2 showed comparable amounts of ligated substrate 7. The initial rate of reactions demonstrate a linear relationship, indicating the constant utilisation of 20 b substrate 7 to form repaired 20 b substrate 7.

4.3.4 Adenylate Kinase Activity

Analysis of cofactor utilisation was carried out using end point ligation with 1 mM of either ADP or ATP added to the buffer. A control sample containing no ADP or ATP was also set up to demonstrate necessity of the co-factor for ligation with substrate 7 (Figure 4.15).

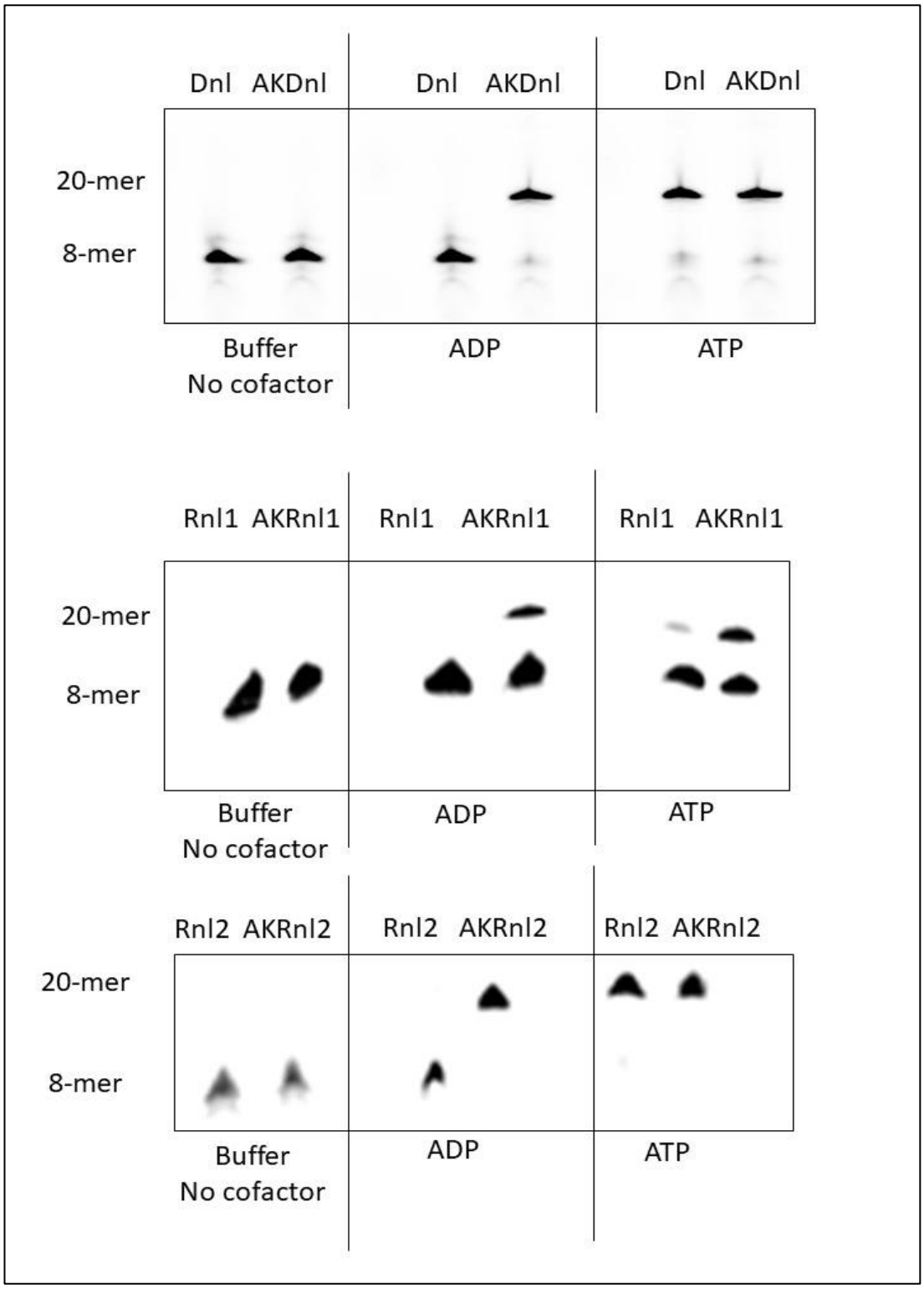


Figure 4.15. End point ligation analysis of the T4 nucleic acid ligases and the AK-ligases with ADP and ATP. To determine whether the AK-ligases could utilise both ADP and ATP for ligation. The end point ligation reactions were set up with the different ligases (70 pmol) and with 45 pmol of substrate 7 at 37 °C using buffer with no co-factor, 1 mM ADP or 1 mM ATP. The reactions were incubated for 30 minutes before quenching with 10 µl 1x stop solution and analysed using urea-PAGE. The AK-ligases could utilise ADP or ATP for ligation, while the T4 ligases could only use ATP. Ligation was not possible when there was no cofactor.

Buffer containing ADP instead of ATP was utilised by the AK-ligases, converting ADP to ATP for incorporation by the ligases for phosphodiester bond formation. Where no co-factor was included in the buffer, no ligation was noted (Figure 4.15).

4.3.5 Ligation Rates with ADP as cofactor

The initial rates of ligation were compared when using both ADP and ATP as a cofactor for the AK-ligases (Figure 4.15). All reactions were repeated in triplicate with error bars showing the standard deviation. The rates of ligation for the AK-ligases when using ADP or ATP are comparable, with AKDnl and AKRnl1 working slightly better with ADP (Table 4.2).

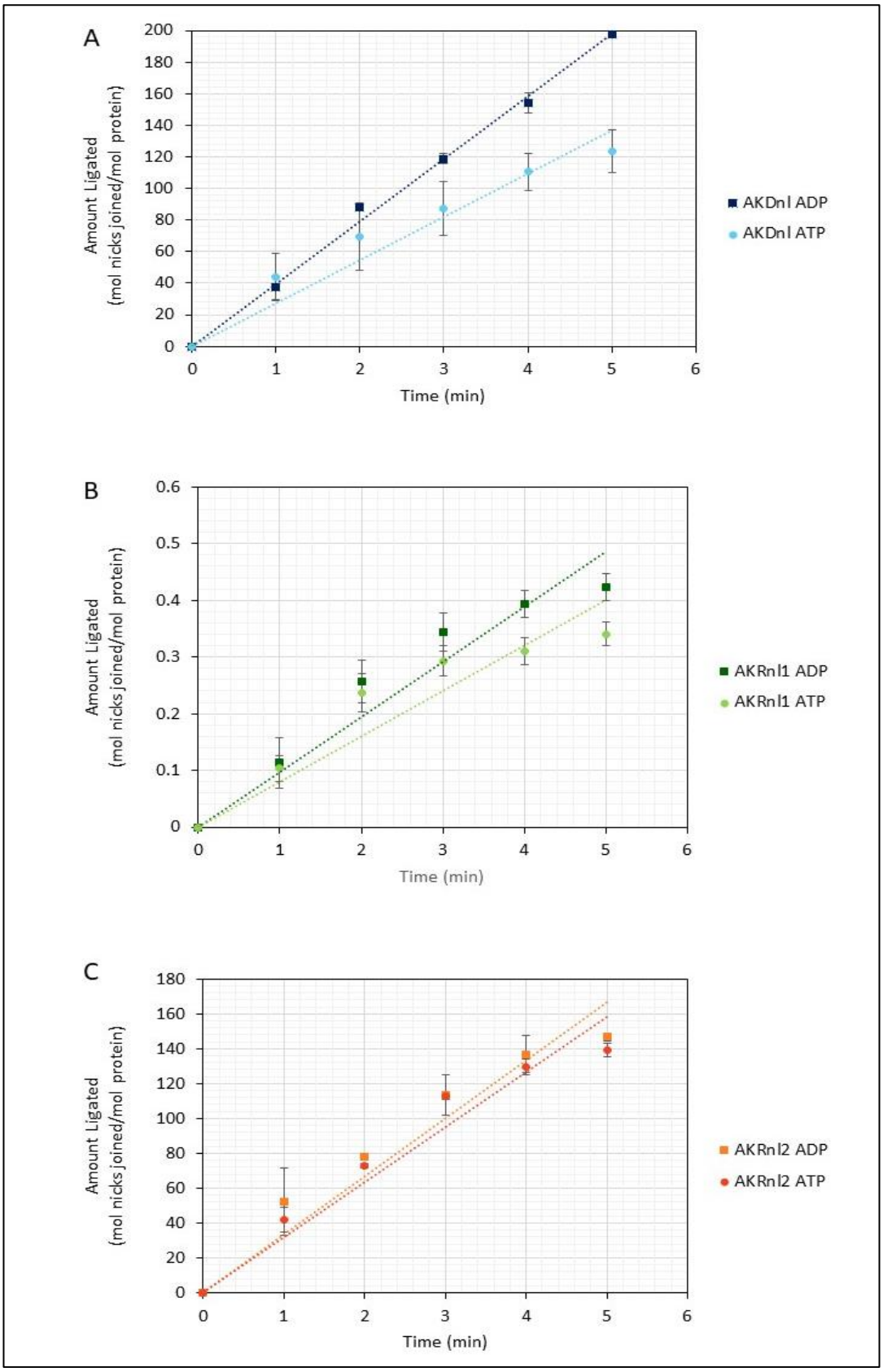


Figure 4.16. Rates of ligation for A) AKDnl, B) AKRnl1 and C) AKRnl2. All rates were carried out in triplicate with double stranded substrate 7 using either ADP or ATP as the co-factor for phosphodiester bond formation. The amount ligated by the AK-ligases is comparable when using either ADP or ATP. AKDnl and AKRnl1 had better rates of ligation when using ADP.

Table 4.2. Rates of nick joining in double-stranded substrate 7. The AK-ligases used both ADP and ATP as the co-factor for ligation. The rates of ligation suggest the AK-ligases are able to utilise both ADP and ATP for phosphodiester bond formation.

Ligase	Rate of ligation at 37 ^{o}C (mol nicks joined/mol protein/min)	
	1 mM ADP	1 mM ATP
T4Dnl	-	28.59 (+/- 2.2)
T4Rnl1	-	0.11 (+/- 0.005)
T4Rnl2	-	26.28 (+/- 7.9)
AKDnl	39.513 (+/- 1.75)	29.54 (+/- 8.3)
AKRnl1	0.11 (+/- 0.018)	0.09 (+/- 0.015)
AKRnl2	34.95 (+/- 4.68)	34.8 (+/- 1.04)

4.3.6 AKRn1 specific analysis

Total ligation percentage with AKRnl1 was notably higher than that of T4Rnl1 when carrying out the time course experiments (Figure 4.13), the addition of AK to T4Dnl and T4Rnl2 resulted in little difference in the total ligation percentage. Further analysis was carried out by using lower amounts of both ligases (60, 150, 300 and 600 pmol) with 540 pmol of substrate 7 over 30 minutes, in the case of AKRnl1 both ADP and ATP were used (Figure 4.17).

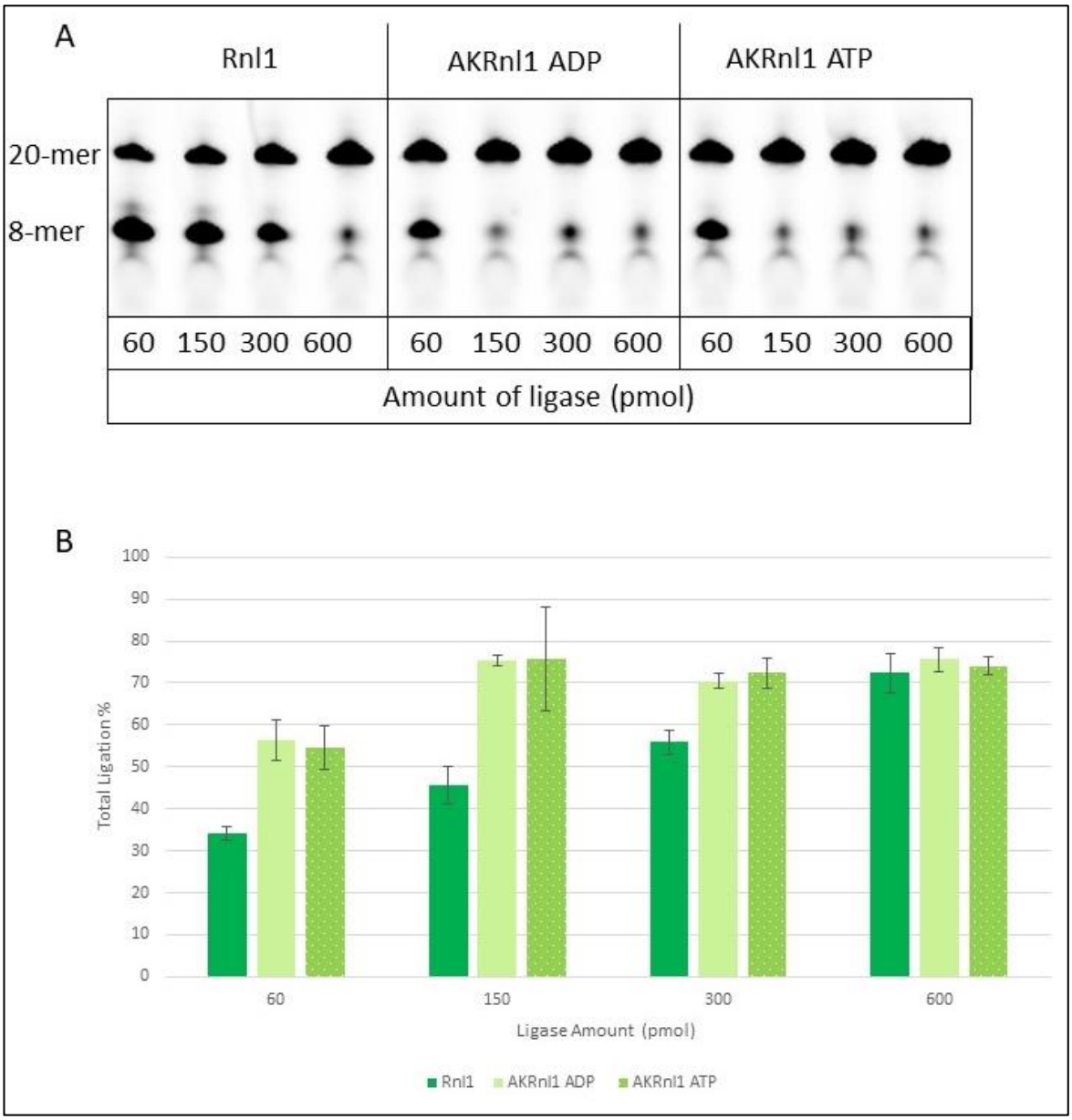


Figure 4.17. The effect of different concentrations of T4Rnl1 and AKRnl1 for ligation of standard substrate 7. A) Urea-PAGE denaturation analysis of Rnl1 and AKRnl1, with 1 mM ADP or 1 mM ATP for end point ligation analysis with 540 pmol of substrate 7; B) triplicate analysis of the end point ligation with different amounts of Rnl1 and AKRnl1. When using 150 pmol of AKRnl1 with both ADP and ATP, total ligation remains at 75 %, compared to 45 % with 150 pmol of T4Rnl1 using 1 mM ATP.

When utilising either ADP or ATP, AKRnl1 was shown to work at the same capacity when using ¼ of the amount of enzyme. To obtain 70 % ligation of substrate 7, 600 pmol of T4Rnl1 was needed, while 150 pmol of AKRnl1 was used. The same result was observed for AKRnl1 with both ADP and ATP (Figure 4.17).

Another substrate to consider is the hairpin substrate which mimics the structure T4Rnl1 would encounter *in vivo*. *In vivo* T4Rnl1 ligates nicks in the tRNA anticodon loop (Amitsur, Levitz and Kaufmann, 1987). A synthesised version of this structure, made up of 18-mer and 15-mer oligonucleotides, with fluorescein attached to the 15-mer oligonucleotide, was also tested with each of the ligases. The addition of AK to T4Rnl1 does not affect the efficiency of ligation for the hairpin substrate (Figure 4.18).

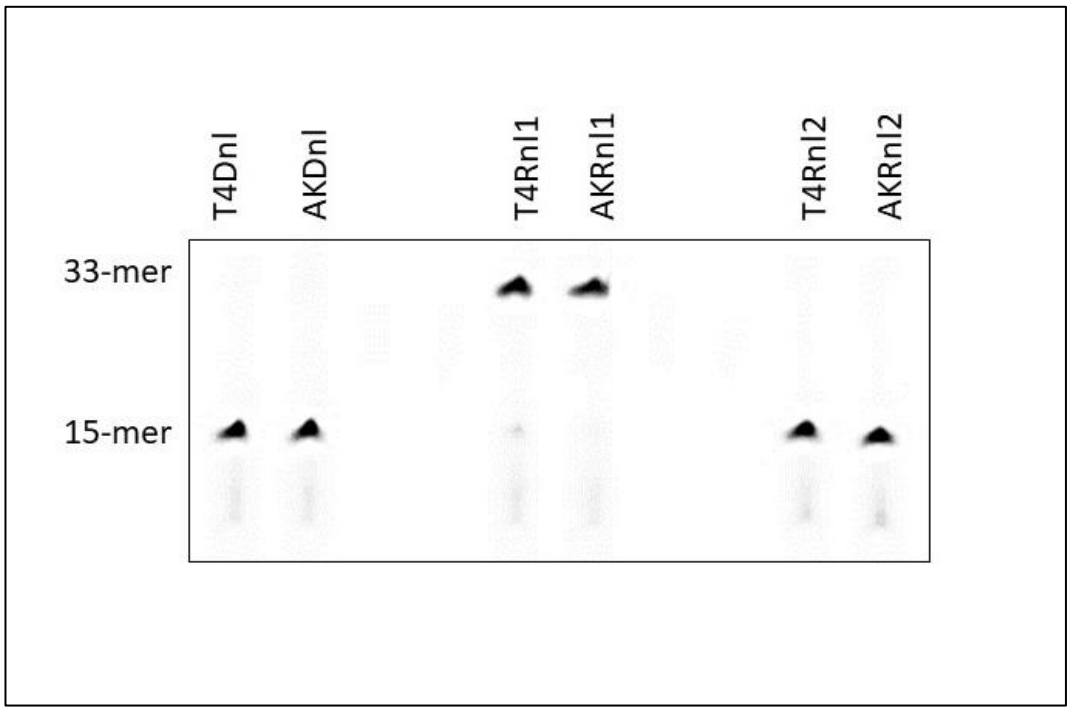


Figure 4.18. Ligation of the hair pin substrate with each of the original T4 nucleic acid ligases and the new AK-ligases. The 33-mer band indicates ligated substrate, 70 pmol of each enzyme was incubated with 45 pmol of hairpin substrate and ligase buffer with ATP at 37 °C for 30 minutes. Urea-PAGE analysis was carried out for each ligase/substrate reaction. The addition of AK to Rnl1 does not affect the ligation of the hairpin substrate.

4. 4 Discussion

The joining of adenylate kinase to the three different T4 ligase enzymes was successful in generating chimeric enzymes with altered biochemical activities. The construction of these novel enzymes has resulted in a new set of T4 RNA ligases that are able to use ADP as a cofactor for ligation instead of ATP at the same, if not a better rate than the original

ligases. The methods used to create these new functional chimeric enzymes demonstrate that traditional cloning methods, restriction enzyme digestion and ligation, are still a robust way to make constructs when inserting whole genes. As each of the T4 ligases are distinctly different from each other it is important to consider the AK-ligases as individual enzymes and asses the rates and function of these new enzymes when compared to their original ligase counterpart.

4.4.1 AK-ligase functionality and protein purification

The original paper published with AK as a solubility tag for T4Dnl reported an increased amount of soluble protein when using AK as an N- terminal tag (Liu *et al.*, 2015). Protein solubility was not something that was calculated when adding AK to the T4-ligases. The effect of the AK-ligase fusions and the concentration of IPTG required for induction was checked (Figure 4.7). It was not possible to calculate solubility from these gel because the gels do not contain known protein concentrations as a standard. Also the gel lanes are quite smeary, making it difficult to single out the bands corresponding to the ligases and ligase fusions only. Obtaining enough purified protein for each of the T4 ligases is not something I've identified as an issue and in all instances more than enough protein was purified for the ligation reactions. Comparison of purified protein for the AK-ligases when compared to the original T4 nucleic acid ligases from this work was difficult to determine because the two preparations and purifications for each protein yield such different amounts of protein. This is an issue that has been well reported (Petsev *et al.*, 2000; Bondos and Bicknell, 2003; Golovanov *et al.*, 2004) and while the addition of AK to the ligases in no way hindered the purification, it was not possible in this instance to comment on improvement to solubility of the T4 nucleic acid ligases. It is possible that differences in purification methods by the published data (Liu *et al.*, 2015) and the methods used in this work may result in differences in solubility and therefore amount of purified protein.

The three adenylate kinase fusions were able to not only ligate double stranded substrates but also used ADP as the co-factor for phosphodiester bond formation. The T4 nucleic acid ligases are usually only able to form this bond using ATP, the T4 nucleic

acid ligases with no co-factor or with ADP as the co-factor were unable to successfully ligate the double stranded substrates. The AK-ligases were able to ligate the double stranded substrates with both ADP and ATP, confirming functionality of both parts of these new fusion ligases. In all instances, both the T4 NA ligases and the AK-ligases were unable to create a phosphodiester bond between the 8-mer and 12-mer oligonucleotides when no co-factor was included in the ligation buffer. There is no difference in ligation profile from end point ligation for the AK-ligases and the original T4 ligases with different double stranded DNA/RNA combination substrates.

4.4.2 Rates of ligation for the AK-ligases

The rates of the new AK-ligases are comparable to the rates of the unmodified T4 nucleic acid ligases. The addition of another functional enzyme at the N- terminus of the T4 nucleic acid ligases permits ligation by the ligase part and also adds the function of converting ADP to ATP for phosphodiester bond formation.

The initial rate of AKRnl1 is similar to that of T4Rnl1, however it is very clearly seen that the total amount of ligation carried out by this enzyme when carrying out the time course ligation was higher than the T4Rnl1. Further analysis into this showed that in fact ligation could continue to go to 70 % using only 150 pmol of AKRnl1 in the reaction, which was ¼ of the amount required when using T4Rnl1 (Figure 4.14). The rate of ligation for AKDnl and AKRnl2 were both slightly increased when compared to the unmodified T4 ligase (Figure 4.11). Addition of AK to the N- terminus of the ligases, does not alter the rate of ligation of the AK-ligases.

4.4.3 ADP vs ATP as a cofactor

For all of the AK-ligases, the ligation assay was carried out using ADP in place of ATP as the cofactor in the buffer. For all AK-ligases it was clear that in the presence of ADP with the adenylate kinase attached the ligase portion of the enzyme was able to ligate, the T4 NA ligases were used as the control for this with all enzymes tested with no cofactor, ADP and ATP (Figure 4.12). The rates of ligation for each new AK-ligase with ADP or ATP were then performed in triplicate, an interesting observation for both AKDnl and AKRnl1

is that these ligases have an improved rate of ligation when using ADP as the co-factor for ligation (Figure 4.13). This may be due to localisation of ATP for these new enzymes, it was immediately ready for the ligases to use in phosphodiester bond formation. The T4 nucleic acid ligases create a ligase-ATP intermediate before forming the phosphodiester bond. The lysine residue in motif I binds to the ATP, localisation of the ATP for use by the ligase may contribute to improved activity as the same reaction using ATP only.

4.5 Conclusion

This is the first time chimeric versions of the T4 RNA ligases have been constructed and shown to be functional. A lot of work has previously focussed on changing or modifying T4 DNA ligase as this is a more commonly used enzyme in molecular biology techniques. The T4 RNA ligases are also useful tools in a number of techniques and as such it is important to explore ways to improve these ligases. This work has demonstrated that addition of another protein at the N- terminus of the T4 nucleic acid ligases yields functional enzyme fusions. In this instance the addition of AK allows for phosphodiester bond formation from ADP, where AK can readily convert ADP to ATP for use by the ligase part of the fusion.

Not only are these fusion enzymes functional, but in the case of AKRnl1 a lower amount of ligase can be used to obtain the same ligation result as the unmodified version. From an industrial perspective, this provides a way to lower costs for production of this ligase, from each round of purification of AKRnl1, ¼ of the amount of ligase is needed for successful ligation, therefore a single purification technically allows for 4 times more ligation activity. The additional function for all of the AK-ligases to utilise ADP and ATP was also desirable and remedies some issues with co-factor degradation in buffer (Hulett, 1970). This successful protein-protein fusion opens the way to other fusions with the T4 nucleic acid ligases and the possibility to remedy some of the issues with these ligases by addition of another functional protein to the ligases.

Chapter 5
Fusion of T4 Polynucleotide Kinase and NA Ligases

Chapter 5: Fusion of T4 Polynucleotide Kinase and NA Ligases

5.1.1 T4 Polynucleotide Kinase

Polynucleotide Kinase (T4Pnk) is an enzyme found in the genomes of both T7 and T4 bacteriophages. It was originally discovered in *E. coli* that had been infected with T-even phage, specifically T2 and T4, which were able to transfer the γ-phosphate from ATP to the 5'-hydroxyl terminus of a polynucleotide (Richardson, 1965). T4Pnk is a transferase responsible for converting 5'OH and 3'PO_4 to 5'PO_4 and 3'OH from broken ends of DNA and RNA which primarily involves the removal and reattachment of phosphate (Wang and Shuman, 2001), once altered it is possible for the break to be repaired by the T4 ligases. Polynucleotide kinases/phosphatases with similar biochemical functions are found in other organisms; there are conserved residues amongst these enzymes but there are significant differences in the DNA binding sites especially when comparing eukaryotic and phage polynucleotide kinases/phosphatases (Galburt *et al.*, 2002; Weinfeld *et al.*, 2011).

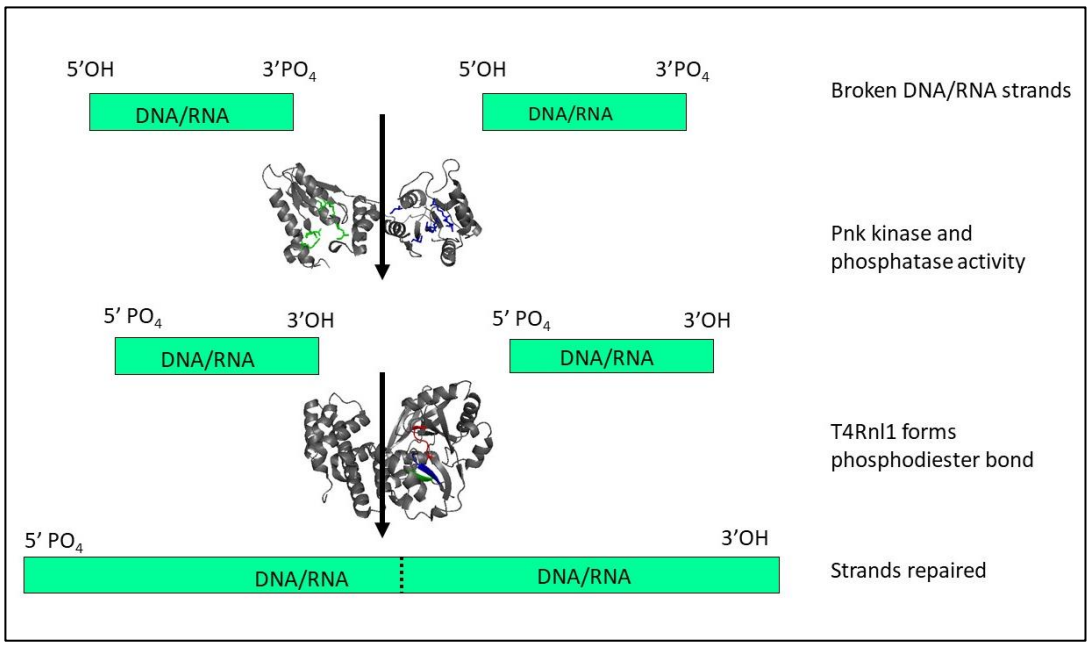


Figure 5.1. T4Pnk and T4Rnl1 DNA/RNA strand repair and ligation. For phosphodiester bond formation between the break, the PO_4 must be at the 5' position. In *E.coli* this DNA/RNA break can be found in the tRNA anticodon loop. Structures, 1LTQ (T4Pnk) and 2C5U (T4Rnl1), illustrated using CCP4MG software.

T4Pnk is a vital enzyme required during viral replication and phosphodiester bond formation alongside the T4 DNA and RNA ligases. The T4Pnk protein *in vivo* is part of the virus response to host defences: when the T4 bacteriophage infects *E. coli*, the bacterium responds by causing breaks in its own tRNA anticodon loop and initiating its own death (Sirotkin *et al.*, 1978; Galburt *et al.*, 2002). T4 phages have short polypeptides on their surface called Stp polypeptides. These Stp polypeptides activate an anticodon nuclease which cleaves the bacterial tRNA and, as a consequence of host cell death, viral protein synthesis is also halted (See section 1.5.2) (Penner *et al.*, 1995; Blanga-Kanfi *et al.*, 2006). To combat this host response system the T4 bacteriophage repairs these breaks using the T4Pnk protein and T4Rnl1 (Figure 5.1). The T4Pnk is able to correctly prime the ends of DNA and RNA to ensure the 5' PO_4 and 3' OH are available for phosphodiester bond formation by the ligase (Penner *et al*., 1995; Galburt *et al.*, 2002; Wang *et al.*, 2002).

```
        10          20          30          40          50
MKKIILTIGC  PGSGKSTWAR  EFIAKNPGFY  NINRDDYRQS  IMAHEERDEY
        60          70          80          90         100
KYTKKKEGIV  TGMQFDTAKS  ILYGGDSVKG  VIISDTNLNP  ERRLAWETFA
       110         120         130         140         150
KEYGWKVEHK  VFDVPWTELV  KRNSKRGTKA  VPIDVLRSMY  KSMREYLGLP
       160         170         180         190         200
VYNGTPGKPK  AVIFDVDGTL  AKMNGRGPYD  LEKCDTDVIN  PMVVELSKMY
       210         220         230         240         250
ALMGYQIVVV  SGRESGTKED  PTKYYRMTRK  WVEDIAGVPL  VMQCQREQGD
       260         270         280         290         300
TRKDDVVKEE  IFWKHIAPHF  DVKLAIDDRT  QVVEMWRRIG  VECWQVASGD

F
```

Figure 5.2. The amino acid sequence for T4Pnk. The residues required for 5' kinase activity (green) and 3' phosphatase activity (blue) (Galburt *et al.*, 2002; Wang, Lima and Shuman, 2002). Amino acids are designated following the standard single letter abbreviations.

T4Pnk is involved in two distinct biochemical activities, kinase and phosphatase which are separated by localisation to either the N- or C- terminus (Figures 5.2 and 5.3). T4Pnk requires magnesium for these two distinct activities (Galburt *et al.*, 2002). The N-terminus is responsible for kinase activity through the addition of phosphate. The amino

acids identified as crucial for this reaction are K15, S16, D35, R38 and R126 (Figure 5.2); mutagenesis of these residues to alanine resulted in a total loss of kinase activity but had no effect on phosphatase activity (Galburt *et al.*, 2002; Wang, 2002). The active site tunnel includes residues K15, S16, D35 and R126, the AMP phosphate makes contact with the side chains of R38 (Galburt *et al.*, 2002).

The phosphatase activity, removal of phosphate, is localised to the C- terminus; mutagenesis studies have shown that the amino acids involved in this site's activity are D165, D167, R176, R213, D254 and D278 (Figure 5.2). Mutation of these residues to alanine inactivate phosphatase activity but had no impact on kinase activity (Galburt *et al.*, 2002; Wang, Lima and Shuman, 2002). When transferring phosphate, D165 is responsible for attacking the 3' phosphorous, the phosphate oxygen's make contact with D167 which goes on to form a salt bridge with R213 (Galburt *et al.*, 2002). D254 is involved in substrate binding, D278 is involved in metal binding and was modelled in close proximity to the magnesium ion; the role of R176 is unclear, it is furthest from the binding site (Galburt *et al.*, 2002).

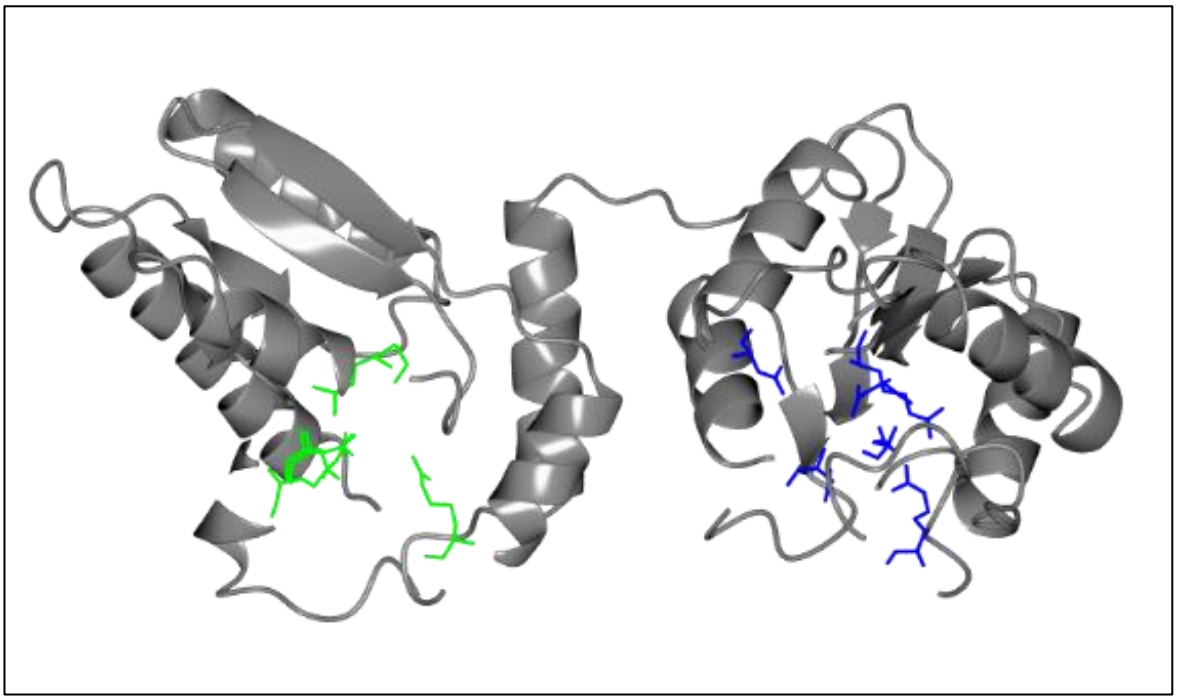


Figure 5.3. The crystallographic structure of T4Pnk. The N- terminus with amino acids responsible for kinase activity highlighted (green) and the C- terminus with the phosphatase activity amino acids highlighted (blue). PDB file 1LTQ illustrated using CCP4MG software.

5.1.2 Molecular techniques that use Polynucleotide Kinase

T4Pnk is also widely used for *in vitro* molecular biology protocols that manipulate DNA and RNA, with the following processes being identified: end-labelling of DNA and RNA for use as probes and nucleic acid sequencing (Chaconas and van de Sande, 1980); addition of 5'-phosphates to oligonucleotides to allow subsequent ligation into DNAs and removal of 3'-phosphoryl groups (Weinfeld *et al.*, 2011).

5.1.3 A T4Pnk-ligase fusion for Next Generation Sequencing

A type of small RNA (sRNA) molecule called micro RNAs (miRNA) have a significant role in gene expression; miRNAs are usually 17-22 nucleotides long and can bind to mRNA to prevent translation or cause mRNA decay (Motameny *et al.*, 2010; Martin, Wani and Steptoe, 2014). A number of pathways involve miRNAs and recent research has started quantifying different miRNAs associated with a number of pathological conditions – offering a non-invasive method for disease identification and monitoring (Zhao *et al.*, 2010; Pogue *et al.*, 2014; Hill and Lukiw, 2016; Agrawal *et al.*, 2018). Identification and

quantification of miRNAs relies heavily on the technique of next generation sequencing (NGS) (Xu *et al.*, 2015). NGS is comprised of 4 main steps: library preparation, cluster generation, sequencing and data analysis. One of the first steps in library preparation of miRNAs is adenylation of adapters and miRNAs, and ligation of the adapters to the miRNAs (Figure 5.4). This process uses the T4 nucleic acid ligases to attach the adapters, and the choice of ligase is often dependent on manufacturer's products, master mixes and protocols (Head *et al.*, 2014).

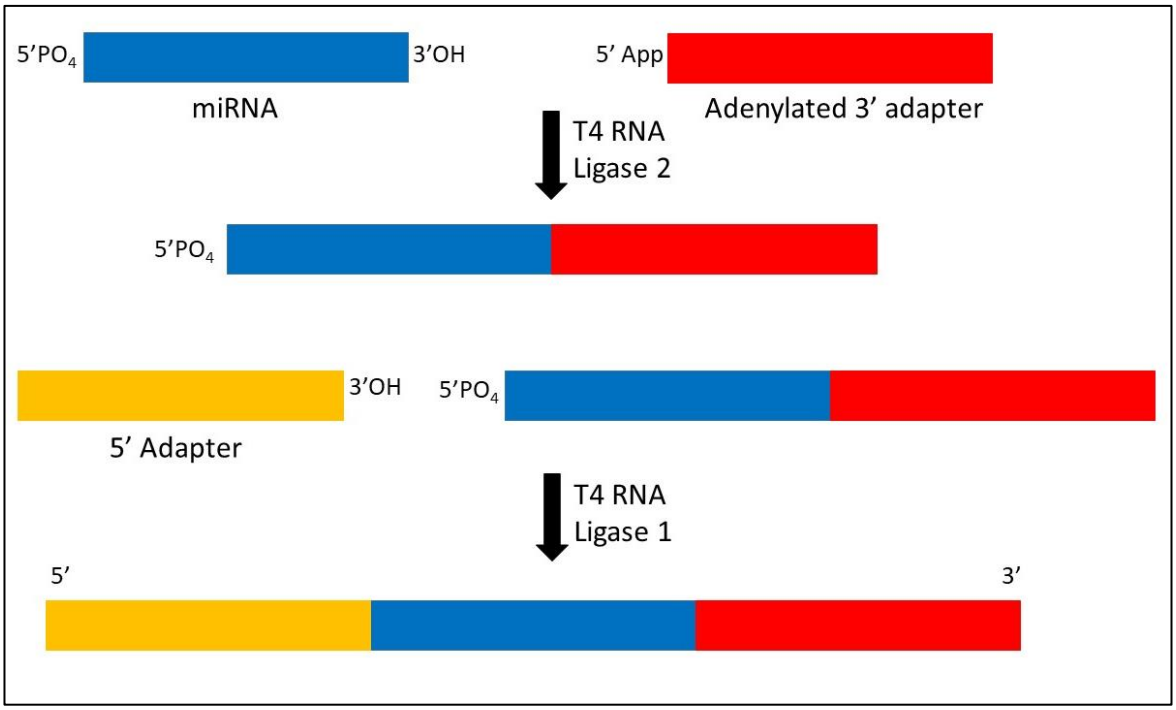


Figure 5.4. The ligation of adapters to miRNAs. The T4 RNA ligases are both used for adapter ligation to miRNAs. A truncated version of T4Rnl2 is used for the first step to ensure only pre-adenylated 3' adapters can be ligated, and this is followed by ligation with T4Rnl1 to ligate the 5' adapter (Head *et al.*, 2014).

The changes in levels of miRNAs have been suggested as a diagnostic tool for different cancers, including breast, pancreatic and ovarian cancers (Mitchell *et al.*, 2008; Taylor and Gercel-Taylor, 2008; Zhao *et al.*, 2010). When sequencing miRNAs though it was noted that there were frequent miscalculations of the number of miRNAs, sometimes by as much as 4 orders of magnitude, resulting in differences between the actual abundance of miRNAs and the calculated amount (Hafner *et al.*, 2011). Adapter addition

to miRNAs is only possible when the ends are correctly primed and the miRNA has 5'PO_4 (Head *et al.*, 2014). In instances where the 5' end is not PO_4 ligation cannot occur and quantification results would show lower levels of miRNAs being detected. From the current NGS methods it is not possible to determine whether low levels of quantified miRNA are attributed to amounts of miRNA produced in the sample or whether the miRNA was not fully ligated to the adapter and, therefore, not detected. This problem would not be apparent with current methodologies as the focus has been on adapter ligation bias (Xu *et al.*, 2015) and modifying adapters to improve adapter/miRNA ligation for miRNA quantification.

Parkinson's Disease (PD) is often an age related condition where in older age the Dicer enzyme, involved in miRNA maturation, may be down regulated resulting in a reduction in the number of miRNAs processed (Vinnikov and Domanskyi, 2017). Another enzyme involved in miRNA maturation is the ribonuclease II enzyme, Drosha, which cleaves and determines the 5' and 3' nucleotides. It seems sensible to deduce that if either Dicer or Drosha are not working correctly then there may be mis-priming of the miRNA (Tulay and Sengupta, 2016). These miRNAs would not be detected using current methodologies as the T4 ligases would not be able to ligate the adapters to the miRNA for quantification or detection. A study has shown that removing the proteins required for preparation of mature miRNAs has an effect on miRNA levels and may lead to incorrect end priming (Kumar *et al.*, 2007). At present the protocols for adapter-miRNA ligation do not include a miRNA priming step, which would ensure that all miRNAs in the sample could be ligated and detected for quantification. As levels of miRNA have been attributed to certain diseases it would be of interest to understand whether this change in miRNA amount is an actual lower amount of miRNA or whether miRNAs are incorrectly processed and, as such, use the latter as a possible therapeutic target.

5.1.4 Other uses for a T4Pnk-ligase fusion

It is proposed that a chimeric enzyme, which is a combined version of T4Pnk and the T4 nucleic acid ligases, would ensure the ligation step in adapter attachment during NGS is successful and any bias encountered is not due to incorrectly primed ends, saving time

and money for researchers using this technique. The T4Pnk-fusions would become a time saving step for any methodologies that use T4Pnk and the T4 ligases at separate steps.

The uses of this chimeric enzyme would not be limited to NGS and could be used for ligation in dumbbell shaped DNA probes too (Qing *et al.*, 2017). Another technique that utilises both T4Pnk and T4Rnl1 is high-throughput sequencing of RNA isolated by crosslinking immunoprecipitation (HITS-CLIP) (Darnell, 2010). HITS-CLIP is a technique that uses both T4 RNA ligases and T4Pnk to map protein-RNA binding sites and RNA modification (Figure 5.5). This technique has successfully demonstrated RNA-protein interactions in the brains of mice and was also used to look at RNA binding capacity of a 90 kDa RNA binding protein (RNABP), a mutated protein in a recessively inherited Parkinson's-like disorder (Chi *et al.*, 2009; Darnell, 2010). HITS-CLIP has also been used to identify miRNA-mRNA interactions (Thomson, Bracken and Goodall, 2011).

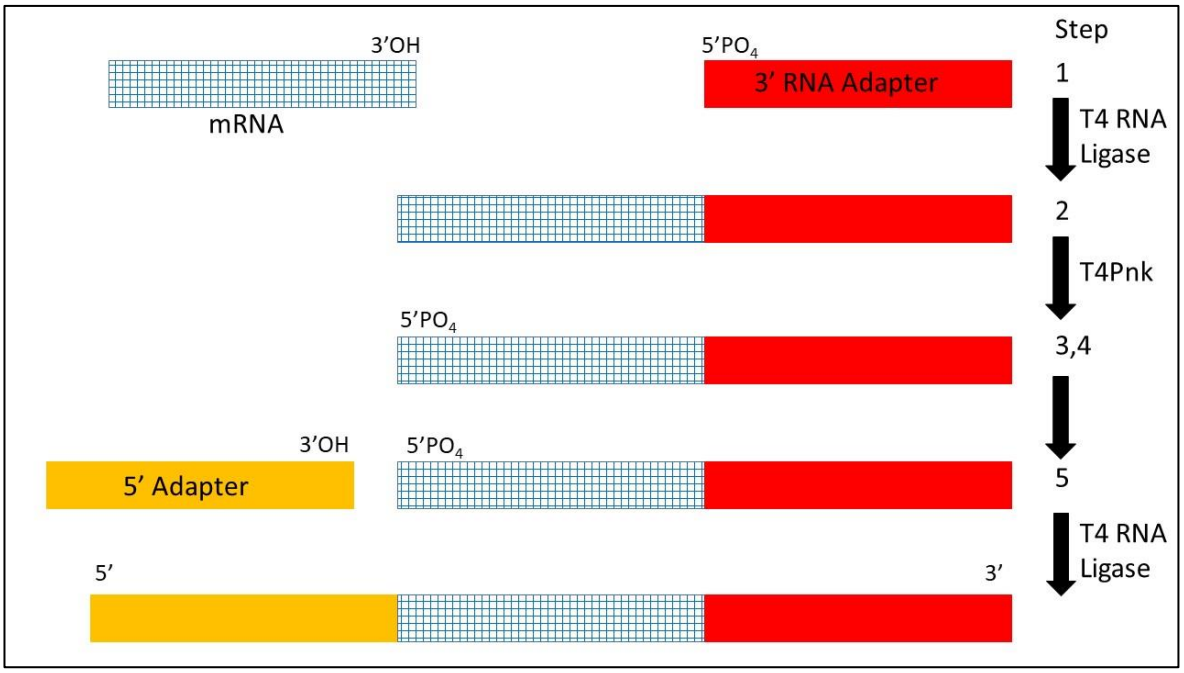


Figure 5.5. HITS-CLIP mRNA protocol (Thomson, Bracken and Goodall, 2011). Each step modifies the mRNA for sequencing. Prior to step 1 the mRNA is cross linked, immunoprecipitated and treated with RNAse A and Alkaline Phosphatase. Step 1: Addition of the 3' RNA adapter with either T4 RNA ligase; 2) T4Pnk addition to prime the 5' end for ligation; 3 and 4) SDS-PAGE and gel extraction; 5) Addition of the 5' adapter using a either T4 RNA ligase. After step 5 the mRNA undergoes reverse transcription, PCR and deep sequencing.

Another more common use for both T4Pnk and the T4 nucleic acid ligases is blunt end ligation (Gao *et al.*, 2015). Blunt ended inserts from restriction enzyme digestion are already phosphorylated but blunt ended inserts produced by PCR need to be phosphorylated and this is often achieved with T4Pnk (Gao *et al.*, 2015). After this step the ligation reaction can be set up and T4Dnl is used for this ligation step. As there are a number of methodologies that use both T4Pnk and the T4 nucleic acid ligases, it would be beneficial to have fusion T4Pnk-ligases. This is particularly the case when these enzymes are often used in subsequent steps because a fusion T4Pnk-ligase would mean a one-step process, saving time and costs.

5.2 Construction of T4Pnk-ligases

5.2.1 Recombinant Plasmid Construction

The T4 polynucleotide kinase gene (*pseT*) was ordered from (Genscript) and synthesised on the plasmid pUC57. The *pseT* gene was added to the N- terminus of the T4 NA ligase plasmids (pRB255, pRB256 and pRB257) by restriction digestion and ligation.

5.2.1.1 Restriction Digestion and PCR

The synthetic gene was designed with the *NcoI* cut site at the N- terminus and *NdeI* at the C- terminus. This allowed double digestion using *NdeI* and *NcoI* (ThermoFisher) and ligation cloning into the pET16b vectors containing the different T4 NA ligases (pRB255, pRB256 and pRB257). The new constructs were denoted pRB560 (PnkDnl) pRB561 (PnkRnl1) and pRB562 (PnkRnl2) (Figures 5.6, 5.7 and 5.8). The *pseT* gene was also cloned into pET16b on its own, named pRB576. To isolate the *pseT* gene for cloning into pRB576, PCR was used to insert the *BamHI* cut site at the C- terminus of the gene (Table 5.1). The NEB Q5 mutagenesis kit was used as per the manufacturer's instructions, the primers were designed and ordered from Eurogentec and an annealing temperature of 59 °C was used. This allowed for double digestion using *NcoI* (ThermoFisher) and *BamHI* (ThermoFisher) to isolate the *pseT* gene, the pET16b plasmid was also digested with these enzymes. The reactions were set up with 12 µl plasmid DNA (up to 1 µg), 1.25 µl

BamHI, 1.25 µl *NcoI*, and 2.5 µl FD buffer. The digestions underwent incubation at 37 ^{o}C, 180 rpm for up to 2.5 hours.

Table 5.1. The primers designed for *BamHI* insertion at the C- terminus of *pseT*; the lower case sequences (red) is the *BamHI* cut site.

Primer Name	Primer sequence 5' to 3'
PnkBamHI F	ggatccCATATGATTCTTAAAATTCTGAACG
PnkBamHI R	AAAATCTCCCGAAGCGAC

5.2.1.2 Agarose Gel Extraction

The digested linearised plasmids (pRB255, pRB256 and pRB257) and the *pseT* gene were extracted using the Bioline Isolate II PCR & Gel Kit (see Section 2.2.4). The manufacturer's protocol was followed and the DNA quantified using the Thermo Scientific Nanodrop one quantification system.

5.2.1.3 Ligating the *pseT* gene into pRB255, pRB256 and pRB257

The *pseT* insert and pET16b cut plasmid were ligated using T4 DNA ligase using the same process as the AK-ligases (see Section 4.2.3). The resultant ligation was transformed into 50 µl α-select *E. coli* cells (available from Bioline, see Section 2.2.6).

5.2.1.4 Confirmation of the recombinant *pseT*-ligase plasmids

The new plasmids containing *pseT* at the N- terminus of each of the T4 nucleic acid ligases (pRB560, pRB561 and pRB562) (Figures 5.6, 5.7 and 5.8) underwent double digestion with *NcoI* and *BamHI* (see Section 2.2.5) to confirm a change in size of the sequence between these cut sites (Figure 5.10). The plasmids were also sent to Eurofins for Sanger sequencing (i54 Business Park, Valiant Way, Wolverhampton, WV9 5BG, UK).

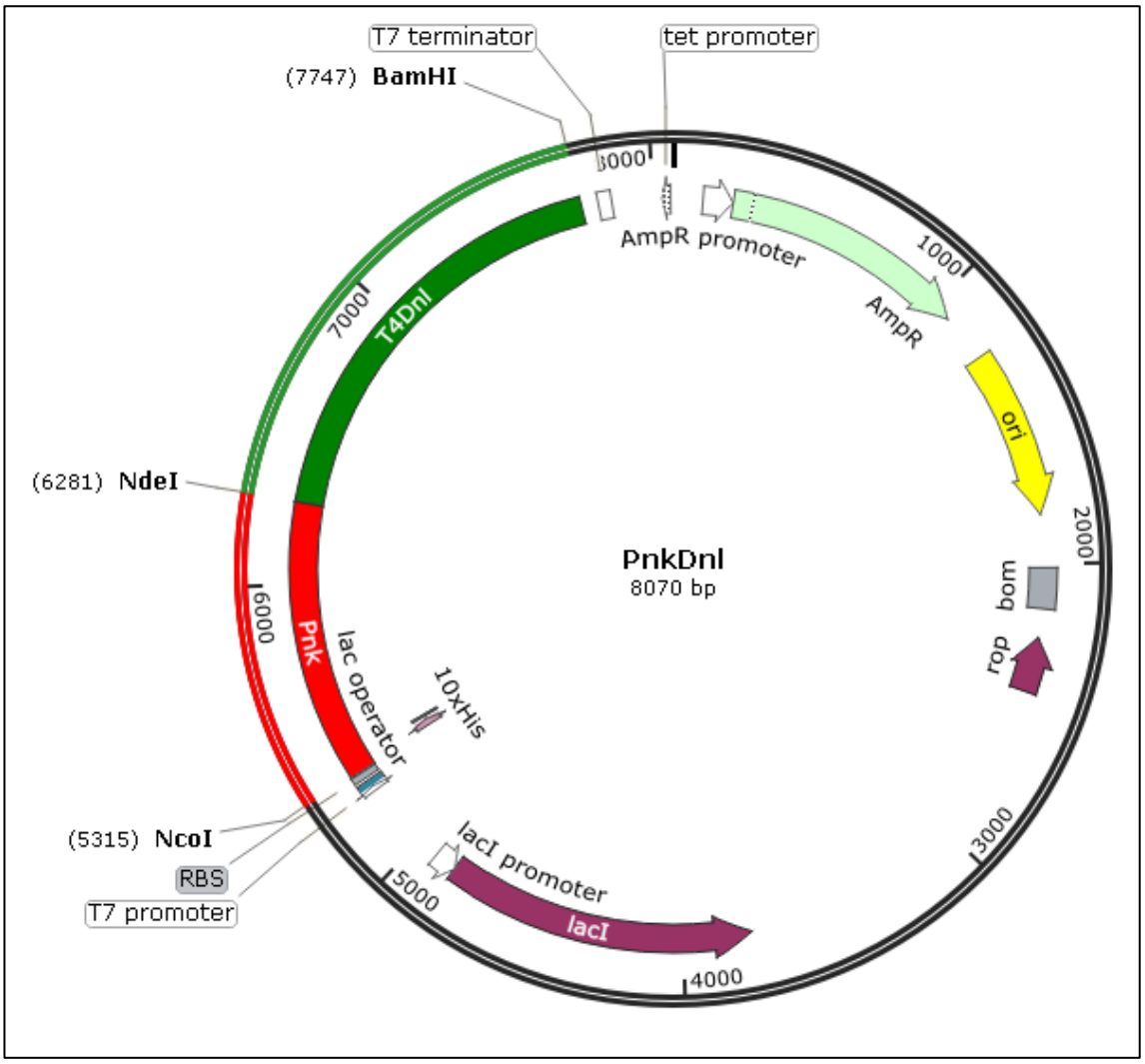


Figure 5.6. Plasmid map for the fusion of Pnk and T4Dnl to make PnkDnl (pRB560). This map details the addition of T4Pnk at the N- terminus of T4Dnl using the *NcoI* and *NdeI* restriction sites. Plasmid maps were created using SnapGene software.

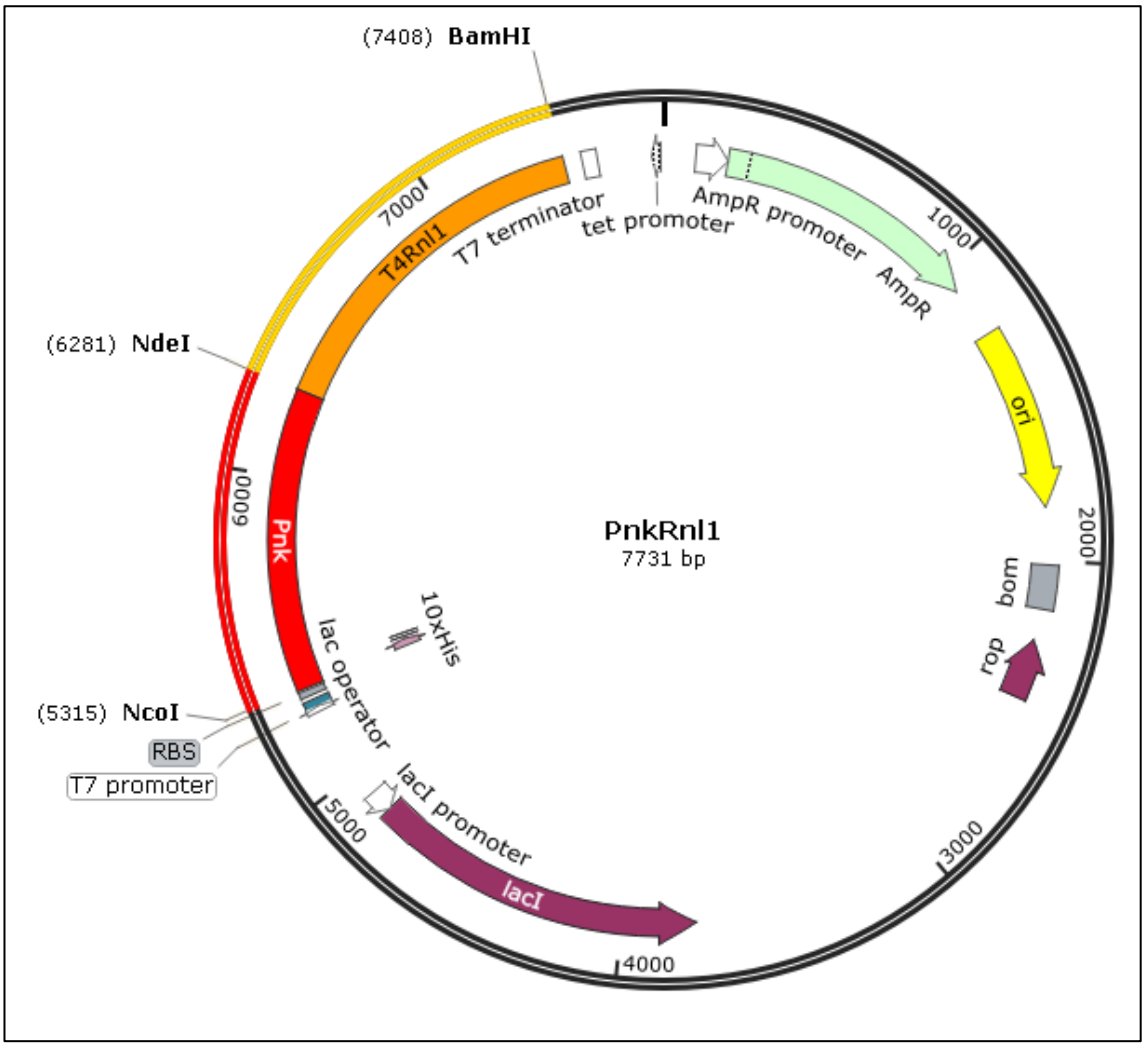


Figure 5.7. Plasmid map for the fusion of Pnk and T4Rnl1 (pRB561). The map details the addition of T4Pnk at the N- terminus of T4Rnl1 using the *NcoI* and *NdeI* restriction sites. Plasmid maps were created using SnapGene software.

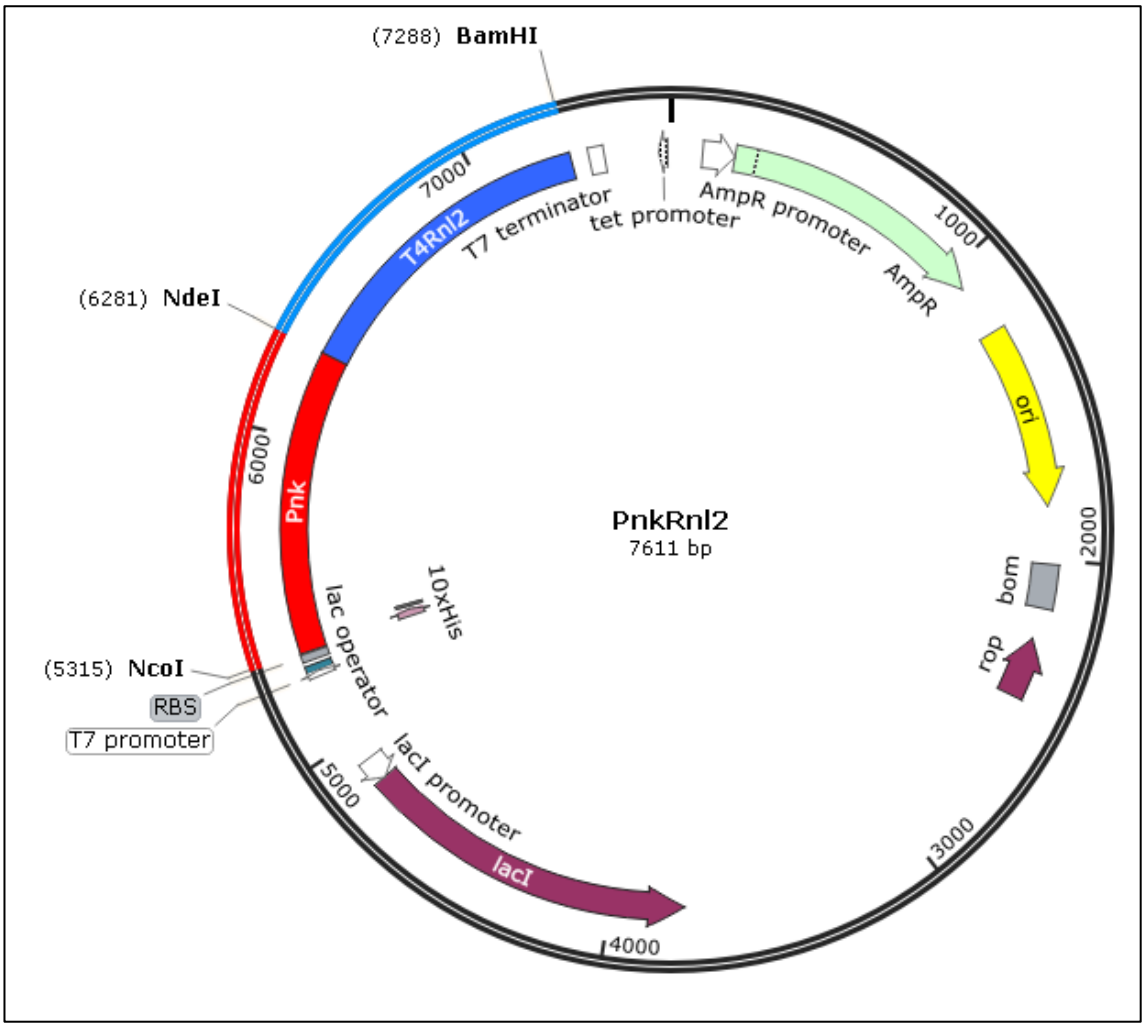


Figure 5.8. Plasmid maps for the fusion of Pnk and T4Rnl2 (pRB562). The map details the addition of T4Pnk at the N- terminus of CT4Rnl2 using the *NcoI* and *NdeI* restriction sites. Plasmid maps were created using SnapGene software.

5.2.2 Protein Characterisation

The plasmids were transformed into *E. coli* BL21 (DE3) cells for protein synthesis (see Section 2.2.6). The new proteins, with T4Pnk at the N- terminus, were induced and synthesised in 1 L of selective LB media (see Sections 2.3.1, 2.3.2 and 2.3.3). Western blot analysis was carried out (see Section 2.3.7) to confirm the 10x His tag was still present for purification (Figure 5.11). The soluble fraction was purified using a Ni-Hi Trap column (see Section 2.3.4) and fractions 4-6 pooled and the salt removed using a PD10 column (see Section 2.3.5). The purification fractions were analysed using 10 % SDS-PAGE (Figures 5.9 and 5.10) (see Section 2.3.6). The amount of purified T4Pnk-ligase protein was quantified using the Bradford Assay (see Section 2.3.8).

5.2.3. Protein Analysis and activity

A number of ligation assays were set up (see Sections 5.2.3.1 and 5.2.3.2) and analysed using urea-PAGE (see Section 2.4.3) and ImageJ software.

5.2.3.1 Endpoint Ligation Analysis

The end point ligation assays were carried out as stated in chapter 2 (see Section 2.4.1). A second set of substrates were made using modified versions of the DNA and RNA oligonucleotides (Table 5.2). The Pnk versions of LigSub1 and LigSub2 contained a phosphate at the 3' end of LigSub1 and no modification to LigSub2. (Tables 5.2 and 5.3, Figure 5.9) This allowed activity of the Pnk part of the fusion to be tested because, for the ligase to form the phosphodiester bond, the phosphate needs to be in the 5' position. These oligonucleotides were combined as stated in Chapter 2 (see Section 2.4, Figures 2.4 and 5.9) to make 8 different 20-mer/40-mer substrates that were a combination of DNA and RNA (Table 5.3). The large oligonucleotides (18b and 22b) were used to make longer versions of substrate 1 and 7 only.

Table 5.2. The oligonucleotides used to create the Pnk substrates. The 20-mer/40mer backbone was the same as the one used for the standard 20 bend point ligation substrates (Table 2.3). Where oligonucleotides include '+Flu' – addition of fluorescein or '+P' – addition of phosphate, denotes a specific modification at the 5' position. The 18 b and 22 b oligonucleotides were used to make a 40 b substrates.

Oligonucleotide Name	Backbone	Sequence 5' – 3'	Modification
PnkLigSub1 (8b)	DNA	GGC-CAG-TG	Flu at 5' end Phosphate at 3' end
PnkLigSub2 (12b)	DNA	AAT-TCG-AGC-TCG	-
PnkLigSub1 (18b)	DNA	GTA-AAA-CGA-CGG-CCA-GTG	Flu at 5' end Phosphate at 3' end
PnkLigSub2 (22b)	DNA	AAT-TCG-AGC-TCG-GTA-CCC-GGG-G	-
PnkRNALigSub1 (8b)	RNA	GGC-CAG-UG	Flu at 5' end Phosphate at 3' end
PnkRNALigSub2 (12b)	RNA	AAU-UCG-AGC-UCG	-
PnkRNALigSub1 (18b)	RNA	GUA-AAA-CGA-CGG-GCA-GUG	Flu at 5' end Phosphate at 3' end
PnkRNALigSub2 (22b)	RNA	AAU-UCG-AGC-UCG-GUA-CCC-GGG-G	-

Table 5.3. The eight 20 b Pnk substrates created using the Pnk oligonucleotides. The three components for each substrate are PnkLigSub1 (blue, Table 5.2), PnkLigSub2 (red, Table 5.2) and the LigSub1+2 (purple) complementary strand from Table 2.3. The phosphate (P) is at the 3' end of PnkLigSub1 (Figure 5.9) for transfer to the 5' end of PnkLigSub2 before ligation can occur.

Name	Substrate (filled = DNA, hatched = RNA)
Pnk Substrate 1	5' P 3' / 3' 5'
Pnk Substrate 2	5' P 3' / 3' 5'
Pnk Substrate 3	5' P 3' / 3' 5'
Pnk Substrate 4	5' P 3' / 3' 5'
Pnk Substrate 5	5' P3' / 3' 5'
Pnk Substrate 6	5' P 3' / 3' 5'
Pnk Substrate 7	5' P 3' / 3' 5'
Pnk Substrate 8	5' P 3' / 3' 5'

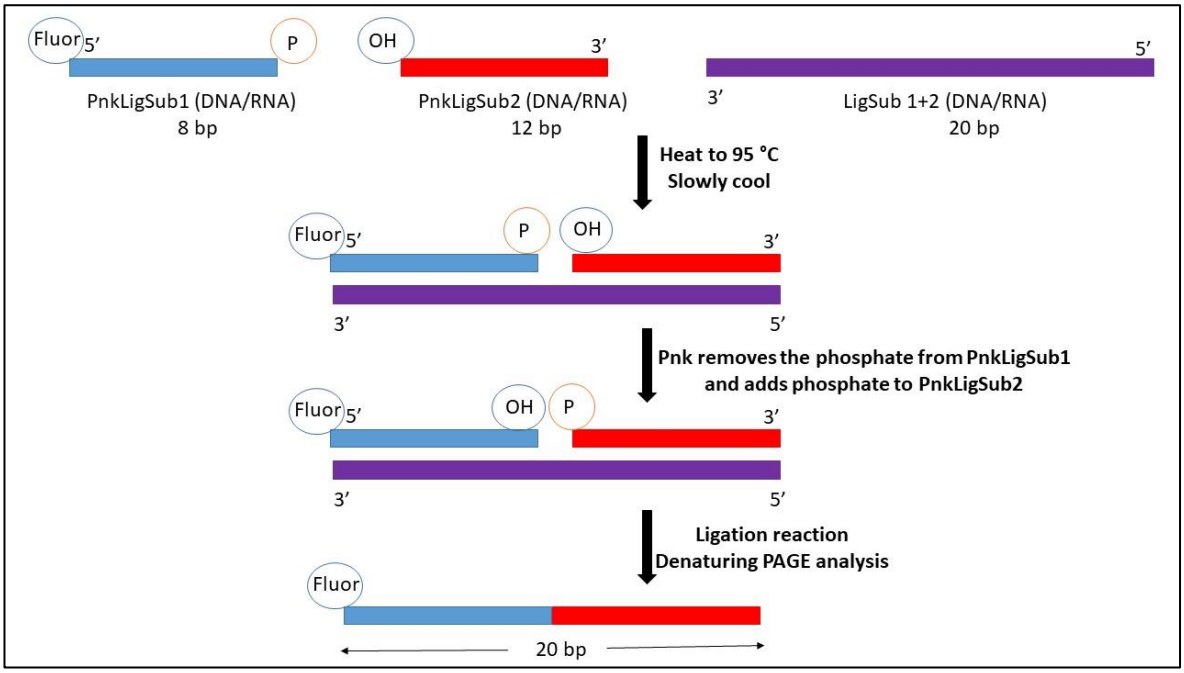


Figure 5.9. Pnk Substrate ligation. When forming the Pnk substrates, the 3 oligonucleotides are heated to 95 °C for 5 minutes and allowed to cool overnight while covered from light. The Pnk substrates used in ligation reactions undergo modification from the Pnk part of the fusion before ligation with the T4 ligases in the fusion.

5.2.3.2 Time Course Ligation Analysis

The end point analysis of different versions of substrate 7 with the Pnk-ligases (Figure 5.14) showed that only PnkRnl2 worked efficiently for ligation when there was an abundance of ligase to each substrate. The ligation rate for PnkRnl2 was determined using time course ligation analysis (see Section 2.4.2). Two versions of substrate 7, the standard substrate 7 (Figure 2.4) and Pnk substrate 7 (Table 5.3, Figure 5.9), were used to determine the initial rate of PnkRnl2. Two types of substrate 7 were required to test each enzyme, T4Pnk and the ligase.

5.2.4. Phosphate release assay

The 3' phosphatase part of T4Pnk should result in release of phosphate (which may be released free in solution or transferred to the 5'OH (Caldecott, 2002; Galburt *et al.*, 2002)). While the 5' kinase part attaches phosphate, which may be provided by ATP or directly transferred from the 3' phosphate, to the 5' OH. Detection of free phosphate was measured when carrying out ligation using Pnk substrate 7 (20-mer and 40-mer) with ligation buffer containing 1 mM ATP. The phosphate release assay was set up with the reaction conditions to be the same as the end point ligation assay (see Section 2.4.1) and nuclease free water added to a final volume of 50 µl in a PCR plate (Figure 5.10), sealed and incubated at 37 ^{o}C for one hour. A standard curve was also constructed by using 10 M KH_2PO_4 to prepare samples at the following concentrations (wells 4 and 5 A – H): 1 mM, 500 µM, 250 µM, 125 µM, 62.5 µM, 31.3 µM, 15.7 µM and 7.85 µM in a total volume of 50 µl. After one hour 50 µl of Phosphate release assay developing solution (4 parts Ammonium molybdate-sulphuric acid solution: 1 part Ferrous sulphate solution, Table 2.5) was added and incubated at room temperature for 10 minutes. The whole reaction mixture was then transferred to a round bottom plate microplate (ThermoFisher). The absorbance was measured at 700 nm using the Hidex Sense Microplate reader. From the recorded absorbance values it was possible to calculate the amount of free phosphate in solution for each of the reactions, and to correct for the absorbance value of the buffer, and buffer + substrate.

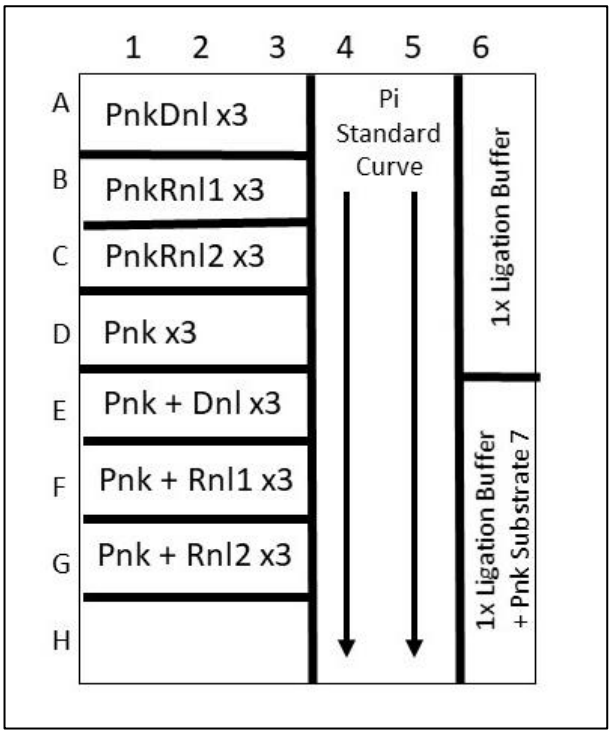


Figure 5.10. The phosphate release plate assay. Performed with each end point ligation repeated in triplicate with both 20-mer and 40-mer Pnk substrate 7.

5.3 Characterisation of the T4Pnk-ligases

5.3.1. Confirmation of the T4Pnk-ligase constructs

To confirm the *pseT* gene had been incorporated into each of the recombinant plasmids, the new plasmids were subjected to double digestion with *NcoI* and *BamHI*, the digested plasmids were analysed by agarose gel electrophoresis (Figure 5.11) and were also sent for Sanger sequencing at Eurofins. The fragment sizes and sequencing results were as expected.

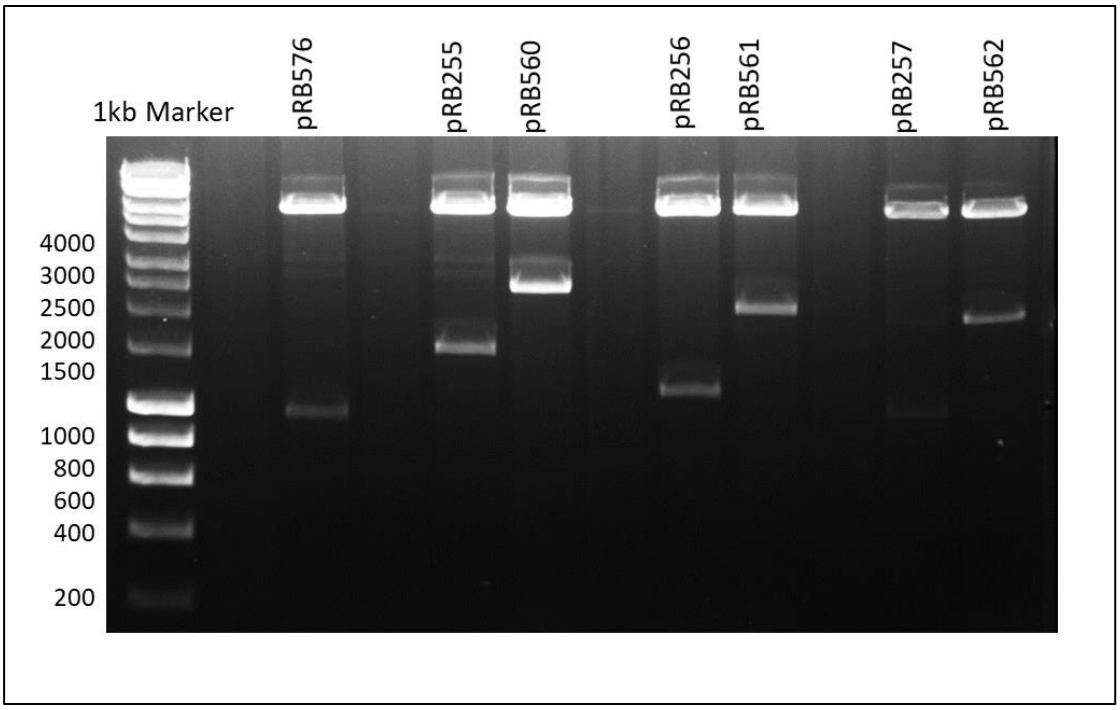


Figure 5.11. Analysis of the new plasmids pRB560, pRB561 and pRB562. The *pseT*-ligase inserts digested with *NcoI* and *BamHI*, compared to the T4 nucleic acid ligases T4Dnl (pRB255), T4Rnl1 (pRB256) and T4Rnl2 (pRB257) to confirm the expected sizes of the cloned genes. The expected sizes were: T4Dnl gene (pRB255) 1473 bp PnkDnl gene (pRB560) 2444 bp, T4Rnl1 gene (pRB256) 1134 bp, PnkRnl1 gene (pRB561) 2105 bp, T4Rnl2 gene (pRB257) 1014 bp, PnkRnl2 gene (pRB562) 1985 bp and the Pnk gene (pRB576) 971 bp. Analysed using a 1 % TAE agarose gel with 5µl 10 mg/ml ethidium bromide (see section 2.24).

5.3.2. Protein purification

The T4Pnk-ligases retained the His-tag for protein purification and were all soluble once purified. Western blot analysis did not detect the extra bands present during purification (Figure 5.12) indicating these proteins/degradation products did not contain the N-terminal His-tag. A clear increase in size for each of the T4Pnk-ligases was observed and the expected sizes were: PnkDnl – 89.9 kDa, PnkRnl1 – 78.1 kDa and PnkRnl2 – 72.2 kDa (Figure 5.13). The final concentrations for the purified proteins were: PnkDnl 177 µg/ml, PnkRnl1 784 µg/ml and PnkRnl2 1.38 mg/ml.

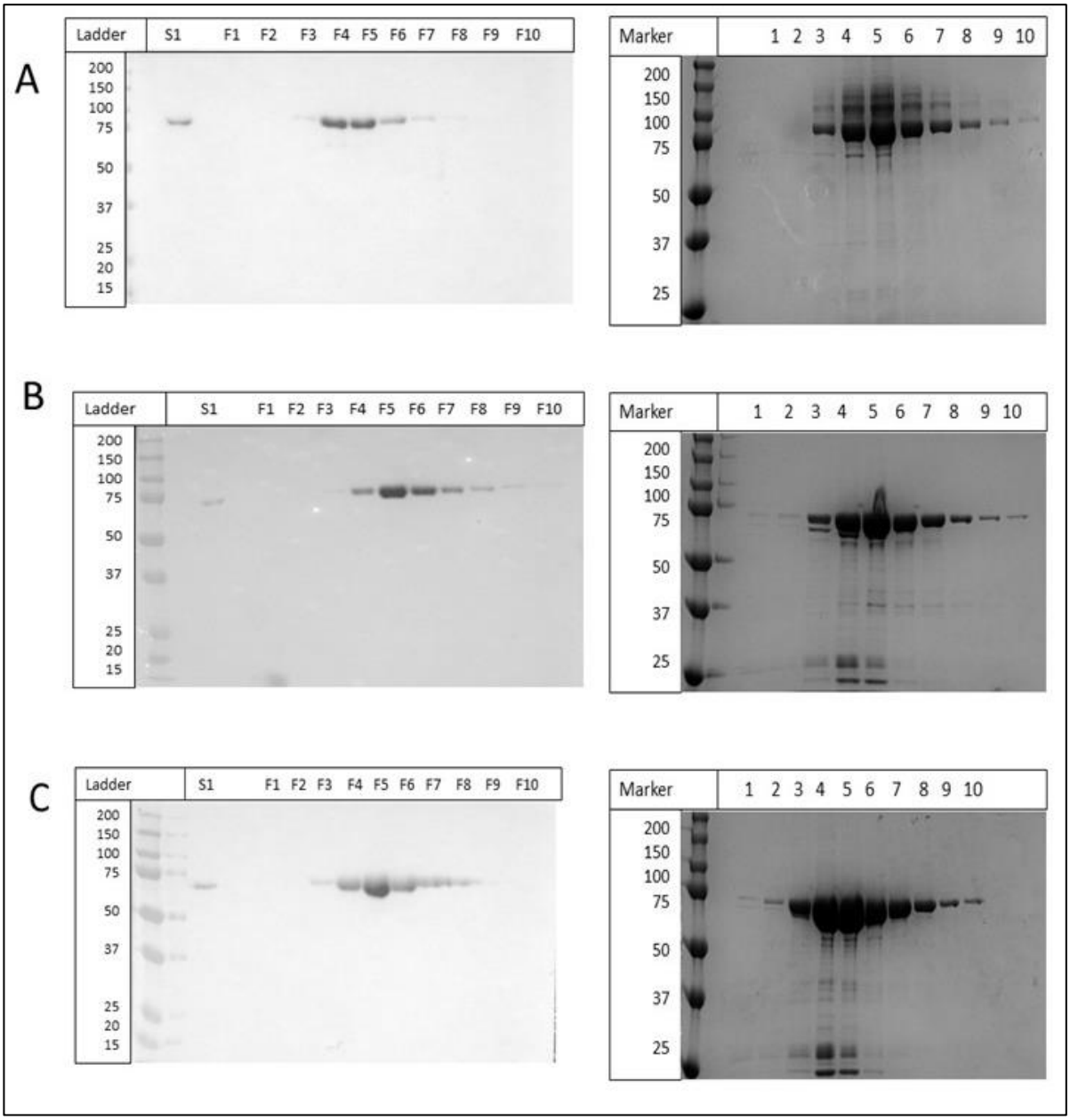


Figure 5.12. Western Blot and SDS-PAGE confirmation of the Pnk-ligases. Left hand side: Western Blot analysis using 2 µl of the monoclonal anti-poly histidine peroxidase antibody (from mouse) of the purification fractions. Right hand side: 10 % SDS-PAGE analysis of the purification elution fractions S1 and F1-F10 (See section 2.3.4) for PnkDnl 89.9 kDa (A), PnkRnl1 78.1 kDa (B) and PnkRnl2 71.6 kDa (C).

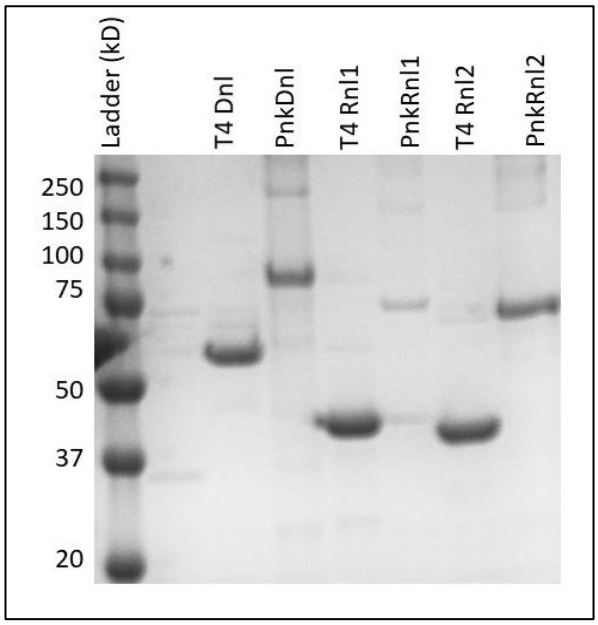


Figure 5.13. Purified T4 and Pnk-ligases. SDS-PAGE analysis (10 % polyacrylamide) highlights the change in size of each T4 nucleic acid ligase and the T4Pnk-ligases. The expected sizes were: T4Dnl 55.3 kDa, PnkDnl 92.5 kDa, T4Rnl1 43.5 kDa, PnkRnl1 80.7 kDa, T4Rnl2 37.6 kDa and PnkRnl2 74.8 kDa.

5.3.3. Characterisation of the T4Pnk-ligases

A number of different substrates were used to check ligation activity of the T4Pnk-ligases. This included the standard substrate set for end point ligation (Figure 5.14, Table 2.4) and Pnk variations of these substrates using modified oligonucleotides (Table 5.2, Figure 5.9). For the time course reactions to determine initial rates, different versions of substrate 7 were used to determine which substrate the T4Pnk-ligases could ligate best (Figure 5.15), including a Pnk version of substrate 7 and two longer 40-mer substrates for the standard and Pnk versions of substrate 7.

5.3.3.1. End point ligation

The end point ligation assay was carried out with all new T4Pnk-ligases. PnkDnl showed the same ligation profile as T4Dnl (Figure 4.10), but some of the substrates (3, 5 and 6) had lower total ligation percentage (Figure 5.15). PnkRnl1 demonstrated the poorest end point ligation, with severely reduced ligation for the substrates that T4Rnl1 is usually able to ligate. Of the new T4Pnk-ligase fusions, the end point ligation analysis of PnkRnl2 confirms the same profile as T4Rnl2.

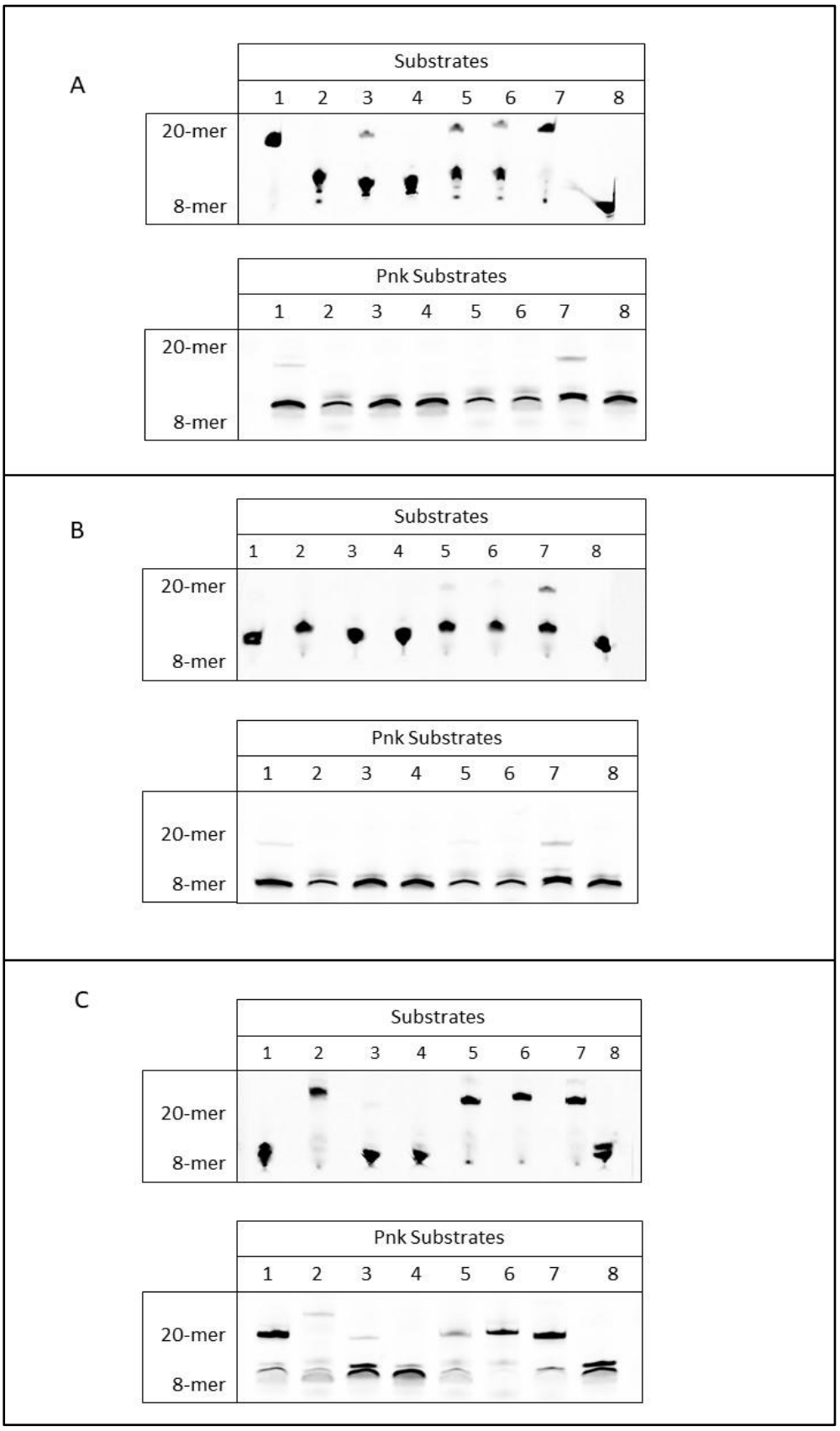


Figure 5.14. End point ligation analysis of the Pnk-ligases with different substrates. Reactions were set up with both the standard 20 b substrates 1-8 and Pnk 20 b substrates 1-8 for one hour at 37 °C and were analysed using 15 % urea polyacrylamide gels for A) PnkDnl, B) PnkRnl1 and C) PnkRnl2.

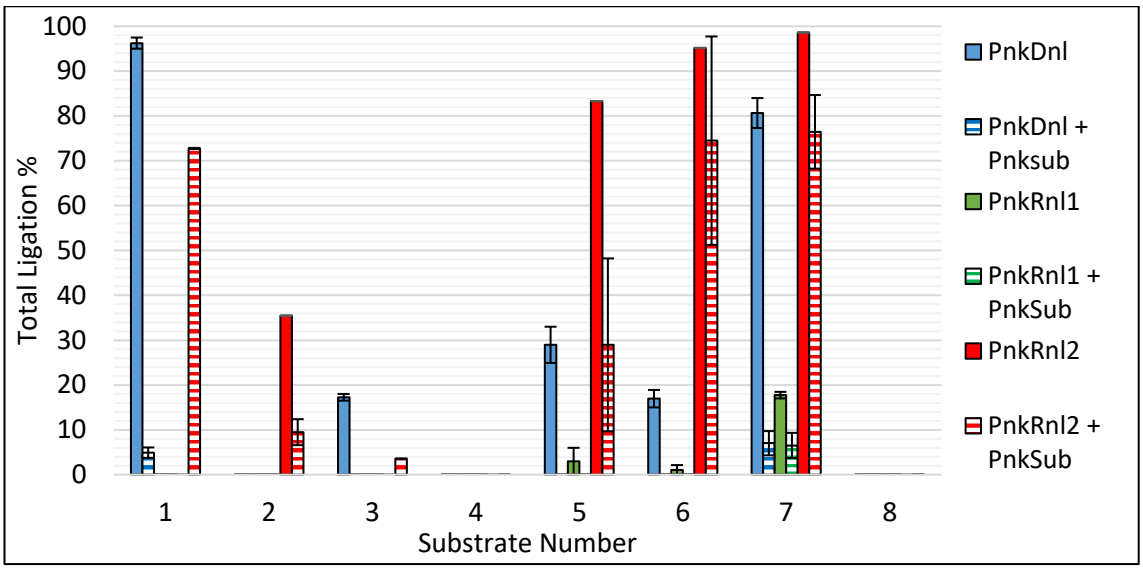


Figure 5.15. Quantification of the end point ligation analysis for the Pnk-ligases and different 20 base substrates. Each Pnk-ligase was tested with the standard 20 b substrates 1-8: PnkDnl (blue), PnkRnl1 (green) and PnkRnl2 (red). The same reaction was also set up with the Pnk substrates 1-8: PnkDnl + PnkSub (blue striped), PnkRnl1 + PnkSub (green striped) and PnkRnl2 + PnkSub (red striped). The ligation reactions were set up in triplicate, as per section 2.4.1 with 45 pmol of substrate and 70 pmol ligase. The gel images from Figure 5.13 were quantified using ImageJ software and the error bars indicate standard deviation.

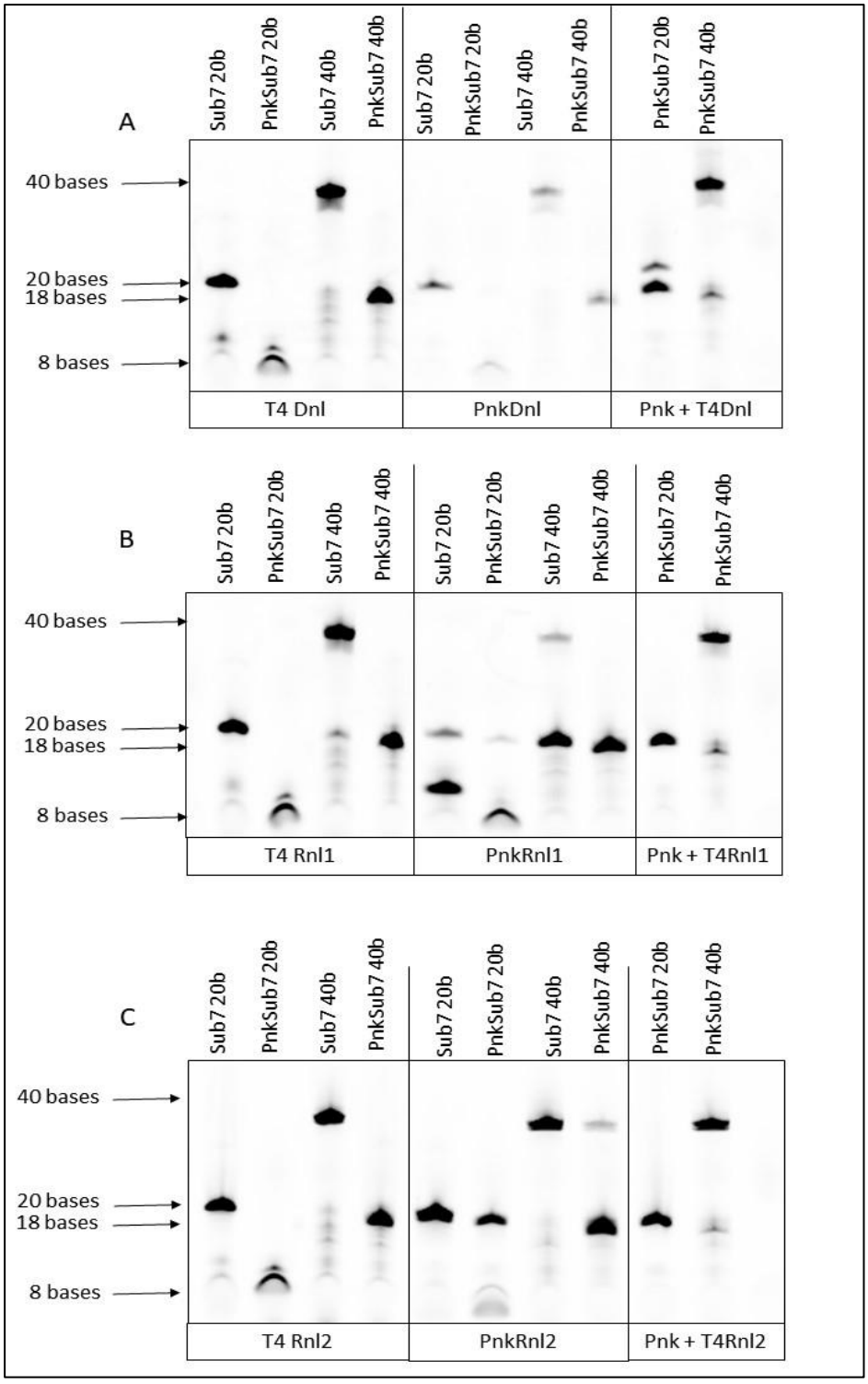


Figure 5.16. Further ligation analysis using 20 base and 40 base variants of substrate 7 and Pnk Substrate7. The T4 ligases and the Pnk-ligases underwent end point ligation analysis with the standard ligation buffer with 1 mM ATP at 37 ^{o}C for one hour (see Section 2.4.1). PnkDnl (A) and PnkRnl1 (B) could ligate a 40 base substrate 7 but not the Pnk version of this substrate (middle panels on A and B respectively). PnkRnl2 (C) could successfully ligate both the standard and Pnk 40 base substrates, the standard 20 b and Pnk 20 b substrate 7 (middle panel). The T4 ligases (left hand panels) and non-fused ligases: T4Dnl, T4Rnl1 and T4Rnl2 + Pnk (right hand panels) were controls for ligation of the standard and Pnk 20 base and 40 base substrates.

The controls, where the Pnk and ligase were not fused, demonstrate the desired ligation capacity for the T4Pnk-ligases (Figure 5.16, last section on each gel). The T4 ligases were also controls to demonstrate that the T4 ligases cannot ligate the Pnk substrates, where there is mis-priming on the DNA/RNA strands, but can ligate the standard 20 b version of substrate 7 (Figure 5.16, first section on each gel). Both PnkDnl and PnkRnl1 had low levels of ligation with the Pnk substrates, for both 20-mer and 40-mer; PnkDnl could ligate the standard 20 b substrates while PnkRnl1 did not ligate the standard 20 b substrates fully (Figure 5.16, middle sections). PnkRnl2 ligated all 4 versions of substrate 7, it worked best with the standard 20 b substrate 7 20-mer, 40-mer and 20-mer Pnk substrate 7 (Figure 5.16, middle section on bottom gel). Since PnkRnl2 was the only T4Pnk-ligase able to ligate the Pnk substrates, it was the only T4Pnk-ligase that underwent time course ligation experiments.

5.3.3.2. Time Course Analysis

The time course ligation assay was carried out to determine the initial rates for the Pnk-ligases. The amount of enzyme used varied from 2-100 pmol and this was then used to determine the amount ligated per mol of protein (Figure 5.17).

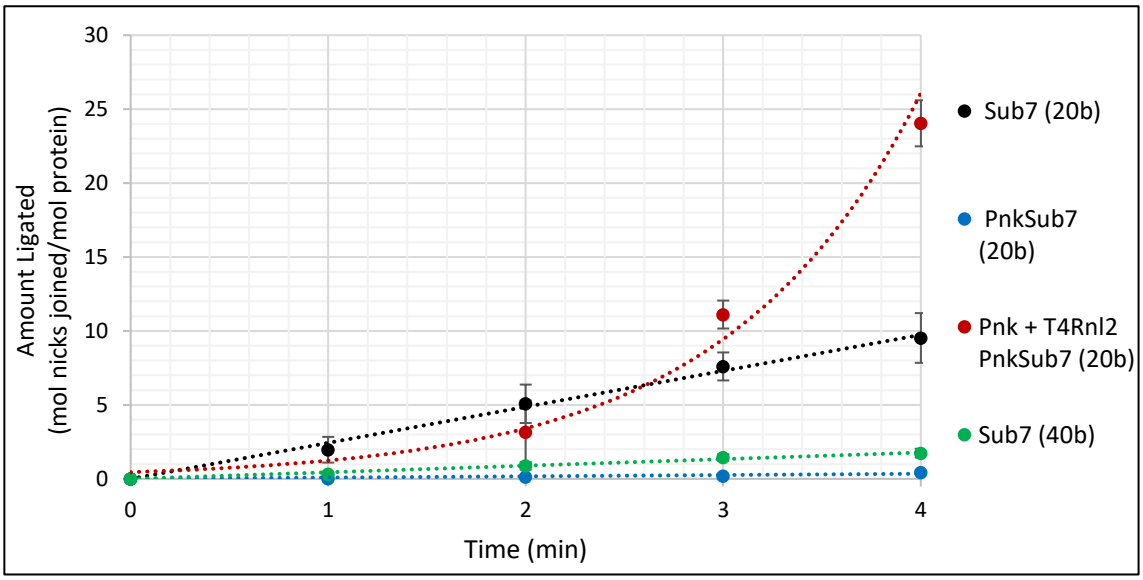


Figure 5.17. Time Course Analysis of PnkRnl2. Initial rates for PnkRnl2 were calculated with different versions of substrate 7, 540 pmol of substrate and 2-100 pmol of enzyme was used with ligation buffer containing 1 mM ATP; reactions were repeated in triplicate with the standard deviation shown as error bars. Non-fused Pnk + T4Rnl2 (red) demonstrates a curve, instead of a linear initial rate of reaction indicating a delay in modification of the DNA for ligation. Fused PnkRnl2 with standard 20 b substrate 7 (black), Pnk 20 b substrate 7 (blue) and 40 b substrate 7 (green) have a linear initial rate of reaction.

Table 5.4. Rates of nick joining for PnkRnl2 and Pnk + T4Rnl2. PnkRnl2 time course analysis with 20 b and 40 b standard substrate 7, 20 b PnkSub7 and non-fused Pnk + T4Rnl2 with standard 20 b substrate 7.

Substrate	Rate of ligation at 37 °C (mol nicks joined/mol protein/min)
Sub 7 (20b)	2.44 (+/- 0.32)
PnkSub7 (20b)	0.09 (+/- 0.02)
Sub7 (40b)	0.45 (+/- 0.08)
Pnk + T4Rnl2 PnkSub7 (20b)	4.53 (+-/ 0.26)

The initial rates for PnkRnl2 with the substrate 7 variants indicated that Pnk and T4Rnl2 separately has a better initial ligation rate than PnkRnl2 with Pnk substrate 7 (20-mer) (Figure 5.17, Table 5.4). The published rate for T4Rnl2 with substrate 7 is 46 (+/- 2.5) (Bullard and Bowater, 2006). T4Rnl2 alone will not ligate the Pnk version of this substrate (Figure 5.16) and this result is the best comparison for the rate of T4Rnl2. PnkRnl2 was able to ligate standard substrate 7 indicating the ligase part was functional, while Pnk

substrate 7 (20-mer) and the longer standard 40-mer version of substrate 7 showed the slowest rates of ligation.

5.3.3.3 Analysis of free phosphate in solution

The incomplete ligation of substrates in both the end point ligation and time course assays indicated that part of the Pnk-ligase fusion was inhibited or not functional, especially in the case of PnkDnl. A phosphate release assay was carried out to determine levels of free phosphate in solution (Figure 5.18). The phosphate release assay uses a developing solution that binds to the free phosphate to produce a colour change, which can be detected at 700 nm. Using the standard curve, it was possible to calculate the amount of free phosphate in solution for each of the reactions (Figure 5.19) and correct for the absorbance of buffer and buffer containing substrate.

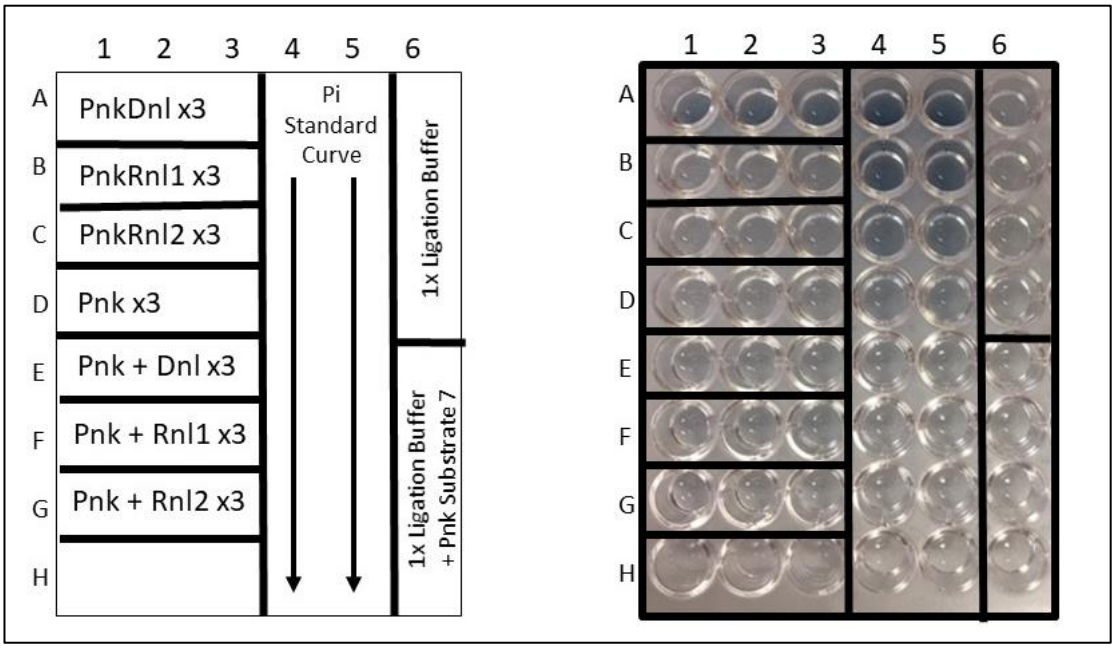


Figure 5.18. The schematic and plate picture of the phosphate release assay. In each reaction 70 pmol of ligase was used with 45 pmol of the 20-mer Pnk substrate 7 with standard ligation buffer containing 1 mM ATP, incubated for one hour at 37°C before the addition of the phosphate developing solution.

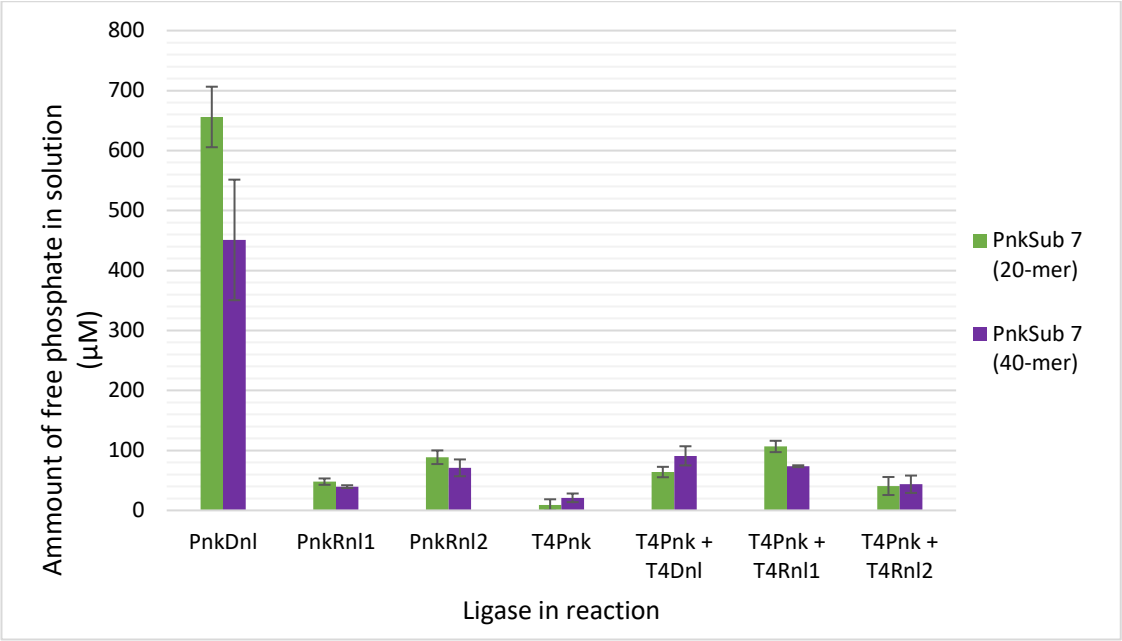


Figure 5.19. The amount of free phosphate in solution from the phosphate release assay. Average free phosphate in solution values with corrections for ligation buffer and substrate phosphate levels, each end point ligation was carried out in triplicate and error bars denote the standard deviation.

Visualisation of the plates with both Pnk versions of substrate 7 (20-mer and 40-mer) show high levels of phosphate release with PnkDnl. All of the reactions release some phosphate into solution but this is 70-fold higher with PnkDnl compared to T4Pnk. T4Pnk on its own is able to attach and remove the phosphate with no hindrance, releasing a small amount into solution (Figure 5.19).

5.4 Discussion

Addition of enzymes at the N- terminus of the T4 nucleic acid ligases was previously successful in producing fusion ligases with added function (see Chapter 4) (Liu *et al.*, 2015). For the experiments described in this chapter though, more work needs to be done to assess the fusion site for Pnk and the T4 ligases. The 3 Pnk-ligases had varied results, PnkDnl and PnkRnl1 had reduced ligation with the standard 20 b and Pnk 20 b substrates 1-8 (Figures 5.14, 5.15 and 5.16). PnkRnl2 effectively ligated a number of the Pnk and standard 20 b substrates (Figures 5.14, 5.15 and 5.16). It is possible that addition of the ligase at the C- terminus of Pnk has inhibited access to one of the active sites,

which in turn produces less substrate that the ligases can ligate. The exception to this is PnkRnl2 and it is likely that, as the smallest of the T4 nucleic acid ligases (37kDa), it is causing less steric hindrance for Pnk. Steric hindrance of T4Pnk has been noted before when using pyrene attached to double stranded DNA (Song and Zhao, 2009).

The end point ligation with the Pnk 20 b substrates (Figures 5.14, 5.15 and 5.16) showed that for PnkDnl and PnkRnl1 nearly all ligation capacity is diminished with the Pnk substrates. Ligation activity with the standard 20 b substrates was comparative to the ligation activity of the T4 ligases (Figure 3.15) indicating the lack of ligation with the Pnk substrates is caused by a problem with the Pnk part of the fusion. PnkDnl and PnkRnl1 ligated the standard 20 b substrates but both had a reduced profile when compared to T4Dnl and T4Rnl1 (Figure 3.15). PnkDnl loses the ability to ligate standard 20 b substrate 2 (double stranded RNA) (Figure 5.14). PnkRnl1 can no longer ligate standard 20 b substrates 2, 5 and 6 and showed lower ligation of standard 20 b substrate 7 when compared to T4Rnl1 (Figures 3.15 and 5.14). The ligase parts of the fusions were able to successfully ligate the standard substrates 1-8 and the profile produced matched that of the non-fused T4 nucleic acid ligases (Figures 4.9 and 5.14). The Pnk Substrate 7 (40-mer) has lowered levels of ligation with all 3 Pnk-ligase fusions, yet these ligases are all able to successfully join the nick for the standard version of substrate 7 (40-mer). Therefore the length of the substrate does not hinder the ligase but likely affects the Pnk part of the fusion (Figure 5.16). T4Pnk normally forms a homotetramer of four subunits (Galburt *et al.*, 2002), a tunnel conformation that allows T4Pnk to remove and attach phosphate at the same time (Caldecott, 2002; Galburt *et al.*, 2002). Further end point analysis of the Pnk-ligases and the T4 ligases (Figure 5.16) with different versions of substrate 7 was carried out to determine which substrates would be suitable for time course ligation analysis. PnkDnl and PnkRnl1 did not ligate the Pnk substrates fully (Figure 5.16) and did not complement the activity seen with the enzymes when not fused. PnkRnl2 did ligate Pnk substrate 7 (20-mer) and a small amount of Pnk substrate 7 (40-mer) (Figure 5.16), and as such only PnkRn2 was then analysed to determine the initial rate.

5.4.1 Phosphate Release Analysis

The phosphate release assay demonstrated that the fusion of T4Pnk and T4Dnl resulted in phosphate removal from the substrate, and a functional C- terminus phosphatase of T4Pnk, but showed that phosphate was not added to the 5' position on the oligonucleotide, indicating the N- terminus kinase activity was inhibited or non-functional. This resulted in high amounts of free phosphate in solution (Figure 5.19) and ultimately affected the ability of the ligase to create the phosphodiester bond for the Pnk substrates (Figure 5.16). The non-bound Pnk and T4 ligases also released phosphate into solution (Figure 5.16), these offer a baseline for the different substrates. When considering enzymatic reactions, substrate (A) is catalysed into a product (B). In this instance there are two reactions occurring, one where the Pnk removes phosphate from the 3' end of the break in the substrate and attaches phosphate to the 5' part of the break (reaction A → B) and then a second reaction where the ligase seals the nick (i.e. product B is converted into a further product, C in reaction B → C). As A is used up, B is created and then used up to make C. A delay in the creation of B (modified DNA/RNA ends) results in a delay in the creation of C (ligated substrate) which may give the result seen in Figure 5.17 – where the rate of reaction isn't linear. The ligase can only ligate the substrate once the phosphate has been removed from the 3' part of the break and attached to the 5' part. There is no way to know whether Pnk is only utilising the phosphate on the substrate or also using the ATP in the buffer and whether the phosphate in solution is from one or both of these sources.

5.4.2 Analysis of PnkRnl2

PnkRnl2 appears to function the most efficiently of the 3 T4Pnk-ligase fusions. The end point ligation profile remained the same as T4Rnl2 with the highest amount of ligation of the 3 T4Pnk-ligases with both versions of substrate 7 (20-mer). The ligation of PnkSub7 (40-mer) is limited by PnkRnl2, the length of the substrate has a very clear impact on ligation capacity of PnkRnl2. This is likely because T4Pnk feeds the DNA/RNA through a positively charged tunnel that runs through the middle of the enzyme, ATP is bound at one end and the DNA/RNA at the other just 3.3 A˚ apart, the phosphate is then transferred from the ATP to the 5'-OH (Caldecott, 2002; Galburt *et al.*, 2002). It is possible to theorise that this mechanism would also work and the substrates with

phosphate at the 3' position passes this to the 5' end as the DNA/RNA is already held in place in the tunnel active site.

The non-fused T4Pnk and T4Rnl2 demonstrated the best initial rate of ligation (Figure 5.18), however the end point ligation after one hour demonstrated that both the fused and non-fused proteins ligate 20-mer Pnk substrate 7 fully, when using 70 pmol of ligase and 45 pmol of substrate (Figure 5.16). The time course analysis indicated slower initial rates of ligation with Pnk substrate 7 (20-mer) and standard substrate 7 (40-mer) (Table 5.3). This is possibly due to the length of the 40-mer substrate and for ligation of the Pnk substrate 7 (20-mer) both the ligase and T4Pnk parts are required to prime and then repair which may take longer than solely ligating the break. It is possible to theorise that, as the smallest of the 3 ligases, the addition of T4 Rnl2 at the 3' end of T4Pnk still allows mobility in the second active site for removal of the phosphate from the 3' position. When Pnk and T4Rnl2 were not fused, both the 20-mer and 40-mer Pnk substrates were ligated fully; likely the Pnk and ligase could catalyse their respective reactions without steric hindrance (Figures 5.16, 5.17 and 5.19). The rate of ligation when Pnk and T4Rnl2 were unbound (Table 5.4) is double the rate of the bound PnkRnl2 with standard 20 b substrate 7: 4.53 mol nicks joined/mol protein/min (+-/ 0.26) and 2.44 mol nicks joined/mol protein/min (+/- 0.32) respectively. The published value for T4Rnl2 activity with this substrate is 46 mol nicks joined/mol protein/min (+/- 2.5) (Bullard and Bowater, 2006); the bound and unbound versions of Pnk and T4Rnl2 have a much lower rate than this because the ligation is delayed until the DNA/RNA substrate has the phosphate removed and reattached by Pnk.

5.4.3. PnkDnl and PnkRnl1

With both PnkDnl and PnkRnl1 there was reduced end point ligation with the standard 20 b and Pnk 20 b substrates (Figures 5.13, 5.14 and 5.15), when compared to the T4 ligases end point ligation (Figure 4.10). PnkDnl and PnkRnl1 retain some ligation activity (Figure 5.15 and 5.16), but as both these ligases did not ligate the substrates fully it was not possible to obtain rates for these reactions. Both fusions with T4Dnl and T4Rnl1 appear to limit the T4Pnk activity, especially with PnkDnl as there is a lot of free

phosphate in the reactions. The tunnel active site in T4Pnk usually localises ATP to the DNA/RNA end requiring repair (Caldecott, 2002; Galburt *et al.*, 2002) and in theory this system could instead transfer the phosphate between the mis-primed strands, as these are already localised by their complementary strand. For PnkDnl though the active site that removes phosphate is functional releasing high levels of phosphate into solution (Table 5.4) but the part that attaches the phosphate at the 5' position was less functional than T4Pnk with the ligases unbound (Figure 5.16). Further optimisation of this system could be carried out, whereby the ligases are joined using a linker between the 2 active sites of T4Pnk, although linker length and placement would need to be optimised.

5.5 Conclusion

The addition of Pnk to the T4 nucleic acid ligases did not totally demonstrate the aimed for result of an efficient fusion. Of the 3 fusions, PnkRnl2 functioned the best and would likely be of interest for industrial application and further study. The phosphate release assay demonstrates that the fusion of T4Pnk and T4Dnl had the capacity to remove phosphate from the Pnk substrates but has lost the capacity to attach phosphate. While the levels of free phosphate in solution for PnkRnl1 were not as high as those for PnkDnl, PnkRnl1 also demonstrated the lowest levels of ligation from end point analysis (Figures 5.13 and 5.14).

Further analysis of PnkRnl2 could be carried out, especially analysis of NGS and adapter ligation to miRNAs. It is possible to theorise that miRNA identification levels may differ with the addition of T4Pnk to correctly prime ends for ligation. This is a new field of research and little work has been conducted to determine whether levels in miRNAs attributed to disease are conclusively lower because of low production levels or whether the levels are lower due to incorrectly primed miRNAs that would not be able to be ligated to the adapters. Theoretically this would only happen if mis-priming resulted in a mixed population of miRNAs – incorrectly and correctly primed. It would be interesting to test a panel of miRNAs under normal adapter-ligation conditions and also with PnkRnl2 in place of T4Rnl2 to see if differences in detectable and quantifiable amounts of miRNA are observed.

Another experiment that could be carried out with the PnkRnl2 fusion protein, would be to use labelled phosphate to show whether incorporation and removal of phosphate is from the ATP or from the 3' end of the oligonucleotide. Another way to check phosphate incorporation would be to fuse Pnk to an NAD dependent ligase, eliminating the need for ATP in the buffer, and repeating the experiment to check whether the phosphate is moved from the 3' part of the break to the 5' part and then ligated. Other factors that could be optimised further include the buffer and its constituents. Salts present in buffers can have an effect on T4Pnk activity (Lillehaug and Kleppe, 1975) and the buffer used for the experiments detailed above was optimised for ligation. The addition of polymers to buffer also reverses activity loss seen with T4Pnk, notably the addition of PEG which acts via molecular crowding, and may help to improve ligation reactions for the T4Pnk-ligases with the Pnk substrates (Harrison and Zimmerman, 1986). Using linkers to join Pnk to the T4 ligases is also an option, but length, linker type and linker location would all need optimisation and the time constraints of this project did not allow for work with linkers.

The activity of T4Pnk was quantified using modified oligonucleotides and the phosphate release assay, which are indicative of T4Pnk activity but are not directly proportional to T4Pnk rate. It would be useful to consider other ways to robustly test T4Pnk activity. T4Pnk has been used as a model for a coupled λ exonuclease reaction that uses a graphene oxide based sensing platform and showed that it could detect T4Pnk activity and inhibition (Lin *et al.*, 2011). Another method for monitoring T4Pnk activity used a labelled hairpin probe coupled with λ exonuclease cleavage; when fluorescence was quenched, and upon phosphorylation by T4Pnk, the quencher was cleaved emitting fluorescence (Song and Zhao, 2009). More recently, detection of T4Pnk activity has been improved further using gold nanoparticles and β-cyclodextrin polymer to enhance fluorescence further (Huang *et al.*, 2013; Song *et al.*, 2015). All of these methodologies would provide a more accurate way of quantifying T4Pnk rate for the T4Pnk-ligase fusions.

In conclusion, there are numerous uses for T4Pnk-ligase fusions, but further work needs to be carried out to determine how best to join Pnk and the T4 ligases for both enzymes to remain functional. In the case of PnkRnl2, further analysis of this new enzyme needs to be implemented to define whether or not it could be used to improve current methodologies that use both T4Pnk and T4 ligases.

Chapter 6
Modifications to *Mycobacterium smegmatis* ligase LigD

Chapter 6: Modifications to *Mycobacterium smegmatis* ligase LigD

6.1.1 Bacterial Ligases

Bacterial ligases can either be NAD dependent or ATP-dependent, with all bacteria containing a replicative NAD-dependent ligase, LigA, and most bacteria containing accessory ATP-dependent ligases (Wilkinson, Day and Bowater, 2001; Pergolizzi, Wagner and Bowater, 2016; Williamson, Hjerde and Kahlke, 2016). The first bacterial ATP-dependent ligase was discovered in *Haemophilus influenenzae* (Cheng and Shuman, 1997) and since this discovery ATP-dependent ligases have been found in many different bacteria. There are several different types of bacterial ATP-dependent ligases, split into 6 sub-groups, LigB, LigC, LigD1, LigD2, LigD3 and LigE. These different classes of bacterial ATP-dependent ligases are characterised by different domain arrangements and domain conservation (Pergolizzi, Wagner and Bowater, 2016; Williamson, Hjerde and Kahlke, 2016) (Figure 6.1).

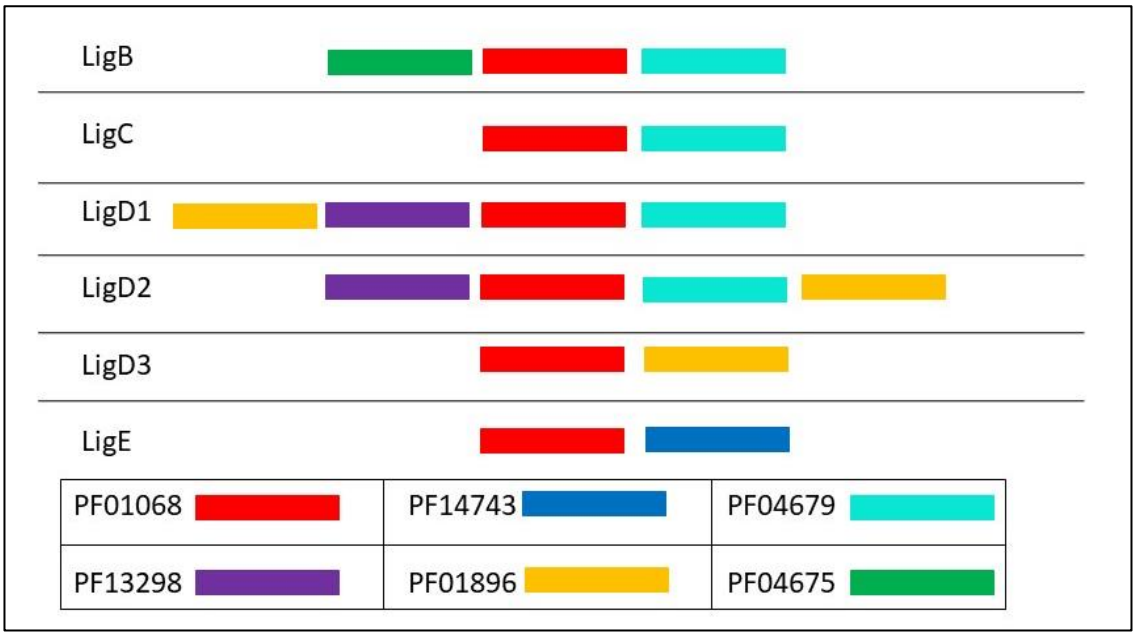


Figure 6.1. The conserved domains in bacterial ATP-dependent ligases. Each ATP-dependent ligase sub group with different conserved domains. The Pfam (PF) domain sequences from the Pfam database (Finn *et al.*, 2014) give specific information on families of proteins. Available at http://pfam.xfam.org/. The different conserved domains have different functions and include combinations of an ATP-dependent DNA ligase domain (PF01068), a DNA ligase OB domain (PF14743), an ATP-dependent ligase C- terminal region (PF04679), a DNA polymerase (PF13298), a DNA primase unit (PF01896) and a DNA ligase N- terminus found in some ATP-dependent ligases (PF04675). PF01068 contains the nucleotidyl transferase conserved polypeptide motifs I, III, IIIa, IV and V.

6.1.2. Non Homologous End Joining with LigD

One system that uses these bacterial ATP-dependent ligases is Non homologous end joining (NHEJ) (Brissett and Doherty, 2009; Matthews and Simmons, 2014). There are a number of causes for DNA damage, including spontaneous breaks during replication and exogenous damaging agents; one type of break that occurs are DNA double stranded breaks (DSBs) (Hefferin and Tomkinson, 2005). DSBs are lethal to the cell when unrepaired and may result in carcinogenesis and cell death in mammalian cells (Pitcher, Brissett and Doherty, 2007; Shuman and Glickman, 2007; Brissett and Doherty, 2009). There are two pathways to repair DSBs: homologous recombination – where a DNA template provides sequence homology to repair the DSBs, leaving no mutation to the sequence, this is considered an accurate form of repair (Shuman and Glickman, 2007); and NHEJ – where end joining occurs without sequence homology and can introduce

changes to the sequence, including addition or deletion of bases, this is considered to be error-prone repair (Hefferin and Tomkinson, 2005). Recent studies have questioned whether the error prone nature of NHEJ occurs for a specific reason and have considered a more detailed classification with highlighted advantages of adaptability of the NHEJ systems (Bétermier, Bertrand and Lopez, 2014).

NHEJ repair systems are not specific to bacteria, they were first characterised in mammals and other eukaryotes and can be found in yeasts too (Matthews and Simmons, 2014). The different NHEJ repair systems utilise different proteins to carry out the end joining process but this usually involves the Ku protein, and a specialised DNA ligase, LigD 1, 2 and 3 in bacteria (Hefferin and Tomkinson, 2005; Shuman and Glickman, 2007). The Ku protein acts as a block on the ends of DSBs, allowing repair and ligation by LigD, and when Ku is missing in *M. smegmatis* two other pathways are upregulated – homologous end joining and single strand annealing (Gupta *et al.*, 2011). *Mycobacteri*a have LigA and 3 additional ATP-dependent ligases: LigB, LigC and LigD. These were first discovered in *M. tuberculosis* where LigB, C and D were shown to be accessory ATP-dependent ligases (Cole *et al.*, 1998; Gong *et al.*, 2004; Płociński *et al.*, 2017). The frequencies of the different proteins involved in NHEJ differs from organism to organism, for example *M. tuberculosis* has single copies of Ku, LigD and LigC, while *M. smegmatis* has one Ku, one LigD and two LigC's (Shuman and Glickman, 2007). LigB was first discovered in *M. tuberculosis* and was found not to be involved in NHEJ, whereas ΔLigD *M. tuberculosis* strains demonstrated an inability to carry out NHEJ(Bhattarai, Gupta and Glickman, 2014). LigC in *M. tuberculosis* and *M. smegmatis* was involved in a NHEJ minor pathway when LigD was deleted (Gong *et al.*, 2004; Bhattarai, Gupta and Glickman, 2014). *M. smegmatis* LigD was especially of interest for this work as it is a naturally occurring, poly-functional enzyme.

6.1.3. *Mycobacterium smegmatis* LigD

LigD from *M. smegmatis* (MsLigD) is comprised of 3 domains (a DNA polymerase, phosphoesterase (PE) and a DNA ligase (Gong *et al.*, 2005)) and is 762 amino acids in length (Figure 6.2). MsLigD doesn't fully fit the Pfam bacterial ligase classification (Figure

6.1), as it is most similar to the LigD1 group containing a polymerase domain (PF13298, purple) and a ligase domain (PF01068, red). MsLigD also contains a PE domain instead of a primase unit (PF01896, yellow). Mutational analysis of MsLigD revealed that its deletion affected NHEJ. When the critical lysine residue at position 484 in motif I (Figure 6.3), was changed to alanine, DSB repair was not greatly affected and was likely supported by MsLigC (Gong *et al.*, 2005; Akey *et al.*, 2006). Removal of both MsLigB and MsLigC eradicated NHEJ in the *M. smegmatis* strain K484A (Shuman and Glickman, 2007). The upstream polymerase and PE are involved in the repair of DSB ends before the ligase seals the nick (Shuman and Glickman, 2007; Aniukwu, Glickman and Shuman, 2008; Williamson, Hjerde and Kahlke, 2016). The addition of single nucleotides at the blunt end of the DSB is catalysed by the polymerase domain (Della *et al.*, 2004; Gong *et al.*, 2005; Shuman and Glickman, 2007). The polymerase domain may also fill in ribonucleotides at a 5' overhang (Gong *et al.*, 2005; Shuman and Glickman, 2007). It has also been noted that the polymerase part of LigD in *Pseudomonas aeruginosa* shows a preference to ribonucleotides over deoxyribonucleotides *in vitro* and inserts ribonucleotides more quickly into DSBs (Yakovleva and Shuman, 2006). Further point mutations of the non-functional polymerase domain, D136A and D138A mutants, to *M. smegmatis* ΔLigD resulted in increased blunt repair of DSBs and eliminated nucleotide insertions, demonstrating a unique role for this polymerase (Zhu *et al.*, 2006; Shuman and Glickman, 2007). The PE domain is responsible for resecting ribonucleotides until the 3' strand has one ribonucleotide. This occurs via a 2-step process whereby the nucleoside is removed leaving a primer strand that has a 3'-PO_4, followed by hydrolysis of the 3'-PO_4 to 3'-OH (Zhu, Wang and Shuman, 2005; Zhu and Shuman, 2006; Shuman and Glickman, 2007). Mutations to PE in MsLigD did not affect NHEJ and it is thought the PE domain in MsLigD offers a way to repair ends that cannot be extended by the polymerase or joined by the ligase, for example DSBs with 3'-PO_4 that would not normally occur in NHEJ (Shuman and Glickman, 2007).

```
        10         20         30         40         50
MARHPWGMER YERVRLTNPD KVLYPATGTT KAEVFDYYLS IAQVMVPHIA
        60         70         80         90        100
GRPVTRKRWP NGVAEEAFFE KQLASSAPSW LERGSITHKS GTTTYPIINT
       110        120        130        140        150
REGLAWVAQQ ASLEVHVPQW RFEDGDQGPA TRIVFDLDPG EGVTMTQLCE
       160        170        180        190        200
IAHEVRALMT DLDLETYPLT SGSKGLHLYV PLAEPISSRG ASVLARRVAQ
       210        220        230        240        250
QLEQAMPKLV TATMTKSLRA GKVFLDWSQN NAAKTTIAPY SLRGRDHPTV
       260        270        280        290        300
AAPRTWDEIA DPELRHLRFD EVLDRLDEYG DLLAPLDADA PIADKLTTYR
       310        320        330        340        350
SMRDASKTPE PVPKEIPKTG NNDKFVIQEH HARRLHYDLR LERDGVLVSF
       360        370        380        390        400
AVPKNLPETT AENRLAVHTE DHPIEYLAFH GSIPKGEYGA GDMVIWDSGS
       410        420        430        440        450
YETEKFRVPE ELDNPDDSHG EIIVTLHGEK VDGRYALIQT KGKNWLAHRM
       460        470        480        490        500
KDQKNARPED FAPMLATEGS VAKYKAKQWA FEGKWDGYRV IIDADHGQLQ
       510        520        530        540        550
IRSRTGREVT GEYPQFKALA ADLAEHHVVL DGEAVALDES GVPSFGQMQN
       560        570        580        590        600
RARSTRVEFW AFDILWLDGR SLLRAKYSDR RKILEALADG GGLIVPDQLP
       610        620        630        640        650
GDGPEAMEHV RKKRFEGVVA KKWDSTYQPG RRSSSWIKDK IWNTQEVVIG
       660        670        680        690        700
GWRQGEGGRS SGIGALVLGI PGPEGLQFVG RVGTGFTEKE LSKLKDMLKP
       710        720        730        740        750
LHTDESPFNA PLPKVDARGV TFVRPELVGE VRYSERTSDG RLRQPSWRGL
       760
RPDKTPDEVV WE
```

Figure 6.2. *M. smegmatis* LigD sequence. MsLigD is made up of 3 domains which code for a DNA repair polymerase domain (purple), a 3' phosphoesterase domain (light green) and a DNA ligase domain (red). Amino acids are designated following the standard single letter abbreviations.

It therefore seems logical that these specific protein activities would be grouped together, although in some organisms the different activities are not always in one protein and, while not essential for bacterial growth, LigD (and in some instances LigC) along with the Ku protein provide some protection against DNA damage. Related NHEJ systems are found in a number of other bacteria and LigD homologs specifically can also be found in *B. subtilis,* P. *aeruginosa*, *M. smegmatis*, *M. tuberculosis* and *M. bovis* and it made sense to align these proteins (Figure 6.3) - previous alignment analysis aligned LigD from *M. smegmatis* and *M. tuberculosis* only (Gong *et al.*, 2004). The ligase part of

LigD is highly conserved and contains the nucleotidyl transferase conserved motifs I, III, IIIa, IV and V, as well as the OB domain and motif VI (Figure 6.3).

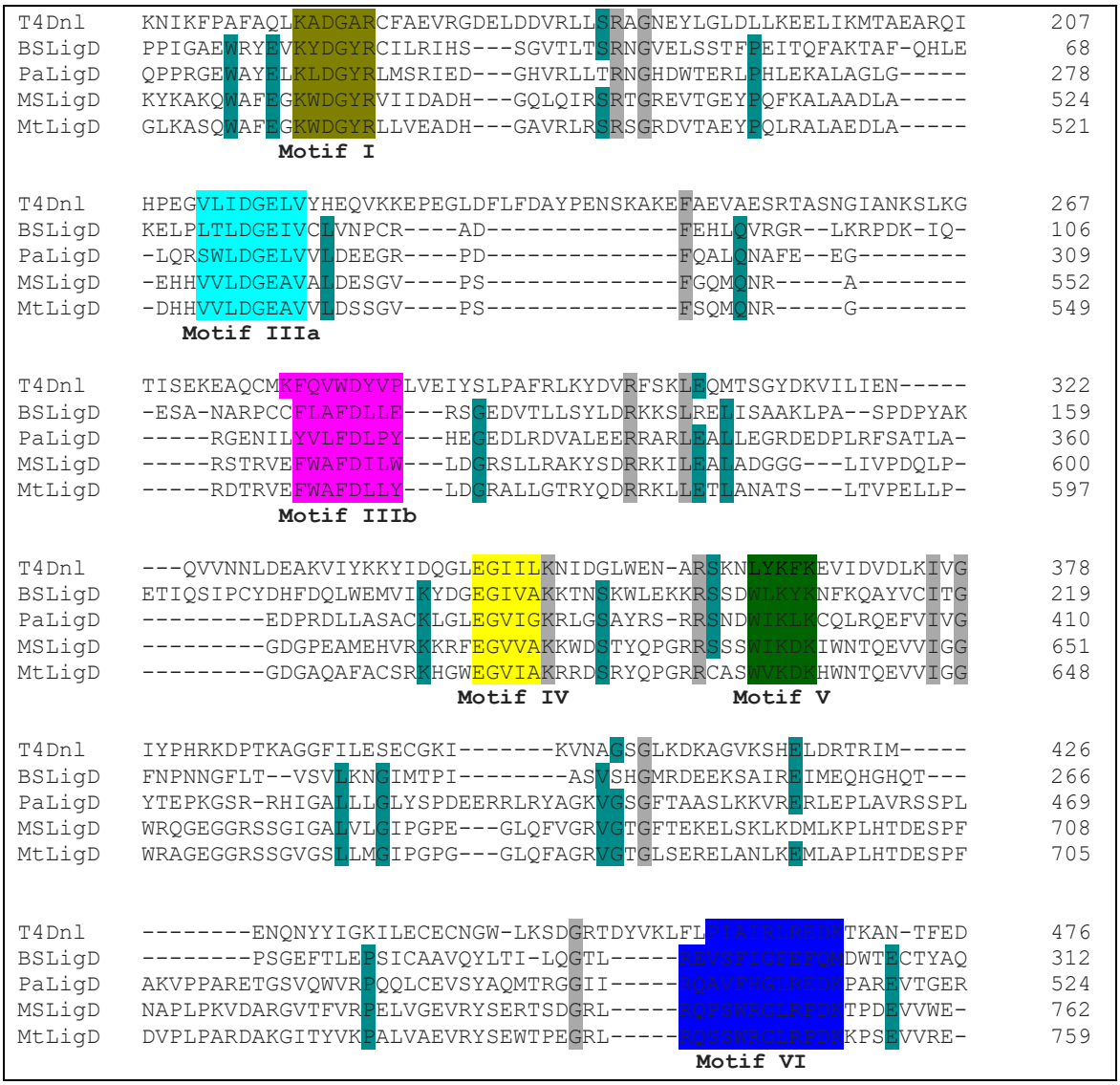


Figure 6.3. Sequence alignment for the ligase part of MsLigD to other LigD proteins and T4DnI. The ligase from LigD proteins involved in NHEJ from *B. subtilis* (BsLigD), *P. aeruginosa* (PaLigD), *M. smegmatis* (MsLigD) and *M. tuberculosis* (MtLigD); and T4DNA ligase (T4DnI) showing conserved motifs I (dark yellow), IIIa (light blue), IIIb (purple), IV (yellow), V (green) and VI (blue). The always conserved residues (grey) and highly conserved residues (teal). Alignments performed using Clustal Omega, https://www.ebi.ac.uk/Tools/msa/clustalo/ amino acids are designated following standard single letter abbreviations.

Mutagenic analysis of the polymerase domain in MsLigD demonstrated its ability to recognise 5'-phosphate on the ends of DNA without the other domains (PE and ligase) or the Ku protein (Pitcher *et al.*, 2007). The MsLigC enzyme can work with the polymerase domain from MsLigD and the Ku protein to form a NHEJ complex when the

PE and ligase domains are non-functional (Bhattarai, Gupta and Glickman, 2014). The polymerase and PE domains have a specific role in repairing DSBs caused by ionizing radiation and the polymerase domain specifically repairs damage caused by ionising radiation in the late stationary phase (Bhattarai, Gupta and Glickman, 2014). As it is possible to separate the activities for the poly-functional protein MsLigD it seemed logical that it would be possible to remove the original DNA ligase part of MsLigD and replace this with the different T4 nucleic acid ligases. MsLigD can only utilise ATP for nick sealing – not NAD and it has been noted that 1 mM of ATP causes an abundance of adenylate DNA (app-DNA) intermediates, whereby the ligase dissociates from app-DNA and reacts with ATP to form ligase-AMP which is unable to rebind to the app-DNA (Gong *et al.*, 2004). This prevents the final ligation step – phosphodiester bond formation. This phenomenon is called the ATP trapping phenomenon and causes low ligation efficiency which can be improved somewhat, but not fully, by lowering the ATP concentration (Gong et al., 2004; Williamson et al., 2016). There was the possibility to improve the efficiency of the ligase portion of this poly-functional enzyme in higher ATP concentrations by replacing the current ligase with one of the T4 nucleic acid ligases. These fusions would also act as proof of principle for fusion with an NAD dependent ligase, omitting the need for ATP in the ligation buffer. These new ligases may also be of interest as molecular biology tools offering an alternative to multiple step reactions to repair double stranded DNA/RNA breaks.

6.2 Construction of LigD with T4 nucleic acid ligases

6.2.1 Recombinant plasmid construction

The gene for MsLigD was pre-ordered and synthesised in pUC57 (Genscript, USA) with the addition of cut sites for *NcoI*, *NdeI* and *BamHI* to allow for cloning into the pET-16b plasmids (Figure 6.4). An *NcoI* site was identified at position 1769 of the gene for MsLigD, 5'-GCCATGGAG-3', coding for Ala-Met-Glu. The sequence was changed to 5'-GCAATGGAG-3', still coding for Ala-Met-Glu while removing the *NcoI* site to make cloning into the pET16b plasmids easier.

6.2.1.1 Restriction Digestion

The T4 NA ligase plasmids (pRB255, pRB256 and pRB257) and the pUC57 plasmid with synthesised MsLigD (pRB571) were digested with *NdeI* (ThermoFisher) and *NcoI* (ThermoFisher). Double digestion with *NcoI* and *BamHI* (ThermoFisher) was also set up with pRB571 to isolate the full length *msLigD* gene for insertion into the pET16b vector system. The double digestions were set up as follows: 10 µl plasmid DNA (up to 1 µg), 10 µl sterile water, 1.25 µl *NcoI*, 1.25 µl *NdeI*/*BamHI* and 2.5 µl ThermoFisher FD buffer. This was then incubated at 37 °C for 2.5 hours. To confirm the digestions had been successful the samples were analysed on a 1 % agarose gel with ethidium bromide (Figures 6.4 and 6.8).

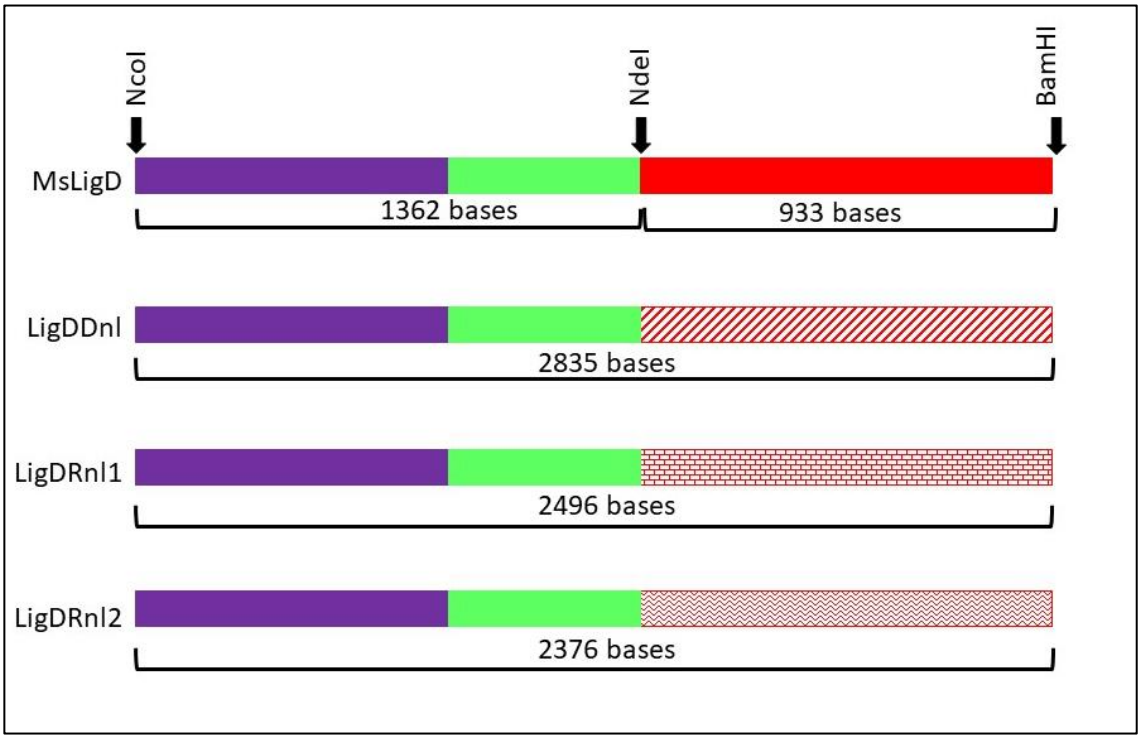


Figure 6.4. The T4 nucleic acid ligase variants of MsLigD. The restriction enzyme cut sites required to move the T4 ligases to the C- terminus of MSLigD, resulting in LigDDnl, LigDRnl1 and LigDRnl2, with the new gene sizes included. Each domain is highlighted: the MsLigD polymerase domain (purple), the PE domain (light green) and the different ligase domains (red). The solid red section is MsLigD ligase, the patterned areas represent T4Dnl (striped) T4Rnl1 (squares) and T4Rnl2 (wavy lines).

6.2.1.2. Gel Extraction

The Bioline isolate II PCR & Gel Kit was used as per the manufacturer's instructions to isolate the DNA for the linearised vector and T4 NA ligase, MsLigD gene and truncated Polymerase and PE domains from MsLigD. The amount of DNA was quantified using the Thermo Scientific Nanodrop one quantification system.

6.2.1.3. Ligating the truncated and normal MsLigD gene into pET16b

The truncated MsLigD was ligated into each of the linearised T4 NA ligase and pET16b vector (Figure 6.4) and full MsLigD was ligated into linearised pET16b, the ratios set up were calculated using the same process as the adenylate kinase-ligases (see Section 4.2.3). The T4 DNA ligase (ThermoFisher) and ligase buffer with ATP (Table 2.5) were used.

6.2.1.4. Confirmation of recombinant LigD plasmids

The new plasmids containing truncated MsLigD + the T4 NA ligases and full length MsLigD were extracted (see Section 2.2.2) and underwent double digestion with *BamHI* and *NcoI* (see Section 2.2.5) to analyse the change in size of the gene fragment (Figure 6.8). The new plasmids (Figures 6.5, 6.6 and 6.7) were also sent to Eurofins for Sanger sequencing (i54 Business Park, Valiant Way, Wolverhampton, WV9 5BG, UK).

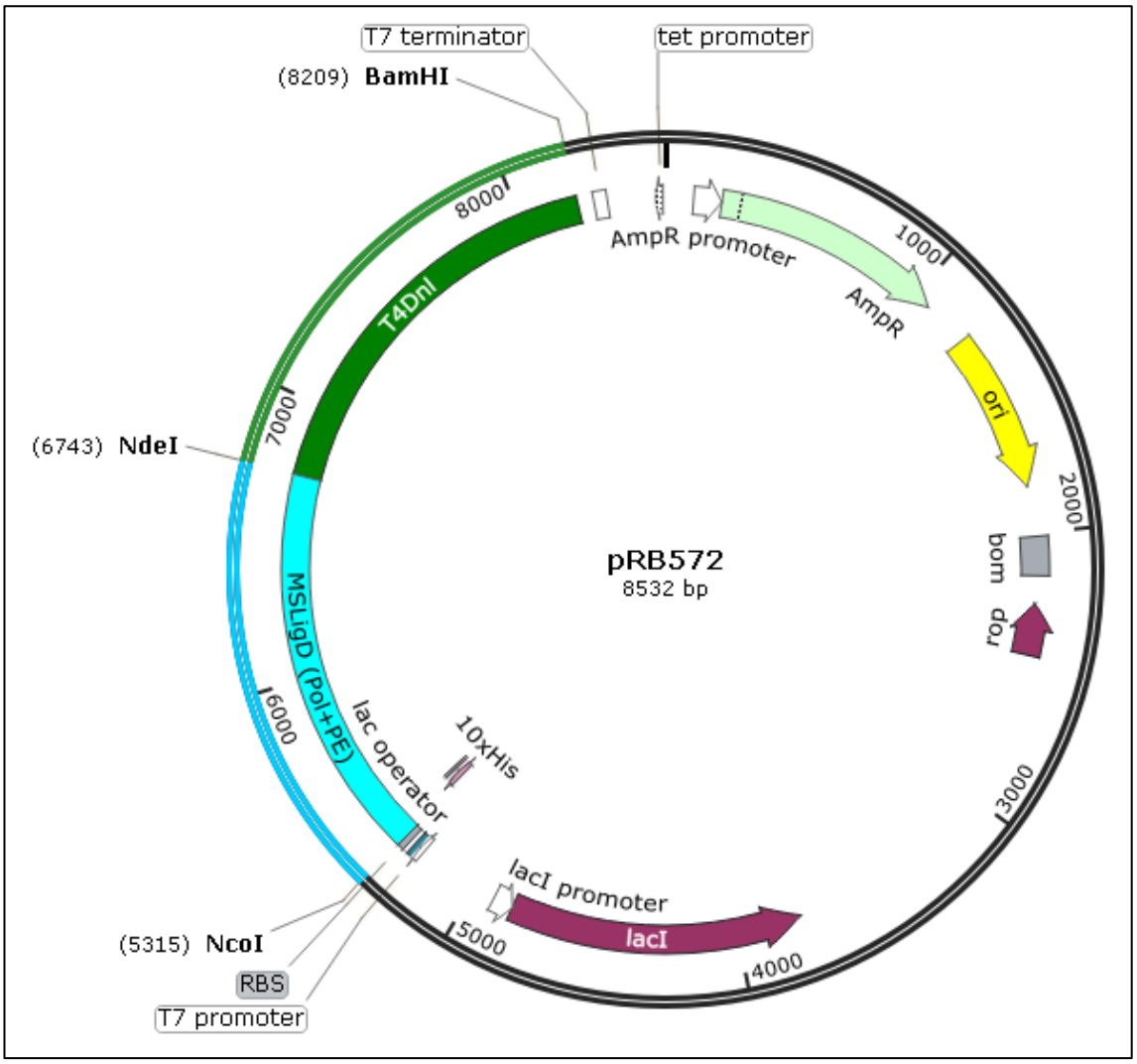


Figure 6.5. The vector map for the fusion of truncated MsLigD and T4Dnl to form LigDDnl (pRB572). The MsLigD polymerase and phosphoesterase domains were between the *NcoI* and *NdeI* restriction sites and the T4Dnl gene between the *NdeI* and *BamHI* restriction sites. This map was created using the SnapGene software.

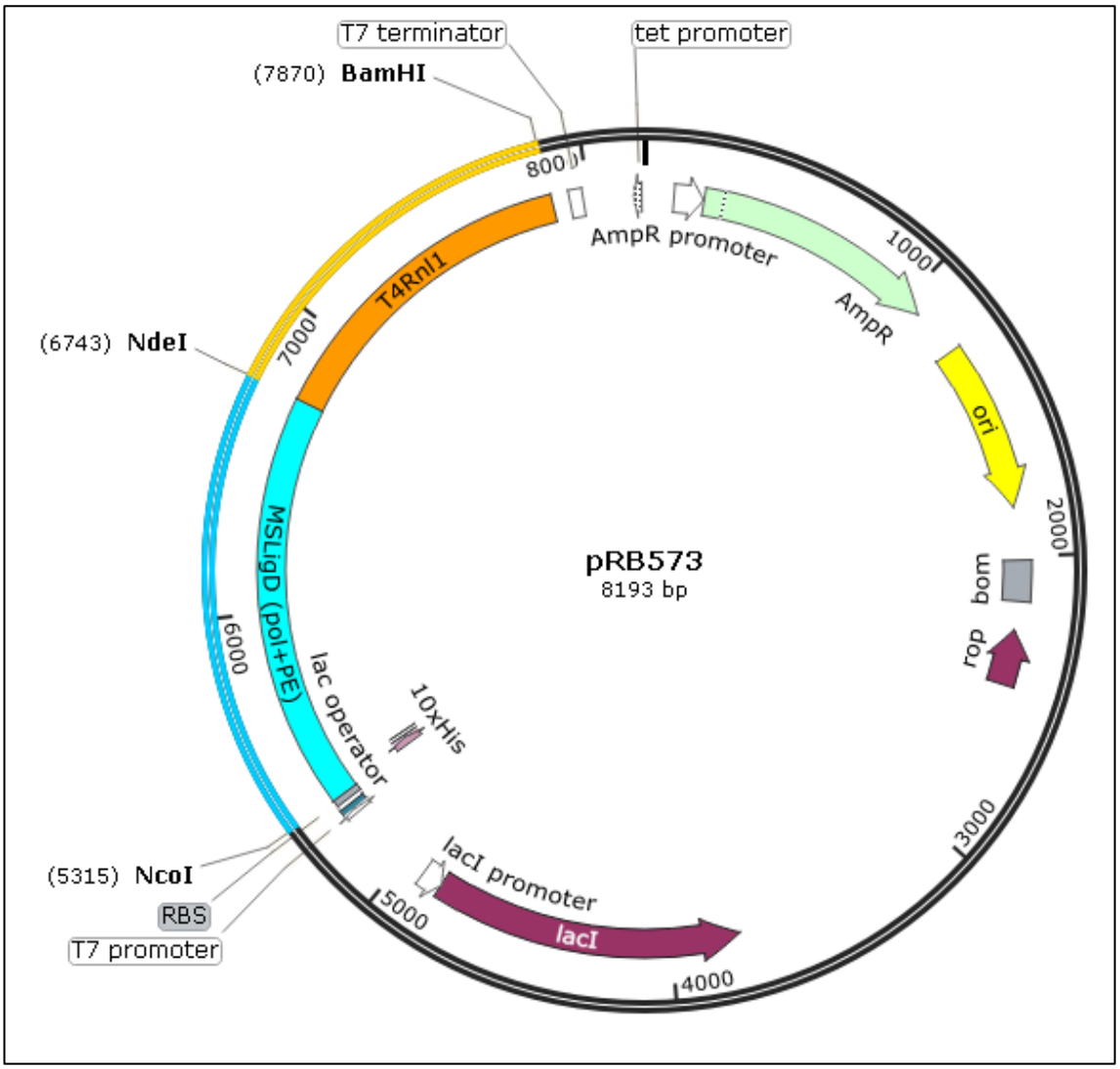


Figure 6.6. The vector map for the fusion of truncated MsLigD and T4Rnl1 to form LigDRnl1 (pRB573). The MsLigD polymerase and phosphoesterase domains were between the *NcoI* and *NdeI* restriction sites and the T4Rnl1 gene between the *NdeI* and *BamHI* restriction sites. This map was created using the SnapGene software.

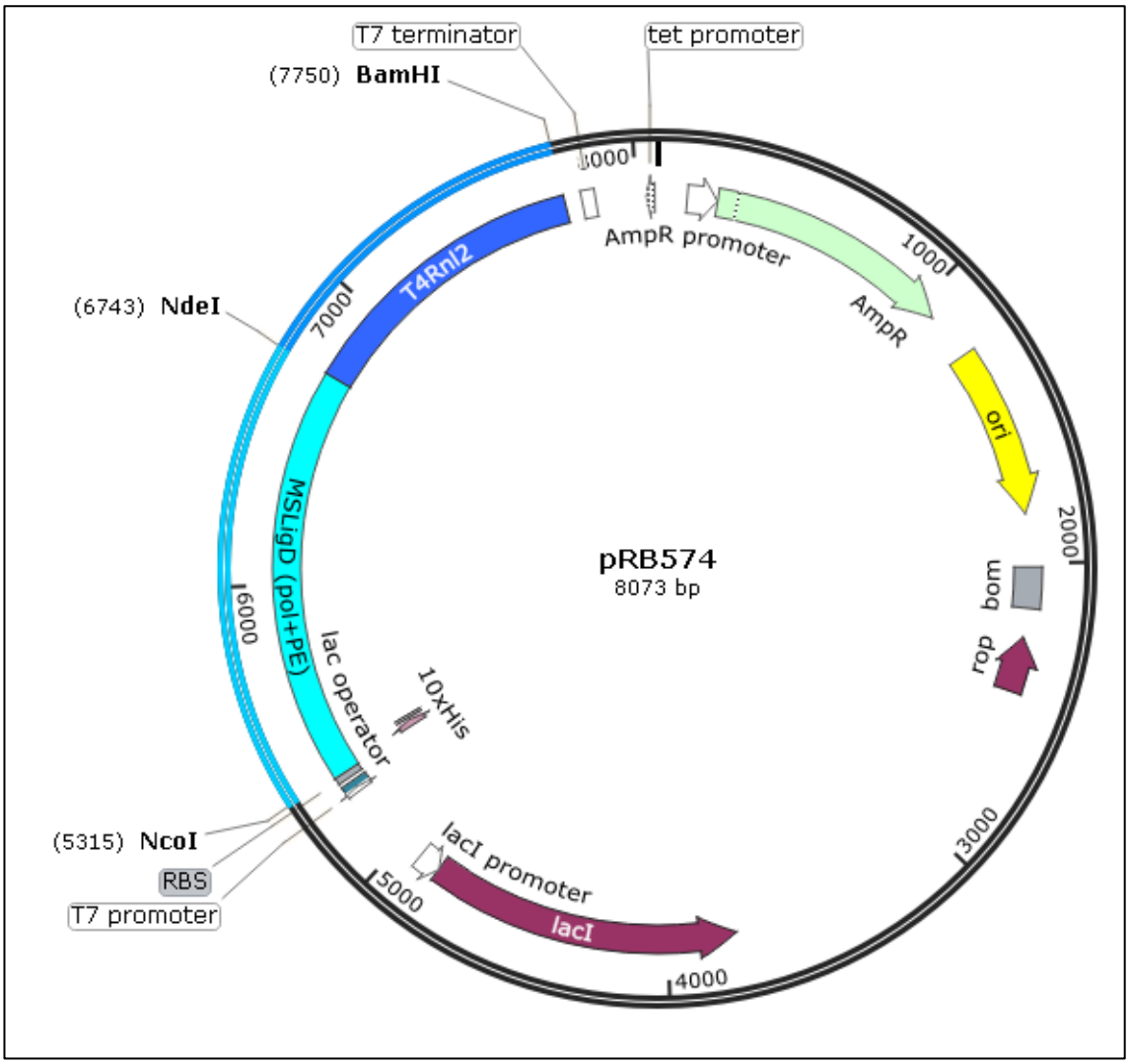


Figure 6.7. The vector map for the fusion of truncated MsLigD to T4Rnl2 to form LigDRnl2 (pRB574). The MsLigD polymerase and phosphoesterase domains were between the *NcoI* and *NdeI* restriction sites and the T4Rnl2 gene between the *NdeI* and *BamHI* restriction sites. This map was created using the SnapGene software.

6.2.2. Protein Characterisation

While purification of *Mycobacterium* proteins encoded by genes from *Mycobacteria* is difficult because mycobacterial genomes are G + C rich, it is not impossible and soluble protein has been obtained using *E. coli* as the bacterium for protein synthesis and expression (Weller *et al.*, 2002). The plasmids were transformed into *E. coli* BL21 (DE3) cells for protein synthesis (see Section 2.2.6). The new proteins were induced and synthesised in 1 L of selective LB media (see Sections 2.3.1, 2.3.2 and 2.3.3). Western blot analysis was carried out (see Section 2.3.7) to confirm the 10x His tag was still present as this is required for purification. The soluble fractions were purified using a Ni-Hi Trap column (see Section 2.3.4) and the salt removed using a PD10 column (see Section 2.3.5). The purification fractions were analysed using 10 % SDS-PAGE (see Section 2.3.6 and Figure 6.7). The amount of each purified LigD protein was quantified using the Bradford Assay (see Section 2.3.8).

6.2.3. Characterisation of biochemical activities

A number of ligation assays were set up (see Sections 6.2.3.1 and 6.2.3.2) and analysed using urea-PAGE (see Section 2.4.3) and ImageJ software. The end point ligation assays were carried out as stated in chapter 2 (see Section 2.4.1) with double stranded substrates 1-8 (Table 2.4). The ligation rate for LigDDnl was determined using time course ligation analysis (see Section 2.4.2) with substrate 1 (double stranded DNA), since this was the one that it worked with best.

6.3 Analysis of the MsLigD-T4 ligase fusions

6.3.1. Recombinant plasmid analysis

To assess the addition of MsLigD polymerase and PE domains, Sanger sequencing (Eurofins) and double digestion analysis were carried out. Double digestions of the new recombinant plasmids and the original T4 ligase plasmids with *NcoI* and *BamHI* (see Section 2.2.5) showed clear and expected changes in size to the gene fragments sites (Figures 6.4 and 6.8).

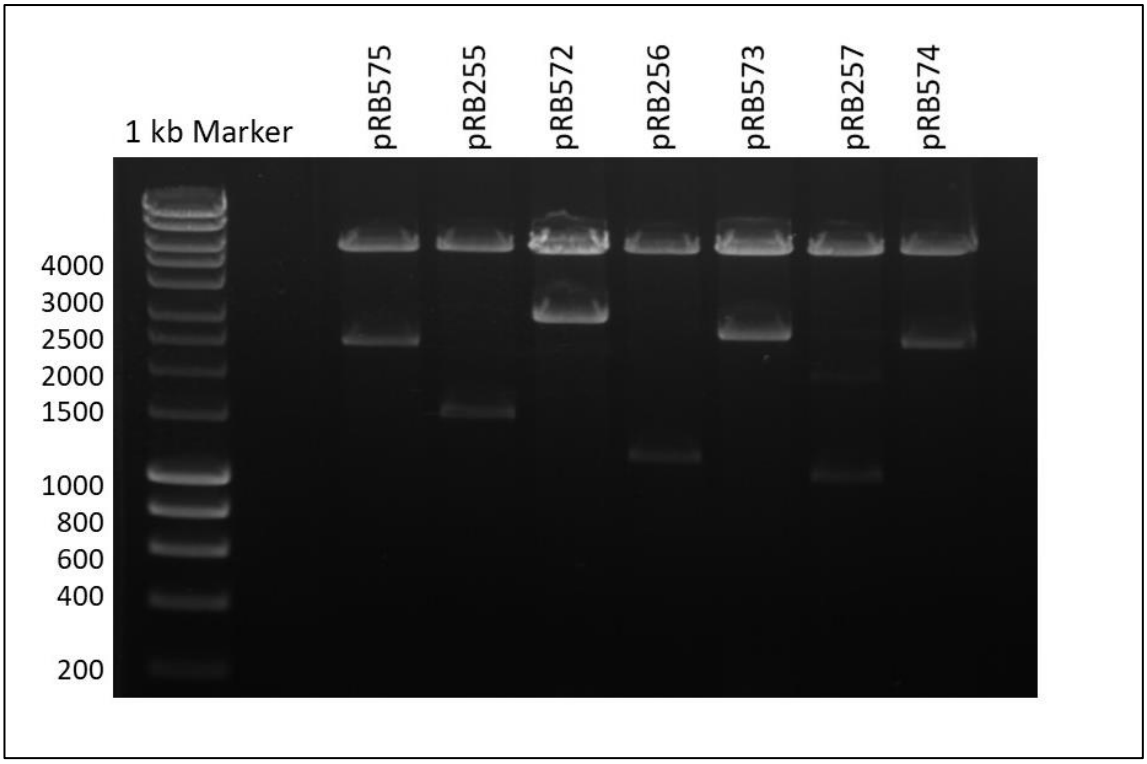


Figure 6.8. Analysis of MSLigD and T4 ligase variants. Double digestions comparisons of full length plasmids containing the genes MsLigD (pRB575), T4Dnl (pRB255), T4Rnl1 (pRB256), T4Rnl2 (pRB257), LigDDnl (pRB572), LigDRnl1 (pRB573) and LigDRnl2 (pRB574). Plasmids were digested with *NcoI* and *BamHI* for 2.5 hours at 37 °C. The expected sizes were: MsLigD 2295 bp, T4Dnl 1474 bp, LigDDnl 2835 bp, T4Rnl1 1134 bp, LigDRnl1 1496 bp T4Rnl2 1014 bp and LigDRnl2 2376 bp.

6.3.2. Protein solubility and purification.

All of the LigD-T4 NA ligase fusions were soluble (Figure 6.9), and the protein concentrations calculated (see Section 2.3.8), the MsLigD purification yield was low (Figure 6.10). The LigD-T4 protein fusions show a clear change in size (Figure 6.10), the expected sizes are: full length MsLigD - 85.5 kDa and the expected sizes for the LigD-T4 fusions are: LigDDnl – 106.5 kDa, LigDRnl1 – 94.8 kDa and LigDRnl2 88.8 kDa. A number of extra bands were present when analysing the LigD fusions, it is not known whether these are contaminating proteins or degradation products.

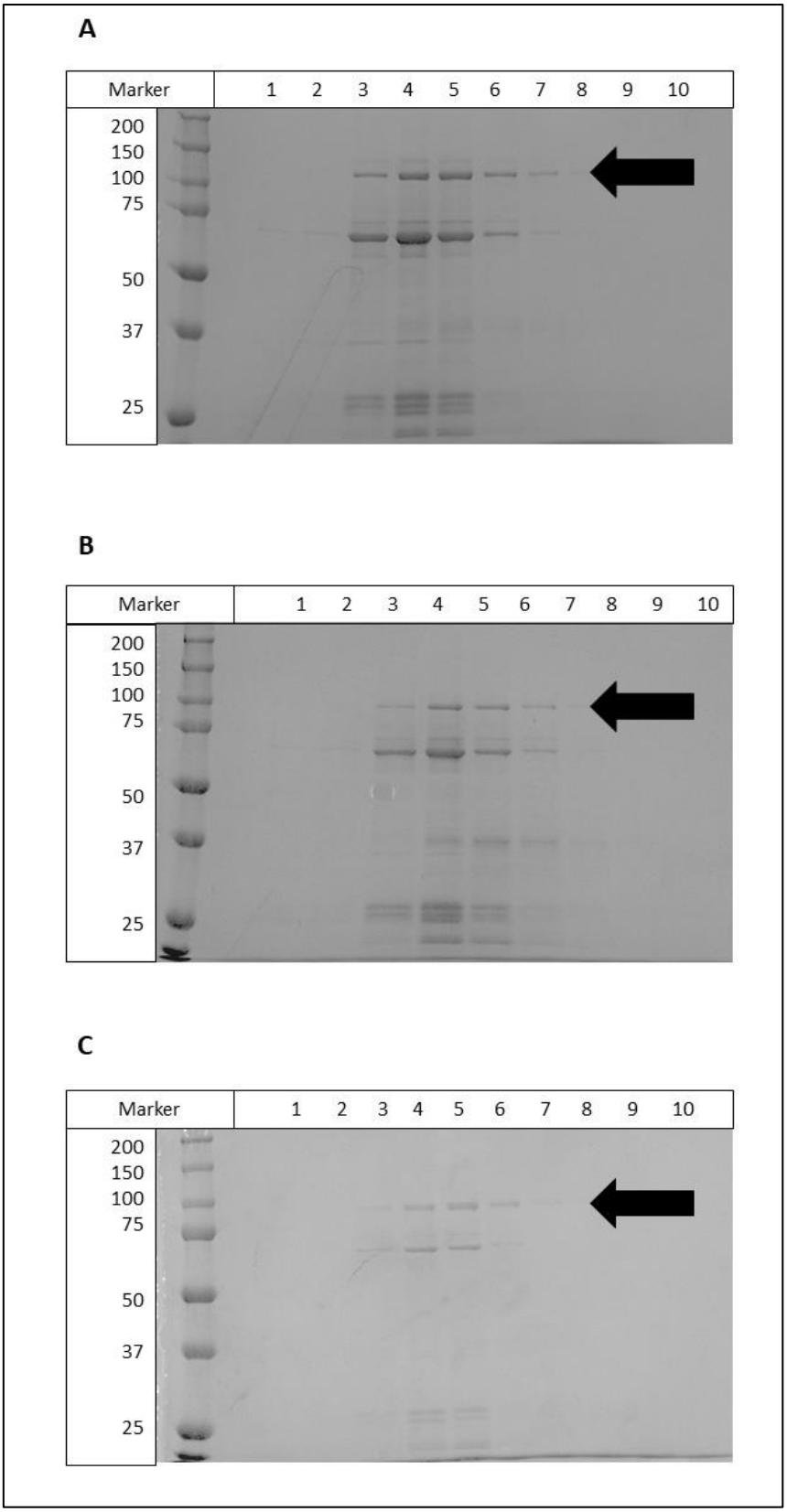


Figure 6.9. Purifications of the LigD-T4 ligases. A) LigDDnl fractions 1-10, approx. protein size 106.5 KDa; B) LigDRnl1 fractions 1-10, approx. protein size 94.7KDa C) LigDRnl2 fractions 1-10, approx. protein size 88.3KDa; all fractions were analysed using 10 % SDS-PAGE, full length proteins indicated by the arrows.

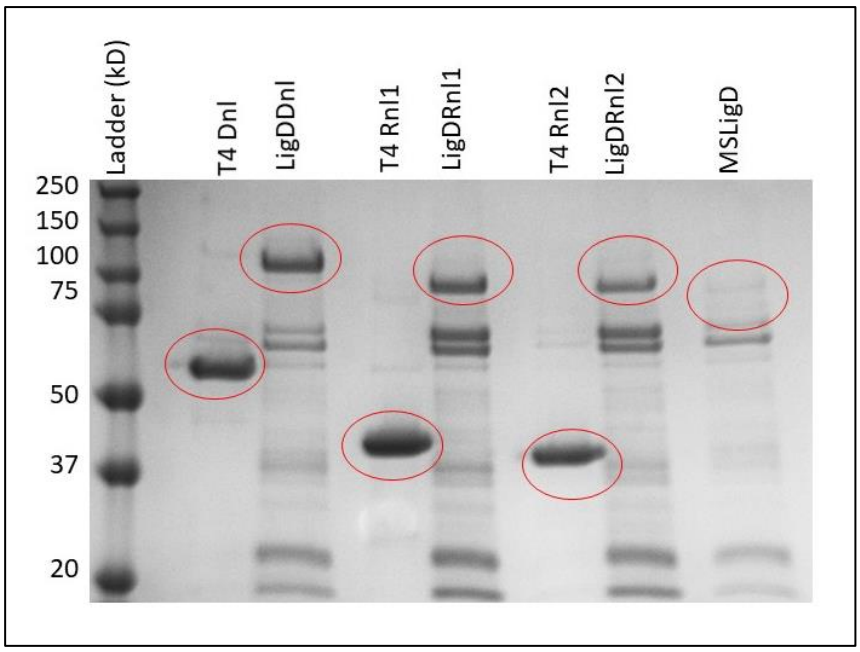


Figure 6.10. Comparison of the purified T4 ligases and variations of MsLigD ligases. SDS-PAGE analysis of the change in size for each of the fully purified T4 NA ligases, MsLigD and the recombinant LigD fusions. The expected molecular weights were: T4Dnl 55.3 kDa, LigDDnl 106.5 kDa, T4Rnl1 43.5 kDa, LigDRnl1 94.76 kDa, T4Rnl2 37.6 kDa and LigDRnl2 88.83 kDa. The correct band sizes are circled. The theoretical molecular weights were calculated using the Expasy compute pI/Mw tool available at https://web.expasy.org/compute_pi/. These proteins were used in the ligation experiments.

6.3.3. Characterisation of the LigD-T4 NA ligases

6.3.3.1 End point ligation analysis

End point ligation analysis was carried out for full length MsLigD, LigDDnl, LigDRnl1 and LigDRnl2 (Figure 6.11). Each of the ligases was incubated with the 8 different 20-mer double stranded substrates (Table 2.4), and MsLigD was also incubated with the 40-mer version of substrate 1 (Figure 6.11 D and E). Of the 8 double stranded 20-mer substrates, LigDDnl could ligate substrate 1 best and ligated a small amount of substrate 7 (Figure 6.11 A and E). There were very low levels of ligation observed with substrates 5 and 7 for LigDRnl1 and LigDRnl2 (Figure 6.11 B, C and E) but the total ligation was too low and no further analysis was conducted.

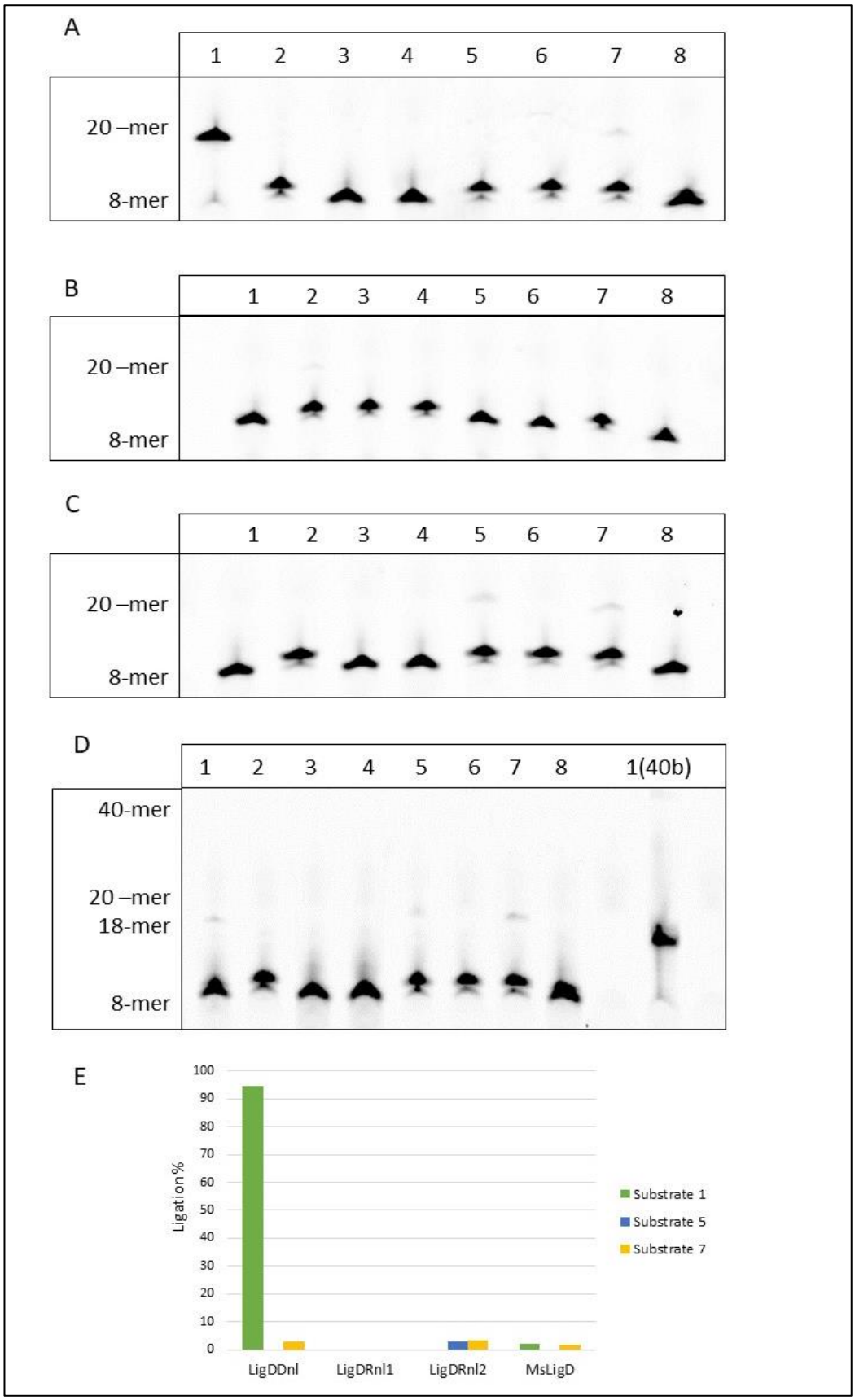


Figure 6.11. End point Ligation Analysis of the LigD-T4 ligases. End point ligation for A) LigDDnl, B) LigDRnl1 and C) LigDRnl2 D) MsLigD with double stranded substrates 1-8 (20-mer) and a 40-mer substrate 1 for MsLigD and E) Analysis of ligation with substrates 1, 5 and 7. All reactions were incubated at 37 °C for 1 hour with ligation buffer containing 1 mM ATP.

6.3.3.2 Initial rate analysis of LigDDnl

The rate of ligation for LigDDnl (Table 6.1) was calculated from quantification of the time course analysis using urea-PAGE (Figure 6.12A) with the ImageJ software (Figure 6.12B).

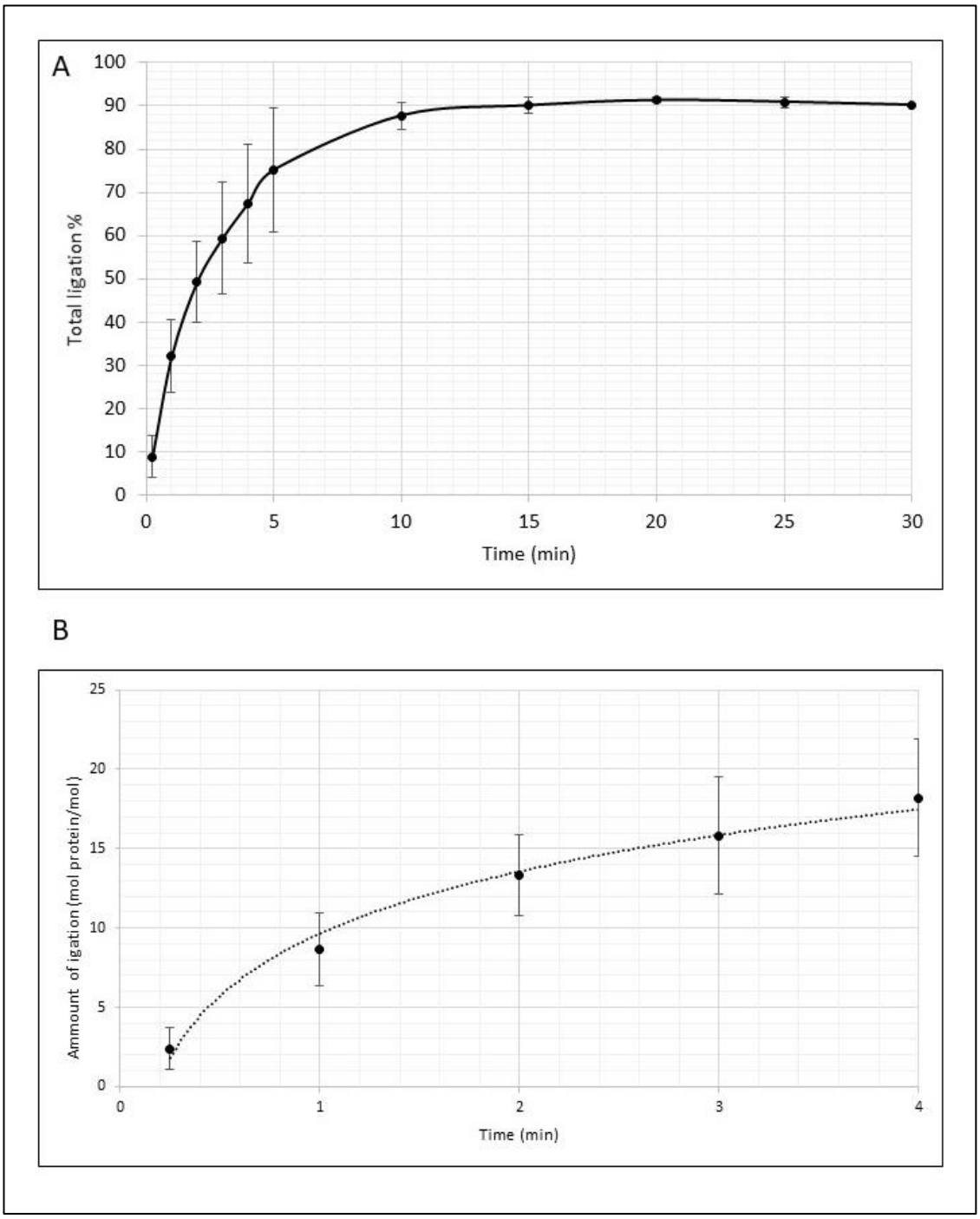


Figure 6.12. Initial rate analysis for LigDDnl. Analysed on 15 % urea-PAGE; A)Quantification of time course ligation analysis repeated in triplicate, with standard deviation included; B) Initial rate analysis, from the triplicate time course ligations with 540 pmol substrate 1 (20-mer) incubated at 37 °C. The line of best fit suggest a non-linear relationship indicating the substrate is ligated quickly at the beginning but the rate then slows after 2 minutes.

Table 6.1. Rate of nick joining in double stranded substrate 1. Ligation of substrate 1 (20-mer) with LigDDnl at 37 °C.

Ligase	Rate of ligation at 37 ^{o}C (mol nicks joined/mol protein/min)
LigDDnl	5.18 (+/- 1.11)

6.4 Discussion

Purification analysis of the 3 new LigD recombinant proteins showed a number of extra bands on the SDS-gels. These bands are likely to be truncated versions of MsLigD without the N- terminal His-tag – there were no other bands present on the Western blot analysis. One knock-on effect of these contaminating MsLigD variants is that the concentration of protein will not be all in the expected full-length (fully active) form. This means that any ligation rates are likely to be under-estimates of the actual rates although it is possible the truncated forms may be active (this wasn't tested for). LigDRnl1 and LigDRnl2 showed severely reduced ligation activity with the double stranded 20-mer substrates (Figure 6.11) and no further analysis was conducted with these proteins. Full length MsLigD ligated small amounts of 20-mer substrates 1 and 7 (Figure 6.11). Even with a large amount of MsLigD compared to the amount of substrate the total ligation was poor (Figure 6.11), and it was therefore not possible to determine a rate for MsLigD with any of the 20-mer substrates or 40-mer substrate 1. LigDDnl showed robust ligation activity with double stranded substrate 1 (DNA) and reduced ligation of substrate 7 (DNA/RNA combination) (Figure 6.11); initial rates of ligation were calculated using the 20-mer double stranded substrate 1.

6.4.1 LigDDnl as an MsLigD alternative

Full optimisation of the conditions required for ligation was not possible, therefore the standard ligation buffer was used in order to compare the MsLigD ligases and the T4 NA ligases. It is possible that the buffer used to determine MsLigD substrate specificity and ligation activity contained too much ATP for optimal ligation activity of the enzyme. Previous studies have suggested that 1 mM of ATP causes ATP trapping which results in reduced ligation efficiency as the final phosphodiester bond formation step is not catalysed. By lowering the concentration of ATP, nick sealing with MsLigD was improved

(Gong *et al.*, 2005; Williamson, Hjerde and Kahlke, 2016). Swapping the ligase part of the enzyme to T4Dnl resulted in an altered substrate profile and, when comparing the profiles for MsLigD and LigDDnl (Figure 6.11), it was observed LigDDnl successfully joins nicks in double stranded substrate 1 with 1 mM ATP and that the ATP trapping phenomenon seen with MsLigD was omitted. Another buffer constituent that may need changing for optimised MsLigD activity is the addition of magnesium as it is noted that *Agrobacterium tumefaciens* LigD utilises manganese more readily than magnesium in nick joining (Zhu and Shuman, 2007; Unciuleac and Shuman, 2013). Both *M. smegmatis* LigD and *A. tumefaciens* LigD consist of polymerase, PE and ligase domains and appear to function similarly (Zhu and Shuman, 2007). It was not possible to align these two LigD proteins as a sequence for *A. tumefaciens* LigD could not be sourced.

LigDDnl was able to successfully ligate double stranded substrate 1 and this substrate was used to calculate the initial ligation rate (Table 6.1, Figure 6.12). All components that make up substrate 1 are DNA and it is well published that MsLigD is a DNA ligase, ligating broken strands of DNA in NHEJ (Della *et al.*, 2004; Gong *et al.*, 2004; Korycka-Machala *et al.*, 2007). Rate analysis was carried out with LigDDnl and 20-mer substrate 1. LigDDnl achieved ~90 % ligation of substrate 1 over 30 minutes at 37 °C (Figure 6.12A) and an initial rate of 5.18 mol nicks joined/mol protein/min (Table 6.1) which is lower than the rate calculated for T4Dnl (Table 4.2). Due to the extra proteins present in the purified fractions this is likely to be an underestimation of the rate for LigDDnl. Published data for the rate of ligation for full length MsLigD is currently not available and it was not possible to calculate a rate for MsLigD in this body of work without first optimising the ligation buffer. Ideally the rate of LigDDnl should be compared to the rate of full length MsLigD to ascertain whether LigDDnl works better as a ligase than full length MsLigD, this preliminary data suggests it is likely to be.

A number of follow up experiments need to be carried out to complete this work, including optimised buffer constituents for MsLigD ligation, improved MsLigD protein solubility studies, initial rate analysis of MsLigD, LigDDnl ligation with 40-mer double stranded substrates and procedures to check the activity of the polymerase and PE

domains with the T4 DNA ligase in LigDDnl. It may also be possible to fuse truncated MsLigD with an NAD dependent ligase to totally remove the need for ATP in buffers and prevent ATP-trapping when ligating DNA/RNA. Fusions with the T4 ligases offer a proof of concept for this type of work and an NAD dependent ligase fusion could also act as the control for altering levels of ATP. This body of work opens up further research that could be conducted with these MsLigD variants and could possibly identify LigDDnl as a novel biotechnology resource. If all parts of LigDDnl are functional, this enzyme could provide an efficient polymerase, PE and ligase for molecular and synthetic biology applications. A number of experiments with *M. smegmatis* have identified a phenotype for knockouts of MsLigD (Korycka-Machala *et al.*, 2006), it would be interesting to see if LigDDnl could complement the phenotype(s) seen in that strain.

Chapter 7

General Discussion

Chapter 7: General Discussion

The results and important conclusions for each chapter have been discussed within them. Therefore, this chapter provides an overview of this body of work and details the possible wider impact of these research outcomes.

This body of work was part of a BBSRC iCASE studentship with Inspiralis, who produce a number of different enzymes for use in academia and the pharmaceutical industry. They retail all of the T4 ligases (T4Dnl, T4Rnl1 and T4Rnl2) and requested novel ligases. The research in this thesis has explored ways to modify the activities of T4 NA ligases using a combination of amino acid substitutions, domain swapping and protein fusions. The T4 NA ligases and different enzymes outlined in this body of work are common place tools for a range of different molecular biological techniques. These techniques are not only crucial for current molecular research methodologies but are also used in protocols for the identification of a number of human and plant diseases (Zhang, Dahlberg and Tam, 2007; Pogue *et al.*, 2014; Hill and Lukiw, 2016). A number of different cloning approaches were used to substitute amino acids in specific motifs, swap whole motifs and create fusion-ligases with enzymes that could provide additional biochemical activities. These modifications to the T4 NA ligase enzymes have demonstrated that chimeric versions of these enzymes have the ability to ligate using new cofactors, to ligate new substrates and that it is possible to create functional chimeras of viral-bacterial enzymes.

7.1. Modifying the T4 NA ligases

The nick joining activities of the T4 NA ligases for double stranded DNA/RNA substrates demonstrated preferences for specific combinations of DNA and RNA (Bullard and Bowater, 2006). Analysis of ligation bias encountered in NGS revealed a structural bias within the T4 RNA ligases (Sorefan *et al.*, 2012; Zhuang *et al.*, 2012; Song, Liu and Wang, 2014). Current publications show attempts to reduce this bias in NGS have focussed on altering the adapters to which miRNAs are ligated (Xu *et al.*, 2015) and little work has been carried out to alter the ligases to reduce this bias.

The amino acid substitutions experiments carried out aimed to identify substrate specific amino acids in motif I and to alter substrate specificity by mutating these amino acids. Previous research had identified that the amino acids in position 6 of motif I, arginine (R) for T4Dnl, leucine (L) for T4Rnl1 and asparagine (N) for T4Rnl2, were involved in interactions with the ribose sugar (Shuman and Lima, 2004). This was based on structural analysis of the T4 RNA ligases and sequence analysis of T4Dnl. Site directed mutational analysis had been carried out extensively on the T4 RNA ligases (Wang *et al.*, 2003; Nandakumar *et al.*, 2004; Pascal, 2008). The other amino acids considered for substitution in motif I of T4Rnl1 and T4Rnl2 – aspartate and histidine, respectively – have not been identified as residues that make contact with the nucleotide substrate but are always conserved in proteins that are Rnl1-like or Rnl2-like. Ligation analysis for each of these ligases demonstrated how single amino acid changes impact ligation activity, notably a total loss of activity for T4Rnl1L104R. Reduced ligation was noted with T4DnlR164N, T4DnlR164L, T4Rnl1D101H and T4Rnl2N40R.

Truncation of T4Rnl2 and substitution of Lys227 to Glu (K227Q) reduced the formation of unwanted ligation products (Viollet *et al.*, 2011). It was this type of substitution that inspired the amino acid substitutions made to motif I. The results (Chapter 3) demonstrated that the residues in the motifs may now be too evolutionarily defined and distinct from each other. Changing these specific residues does not positively alter the ligase activity or specificity but it may be possible using other residues. It is likely the residues selected in Chapter 3 interact so specifically with other residues in the nucleotide binding pocket that altering them affects one or multiple steps in the ligation reaction. The Arg164 residue in T4Dnl interacts with the nucleotide phosphate group at the 3' position (Shi *et al.*, 2018) during step 2 of the ligation reaction. Changing this residue to leucine and asparagine may have affected this interaction resulting in the reduced ligation profiles seen (Figures 3.15 and 3.16). Structural data of the ligases may offer a way around this, for example, using structures where DNA/RNA is bound or unbound would give insight into conformational changes that occur with specific amino acid residues. In T4Rnl2, the phosphate from AMP makes contact with both Lys255 in motif V and Lys35, which then make contact with Lys227 during step 1 of the ligation reaction (Nandakumar, Shuman and Lima, 2006). Targeting residues in the other motifs,

e.g. Lys255, that interact in the reaction may be a way to alter substrate specificity. In T4Rnl2, Arg55 is involved in nick recognition, by severing contact with the ribose in step 2 of the ligation reaction (Nandakumar, Shuman and Lima, 2006). Arg55 isn't in one of the conserved polypeptide motifs and raises an interesting point about the residues between motifs also playing a key role in the ligation reaction and substrate recognition.

Motif switching and addition of the OB domain to T4Rnl1 and T4Rnl2 was achieved using GGC. Creating the recombinant DNA was relatively straightforward and GGC as a technique is quick and reliable, the most time consuming part was primer design. While engineering the recombinant DNA was successful, the new proteins were either insoluble (T4Rnl2M1 and T4Rnl2OB) or inactive (T4Rnl1M1 and T4Rnl1OB). This type of inactivity has been seen before when directly binding protein domains (Bai, Ann and Shen, 2005; Klein *et al.*, 2014). While some success has been seen with addition of archaeal DNA binding domains to T4Dnl to improve ligation (Wilson *et al.*, 2013), it is apparent that a more targeted approach needs to be applied to modify the T4 RNA ligases. There are many more substitutions of specific amino acids in each of the different motifs that could be tried. For example, by changing amino acids one at a time it could be possible to understand better how each residue interacts with the substrate and other residues in the nucleotide binding pocket.

7.2. Chimeric T4 NA ligases

Generation of chimeric ligases seemed a logical next step, due to previous successful examples, adenylate kinase fused to T4Dnl (Liu *et al.*, 2015), T4Dnl fused to DNA binding proteins (Wilson *et al.*, 2013) and fusion of a polymerase to a DNA binding protein to improve nucleic acid modifying proteins (Wang *et al.*, 2004). Published work had focussed heavily on fusions with T4Dnl, but there were no efforts to do the same with the T4 RNA ligases. The T4 RNA ligases are used less frequently in molecular biology techniques but are still important tools – especially in NGS and other sequencing techniques. Two different sets of chimeric ligases were created, one set with adenylate kinase at the N- terminus, which generated chimeric AK-ligases that could ligate using ADP instead of ATP. The second set had polynucleotide kinase at the N- terminus of the

ligases, whereby the chimeric enzyme would be able to transfer phosphate and correctly prime the ends of the DNA/RNA for ligation. There was noted success with both sets of these fusion proteins, all of the fusion ligases were soluble and most retained ligation activity. These fusions resulted in novel ligases with specific new activities. Notably all of the AK-ligases could utilise ADP for nick sealing double stranded substrates (Figure 4.13, Table 4.2). In the instance of AKRnl1, a much smaller amount of the new fusion enzyme was used to achieve the same total ligation percentage, 150 pmol could be used instead of 600 pmol (Figure 4.14).

The fusion of Pnk to the N- terminus of the T4 ligases elicited mixed results. While all of these fusion enzymes were soluble, there were large differences in the activities of the Pnk-ligases. The ligase part of the fusion remained functional in PnkDnl and PnkRnl2 (Figures 5.14 and 5.15), but PnkRnl1 displayed a severe reduction in ligation of the double stranded substrates (Figure 5.14). A second set of substrates were also used to test the Pnk part of these fusions, and of the three Pnk-ligase fusions only PnkRnl2 was able to successfully ligate these substrates under the conditions tested. These Pnk substrates were double stranded and included nicks where the ends were 3’ phosphate and 5’ OH, and required the Pnk to remove the phosphate from the 3’ end and add phosphate to the 5’ end for ligation to be possible. Further analysis of the Pnk-ligases was carried out and the amount of phosphate in solution measured (Figure 5.19). A more structural based-design approach should be used when considering ways to improve these Pnk-ligases. The use of linkers to join Pnk to the T4 ligases is one possible method. Often linkers are used as a way to immobilise enzymes to a structure or surface, these linkers can be chemical or peptide linkers (Wu *et al.*, 2018). A number of factors need to be considered when using linkers, e.g. the linker length affects the distance between the two proteins and can affect substrate binding (Bai, Ann and Shen, 2005; Klein *et al.*, 2014). Naturally occurring linkers average a length of 10 +/- 5.8 residues (George and Heringa, 2003). Linkers fall into 3 groups: flexible, cleavable and rigid (Klein *et al.*, 2014). Flexible linkers contain Gly, Ser and Thr; a common flexible linker is Gly_4Ser_n although flexible linkers can limit domain separation (Klein *et al.*, 2014). A factor to consider when designing or choosing rigid linkers is whether they are helical or non-helical. Helical linkers rigidly separate two domains; non-helical linkers, while also rigid,

are proline-rich which isolates the linker from the domains it joins (George and Heringa, 2003). These points offer a multitude of options to consider when joining Pnk to the T4 ligases – especially T4Dnl and T4Rnl1 which are larger and may require a rigid, non-helical linker to keep the nucleotide binding sites available.

It was noted that PnkDnl had high levels of free phosphate in solution, indicating that the part of the enzyme that removed phosphate was functional but the second active site, for attaching phosphate, was not. As the largest of the T4 NA ligases it is possible that T4Dnl is causing steric inhibition of Pnk resulting in phosphate removal but not phosphate attachment. While AKDnl demonstrated a successful T4Dnl fusion, a high resolution, crystallographic structure for T4Dnl would help in identifying where best to create fusions with larger more complex enzymes, like Pnk. Considering how readily available T4Dnl is and that it is a commonly used molecular biology tool, it is surprising that there is no crystallographic structure available. Efforts to create a theoretical model using the Phyre2 software (Appendix 2) yielded questionable results because T4Dnl is modelled against human DNA ligase which is a much larger eukaryotic enzyme. Given the importance of T4Dnl it would be incredibly worthwhile for someone to obtain the structure; however the fact that a structure is not available suggests it may be difficult to achieve.

At the time of submission of this work a crystal structure at 1.40 Å resolution for T4Dnl in complex with a 21 base duplex DNA was published (Shi *et al.*, 2018). This study is important because it not only demonstrates the structure of T4Dnl but also how T4Dnl interacts with DNA (Figure 7.1A) (Shi *et al.*, 2018). Had this structure been available before completion of this work, it would have provided much needed guidance and insight into the binding capacity of T4Dnl and where best to make modifications to this ligase. The Arg164 residue, which was mutagenized in Chapter 3, was previously thought to interact with the ribose sugar on the DNA (Shuman and Lima, 2004). The structural model of DNA bound to T4Dnl shows Arg164 in fact interacts with the nucleotides phosphate group at 3' position (Shi *et al.*, 2018). This structural model also provides information about the OB domain and motif VI residues which make contact with the

DNA substrate (Shi *et al.*, 2018). The OB domain is essential for transferring AMP to the lysine in motif I, Lys159, in step 1 of the ligation reaction. The OB domain also positions the DNA for steps 2 and 3 of the ligation reaction by making contacts with the DNA backbone (Shi *et al.*, 2018). This information would have provided much needed guidance when trying to make changes to T4Dnl and would have impacted the decision to move the OB domain, including motif VI, to the T4 RNA ligases (Chapter 3). A number of residues in the OB domain (residues 387-487) have specific roles in the ligation reaction. Asn404, Ser447, Asp448, Arg450 and Lys455, interact with the DNA backbone (Shi *et al.*, 2018) and may be potential targets to alter the substrate specificity of T4Dnl. Three residues identified in motif VI Arg465, Asp476 and Lys468 also make contact with the DNA strands (Shi *et al.*, 2018). These residues were not targeted and on reflection it seems this domain may be the best way to change the substrate specificity of T4Dnl. Due to the significant interactions of the OB domain and its interactions with the DNA, it would be better to look at how the OB domain in other T4Dnl-like ligases functions. It does not seem logical to reattempt to attach this large domain to the T4 RNA ligases which lack this region and are sequentially different from T4Dnl (Figure 3.3).

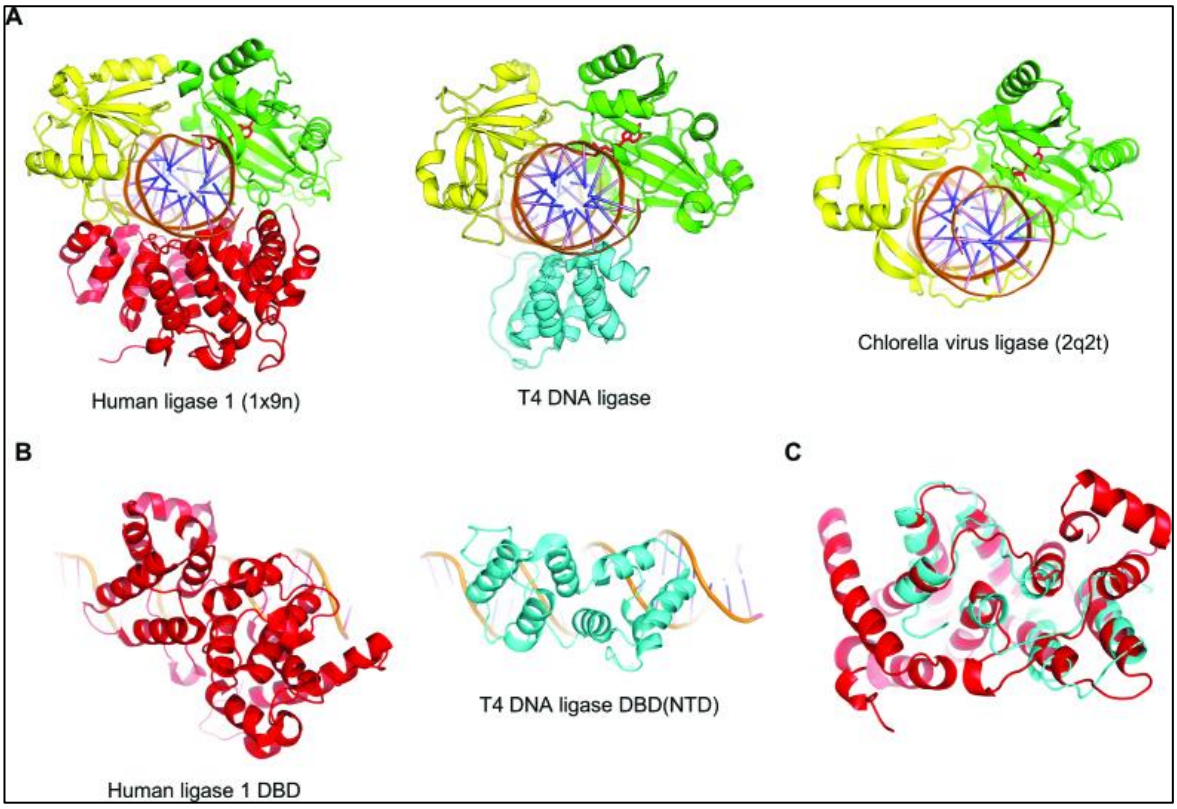


Figure 7.1. Comparison of T4Dnl, Chlorella virus ligase and Human Ligase I. A) Human ligase I, T4Dnl and Chlorella virus ligase in complex with DNA. The DNA binding domains of Human liase I (red) and T4Dnl (blue). B) The DNA binding domains of human ligase I and T4Dnl and C) superposition of these two binding domains (Shi *et al.*, 2018). Comparison of these ligases and their domains demonstrates the similarity of T4Dnl to Human ligase 1.

The model constructed using Phyre2 software (Appendix 2) was most similar to human ligase I, this is supported by the published structural model which was described as being evolutionarily similar to eukaryotic ligases DNA-binding mechanism (Figure 7.1) (Shi *et al.*, 2018). Comparison of Human ligase I, T4Dnl and *Chlorella* virus ligase revealed a DNA binding domain that is absent in *Chlorella* virus (Shi *et al.*, 2018). Furthermore, superposition of the DNA binding domains from Human ligase I and T4Dnl (Figure 7.1) demonstrated the high degree of similarity between these two protein domains. This DNA binding domain in T4Dnl lacks some of the structural elements found in eukaryotic and archaeal ligases (Shi *et al.*, 2018).

As successful fusion ligases were created with proteins from different organisms, it seemed sensible to consider naturally occurring fusions (with multiple activities) for biological engineering. An enzyme that demonstrates multiple biochemical activities is LigD from *M. smegmatis*. MsLigD contains a primase, phosphoesterase and a ligase (Gong *et al.*, 2004) and plays a vital role in NHEJ of DSBs; each of the different enzyme activities that make up MsLigD have different functions in the nick sealing process (Aniukwu, Glickman and Shuman, 2008). Studies have demonstrated that, while functional, the ligase part of MsLigD is not very active (Gong *et al.*, 2005; Zhu and Shuman, 2006; Aniukwu, Glickman and Shuman, 2008) and, as such, it is of little value as a molecular biological tool. It may become valuable if the ligation activity could be improved, by replacing the ligase part of MsLigD with the T4 NA ligases it may be possible to achieve improved ligation. All of the T4LigD fusions were soluble but significant ligation activity was only seen with LigDDnl (Figure 6.11 A and E), which was able to ligate 90 % of substrate 1 over a 30 minute incubation at 37 °C (Figure 6.12), a noted improvement as MsLigD ligated very small amounts of substrates 1 and 7 in comparison (Figure 6.11). LigDRnl2 showed some nick sealing of substrates 5 and 7, but this was at a very low level considering there was an abundance of ligase in the end point reaction (Figure 6.11 D and E).

LigDDnl offers a possible solution to the ATP trapping-phenomenon detailed in Chapter 6. ATP trapping occurs when the ligase dissociates from the app-DNA and reacts with ATP to form ligase-AMP which is unable to rebind to the app-DNA and complete the ligation reaction. This results in no phosphodiester bond formation (Gong *et al.*, 2004). It is well known that T4Dnl works well with 1 mM ATP – nearly all commercial buffers contain ATP at this concentration (New England Biolabs and Sigma both suggest using this concentration of ATP in ligation reactions). Creating LigDDnl provides an enzyme that has three separate nucleic acid modifying functions that work well at an ATP concentration of 1 mM. Another way to omit ATP-trapping would be the addition of truncated MsLigD to an NAD-dependent ligase.NAD-dependent ligases use NAD instead of ATP for phosphodiester bond formation in DNA breaks, removing the need for ATP in the reaction. The polymerase and PE domains in MsLigD do not require ATP to carry out their modifications to the DNA break. The addition of the T4 ligases offers proof of

principle that the ligase part of MsLigD can be replaced and the ligases remain functional.

7.3. Future work

The work detailed in this thesis has started the biochemical analysis of these different enzyme fusions, but there is further work that could be carried out. It would be interesting to see if, like AKRnl1, lower amounts of AKDnl and AKRnl2 can be used to achieve the same total ligation. It would also be beneficial to check whether the addition of AK to each of the T4 ligases has improved solubility. As an iCASE funded studentship it would also be interesting to note how desirable these new fusion ligases are for research and commercial purposes – by contacting different institutions (both in academia and in pharmaceuticals) for feedback on ligase issues. For PnkRnl2, more work could be carried out to determine the quantifiable data for NGS using this enzyme. It would be interesting to test a panel of miRNAs, using both normal adapter ligation protocols and PnkRnl2 in place of T4Rnl2. Analysing both sets of quantifiable data would help to determine whether the differences in miRNA levels used to identify disease (Zhao *et al.*, 2010) are caused by definite mis-regulated miRNA production or by a problem with the maturation process of miRNAs resulting in different forms of miRNA. There are a number of enzymes involved in the miRNA maturation process, with one pathway in eukaryotes (Figure 7.2) utilising Dicer and Drosha to form the 22-nt mature miRNA needed for mRNA transcription (Mendes, Freitas and Sagot, 2009). Analysis of a number of miRNAs showed that some play a role in a number of cancers and that it is unclear whether different cancers associated with the same miRNAs have the same miRNA dysfunction (deletion, up-regulation, down-regulation and mutation) (Lu *et al.*, 2008). It is important to remember the T4 ligases require 5’ phosphate and 3’OH for ligation (Sirotkin *et al.*, 1978; Wang and Shuman, 2001; Sriskanda, 2002; Wang *et al.*, 2003) and cannot ligate nicks in DNA or RNA when the nicked strands do not have this. Similarly, miRNAs that do not mature correctly and do not have a 5’ phosphate will never be ligated to the adapter, giving a false result for down regulation of the miRNA – when in fact the miRNA is incorrectly formed. There are a number of pathways, such as signalling pathways, which utilise a number of genes, dysfunction of any single gene in the pathway can result in the same disease and a similar rationale has been revealed

with miRNA sets (miRNAs implicated in the same disease). If a large proportion of the miRNAs in the set are associated with a disease, the other members also have a higher probability to be associated with that disease too (Lu *et al.*, 2008).

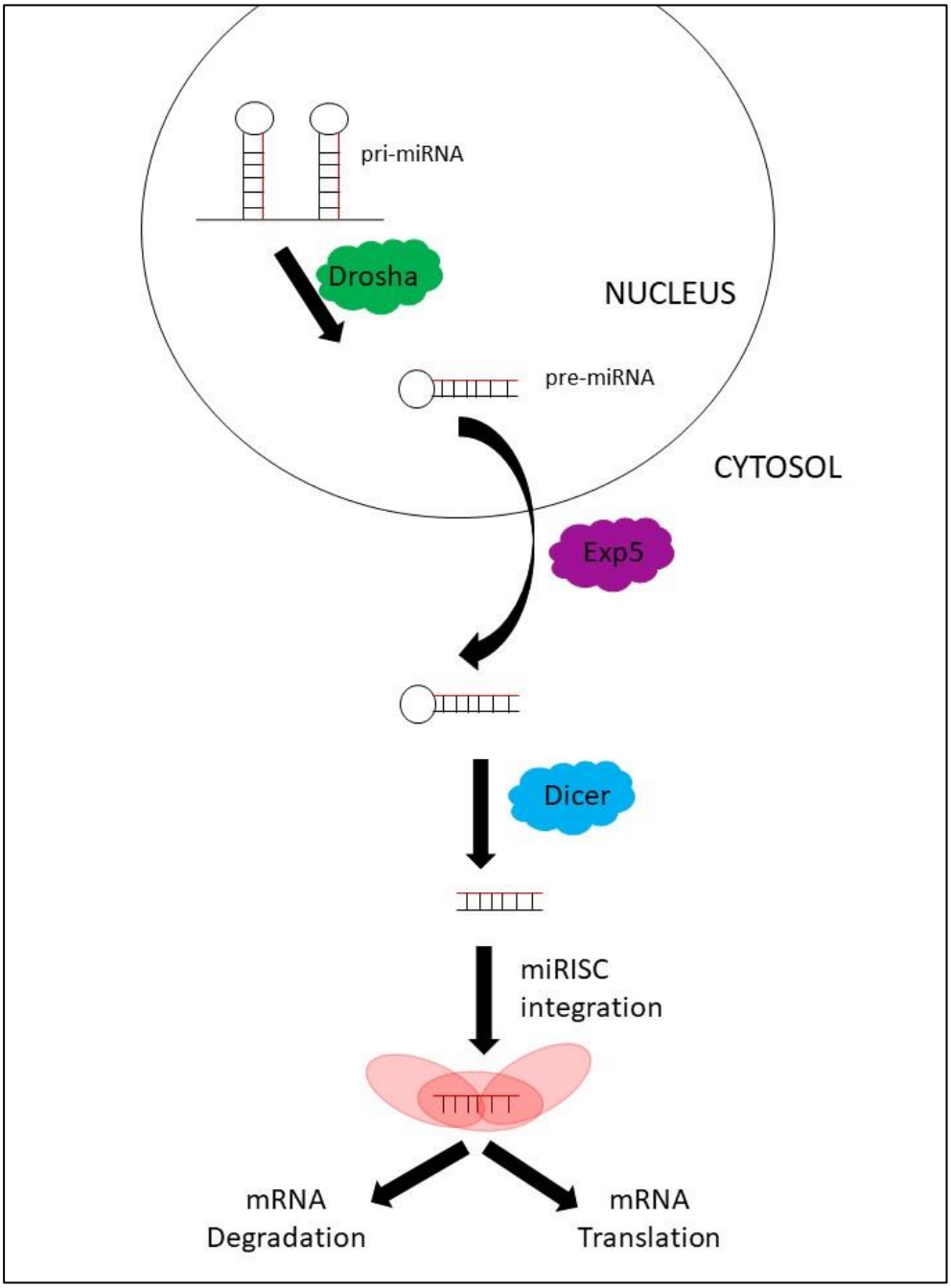


Figure 7.2. A pathway for miRNA maturation in eukaryotic cells. In the nucleus, the pri-miRNA is cleaved by Drosha to give rise to pre-miRNA; the pre-miRNA is transported to the cytosol by Exportin 5 (Exp5). Dicer then cleaves the double stranded stem to give a 22-nt miRNA:miRNA duplex. The duplex unwinds and the miRNA strand is incorporated into the miRNA Induced silencing Complex (miRISC).

Accurate miRNA detection is vital in not only life threatening conditions but also conditions that are debilitating and hard to diagnose. For example, endometriosis is an inflammatory condition which causes severe and chronic pain in sufferers and can take anywhere between 8 and 12 years to diagnose (Agrawal *et al.*, 2018). A number of mis-regulated miRNAs are associated with endometriosis and offer a promising non-invasive detection method (Agrawal *et al.*, 2018) but only if current NGS miRNA detection and quantification methods improve. As stated above, for PnkDnl and PnkRnl1 it may be possible to create fusion enzymes using a number of different linkers. T4Dnl and T4Rnl1 are slightly bigger ligases and a linker would allow for more flexibility and distance between the two enzymes which may prevent steric inhibition. Steric inhibition was identified with *Chlorella* virus DNA ligase when the DNA contained a bulky methyl group at the nick (Zhu and Shuman, 2008), a much smaller structure than addition of another protein to the ligase. This addition of a linker is something that would be possible with a large industrial partner who could screen ligases with linkers using high-throughput analysis.

It would be useful to see whether the primase and phosphoesterase parts of the T4LigD variants are still functional, especially for LigDDnl as this was the most efficient LigD fusion, and to identify the different types of substrates that this new fusion could ligate. Experiments could also be carried out with LigDRnl1 to ascertain if it is able to ligate the hairpin structure or if all activity has been lost. The hairpin substrate relates to tRNAs which seem to be an *in vivo* target for T4Rnl1 and it would be interesting to see if these new ligases could work in vivo.

The ligation buffer used for MsLigD and the LigD variants may need optimising, as previous research has shown that MsLigD prefers manganese to magnesium (Unciuleac and Shuman, 2013) and that 1 mM ATP causes an ATP trapping phenomenon (Gong *et al.*, 2004; Williamson, Hjerde and Kahlke, 2016). The buffer used may be the reason MsLigD could not ligate any of the double stranded substrates fully. There is also the option to carry out complementation assays with LigDDnl in *M. smegmatis* to observe the effect this new ligase has on NHEJ *in vivo*. Another way to check the effect of ATP

would be to use an NAD dependent enzyme with the truncated version of LigD. This would omit the need for ATP for ligation and could act as a control when altering the concentrations of ATP in the buffer.

To summarise, the work outlined in this thesis demonstrates the ability to create novel ligases using a variety of molecular and synthetic biology methods. Amino acid substitutions produced a variety of active ligases but none with novel ligation activities. Using a more targeted, structural approach to amino acid substitutions might be a better way to create novel ligases. Domain swapping was logical on paper, but it often resulted in insoluble and inactive enzymes; considering the recently published structure of T4Dnl (Shi *et al.*, 2018) and the structural differences of the T4 ligases, a more targeted approach could now be used to attempt to successfully carry out these types of domain swaps. Chimeric fusion enzymes were a more successful way of creating novel ligases where both parts of the fusion were functional. This also opens up the possibility of exploring the use of linkers in the future, as a way to join the ligases to larger enzymes and create proteins with new ligation activities or improved efficiencies. These new ligases spark endless possibilities of chimeric enzymes, where both enzymes are used in specific protocols or utilising specific activities of different enzymes to improve the T4 NA ligases.

Chapter 8

References

Chapter 8: References

Abele, U. and Schulz, G. E. (1995) 'High-resolution structures of adenylate kinase from yeast ligated with inhibitor Ap $_5$ A, showing the pathway of phosphoryl transfer', *Protein Science*, 4(7), pp. 1262–1271. doi: 10.1002/pro.5560040702.

Agrawal, S. *et al.* (2018) 'The miRNA mirage: How close are we to finding a non-invasive diagnostic biomarker in endometriosis? a systematic review', *International Journal of Molecular Sciences*. Multidisciplinary Digital Publishing Institute, p. 599. doi: 10.3390/ijms19020599.

Akey, D. *et al.* (2006) 'Crystal structure and nonhomologous end-joining function of the ligase component of Mycobacterium DNA ligase D', *Journal of Biological Chemistry*. American Society for Biochemistry and Molecular Biology, 281(19), pp. 13412–13423. doi: 10.1074/jbc.M513550200.

Amitsur, M., Levitz, R. and Kaufmann, G. (1987) 'Bacteriophage T4 anticodon nuclease, polynucleotide kinase and RNA ligase reprocess the host lysine tRNA.', *The EMBO journal*. Wiley-Blackwell, 6(8), pp. 2499–2503. doi: 10.1002/J.1460-2075.1987.TB02532.X.

Aniukwu, J., Glickman, M. S. and Shuman, S. (2008) 'The pathways and outcomes of mycobacterial NHEJ depend on the structure of the broken DNA ends', *Genes and Development*. Cold Spring Harbor Laboratory Press, 22(4), pp. 512–527. doi: 10.1101/gad.1631908.

Ansorge, W. *et al.* (1987) 'Automated DNA sequencing: Ultrasensitive detection of fluorescent bands during electrophoresis', *Nucleic Acids Research*. Oxford University Press, 15(11), pp. 4593–4602. doi: 10.1093/nar/15.11.4593.

Arber, W. (1965) 'Host-controlled modification of bacteriophage', *Annual Review of Microbiology*. Annual Reviews 4139 El Camino Way, P.O. Box 10139, Palo Alto, CA 94303-0139, USA, 19(1), pp. 365–378. doi: 10.1146/annurev.mi.19.100165.002053.

Armstrong, J., Brown, R. S. and Tsugita, A. (1983) 'Primary structure and genetic organization of phage T4 DNA ligase', *Nucleic Acids Research*. Oxford University Press, 11(20), pp. 7145–7156. doi: 10.1093/nar/11.20.7145.

Bai, Y., Ann, D. K. and Shen, W.-C. (2005) 'Recombinant granulocyte colony-stimulating factor-transferrin fusion protein as an oral myelopoietic agent.', *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, 102(20), pp. 7292–6. doi: 10.1073/pnas.0500062102.

Balasubramanian, A. and Ponnuraj, K. (2010) 'Crystal Structure of the First Plant Urease from Jack Bean: 83 Years of Journey from Its First Crystal to Molecular Structure', *Journal of Molecular*

Biology. Academic Press, 400(3), pp. 274–283. doi: 10.1016/j.jmb.2010.05.009.

Bellacosa, A. and Moss, E. G. (2003) 'RNA repair: Damage control', *Current Biology*. Cell Press, pp. R482–R484. doi: 10.1016/S0960-9822(03)00408-1.

Berry, M. B. *et al.* (2006) 'Crystal structure of ADP/AMP complex of Escherichia coli adenylate kinase', *Proteins: Structure, Function and Genetics*, 62(2), pp. 555–556. doi: 10.2471/BLT.06.032334.

Bétermier, M., Bertrand, P. and Lopez, B. S. (2014) 'Is Non-Homologous End-Joining Really an Inherently Error-Prone Process?', *PLoS Genetics*. Edited by S. Jinks-Robertson. Public Library of Science, p. e1004086. doi: 10.1371/journal.pgen.1004086.

Betts, M. J. and Russell, R. B. (2003) 'Amino-Acid Properties and Consequences of Substitutions', in *Bioinformatics for Geneticists: A Bioinformatics Primer for the Analysis of Genetic Data: Second Edition*, pp. 311–342. doi: 10.1002/9780470059180.ch13.

Bhattarai, H., Gupta, R. and Glickman, M. S. (2014) 'DNA ligase C1 mediates the LigD-independent nonhomologous end-joining pathway of Mycobacterium smegmatis.', *Journal of bacteriology*. American Society for Microbiology, 196(19), pp. 3366–76. doi: 10.1128/JB.01832-14.

Blanga-Kanfi, S. *et al.* (2006) 'PrrC-anticodon nuclease: Functional organization of a prototypical bacterial restriction RNase', *Nucleic Acids Research*. Oxford University Press, 34(11), pp. 3209–3219. doi: 10.1093/nar/gkl415.

Bondos, S. E. and Bicknell, A. (2003) 'Detection and prevention of protein aggregation before, during, and after purification', *Analytical Biochemistry*. Academic Press, 316(2), pp. 223–231. doi: 10.1016/S0003-2697(03)00059-9.

Boyce, S. and Tipton, K. F. (2001) 'Enzyme Classification and Nomenclature', in *Encyclopedia of Life Sciences*. Chichester: John Wiley & Sons, Ltd. doi: 10.1038/npg.els.0000710.

Branton, D. *et al.* (2008) 'The potential and challenges of nanopore sequencing', *Nature Biotechnology*. Nature Publishing Group, pp. 1146–1153. doi: 10.1038/nbt.1495.

Brissett, N. C. and Doherty, A. J. (2009) 'Repairing DNA double-strand breaks by the prokaryotic non-homologous end-joining pathway', *Biochemical Society Transactions*, 37(3), pp. 539–545. doi: 10.1042/BST0370539.

Bullard, D. R. and Bowater, R. P. (2006) 'Direct comparison of nick-joining activity of the nucleic acid ligases from bacteriophage T4', *The Biochemical journal*. Portland Press Ltd., 398(1), pp.

135–44. doi: 10.1042/BJ20060313.

Caldecott, K. W. (2002) 'Polynucleotide Kinase: A Versatile Molecule Makes a Clean Break', *Structure*. Cell Press, 10(9), pp. 1151–1152. doi: 10.1016/S0969-2126(02)00833-X.

Cao, M. *et al.* (2009) 'Chlamydomonas (Chlorophyceae) colony PCR.', *Protoplasma*, 235(1–4), pp. 107–10. doi: 10.1007/s00709-009-0036-9.

Carrigan, P. E., Ballar, P. and Tuzmen, S. (2011) 'Site-directed mutagenesis.', *Methods in molecular biology (Clifton, N.J.)*, 700, pp. 107–24. doi: 10.1007/978-1-61737-954-3_8.

Carson, M. *et al.* (2007) 'His-tag impact on structure', *Acta Crystallographica Section D Biological Crystallography*, 63(3), pp. 295–301. doi: 10.1107/S0907444906052024.

Cermak, T., Starker, C. G. and Voytas, D. F. (2014) 'Efficient design and assembly of custom talens using the golden gate platform', in *Chromosomal Mutagenesis: Second Edition*. Humana Press, New York, NY, pp. 133–159. doi: 10.1007/978-1-4939-1862-1_7.

Chaconas, G. and van de Sande, J. H. (1980) '[10] 5'-32P labeling of RNA and DNA restriction fragments', *Methods in Enzymology*. Academic Press, 65, pp. 75–85. doi: 10.1016/S0076-6879(80)65012-5.

Chargaff, E. (1951) 'Structure and function of nucleic acids as cell constituents.', *Federation proceedings*, 10(3), pp. 654–659. Available at: http://www.ncbi.nlm.nih.gov/pubmed/14887699 (Accessed: 22 May 2018).

Chargraff, E. *et al.* (1951) 'The composition of the deoxyribonucleic acid of salmon sperm.', *The Journal of biological chemistry*, 192(1), pp. 223–230. Available at: http://www.ncbi.nlm.nih.gov/pubmed/14917668 (Accessed: 15 February 2019).

Chauleau, M. and Shuman, S. (2013) 'Kinetic mechanism of nick sealing by T4 RNA ligase 2 and effects of 3'-OH base mispairs and damaged base lesions.', *RNA (New York, N.Y.)*, 19(12), pp. 1840–7. doi: 10.1261/rna.041731.113.

Check, E. (2005) 'Designs on life', *Nature*, pp. 417–418. doi: 10.1038/438417a.

Cheng, C. and Shuman, S. (1997) 'Characterization of an ATP-dependent DNA ligase encoded by Haemophilus influenzae', *Nucleic Acids Research*. Oxford University Press, 25(7), pp. 1369–1374. doi: 10.1093/nar/25.7.1369.

Cherepanov, A. V. and De Vries, S. (2003) 'Kinetics and thermodynamics of nick sealing by T4 DNA ligase', *European Journal of Biochemistry*. Wiley/Blackwell (10.1111), 270(21), pp. 4315–4325. doi: 10.1046/j.1432-1033.2003.03824.x.

Chi, S. W. *et al.* (2009) 'Argonaute HITS-CLIP decodes microRNA-mRNA interaction maps.', *Nature*. NIH Public Access, 460(7254), pp. 479–86. doi: 10.1038/nature08170.

Cole, S. T. *et al.* (1998) 'Deciphering the biology of Mycobacterium tuberculosis from the complete genome sequence', *Nature*, 393(6685), pp. 537–544. doi: 10.1038/31159.

Corsi, P. (2015) 'Jean-Baptiste Lamarck', in *eLS*. Chichester, UK: John Wiley & Sons, Ltd, pp. 1–3. doi: 10.1002/9780470015902.a0002468.pub2.

Crick, F. (1970) 'Central dogma of molecular biology.', *Nature*. Nature Publishing Group, 227(5258), pp. 561–3. doi: 10.1038/227561a0.

Cummings, P. J. *et al.* (2013) 'Pyrosequencing for microbial identification and characterization.', *Journal of visualized experiments : JoVE*. MyJoVE Corporation, (78), p. e50405. doi: 10.3791/50405.

Dahm, R. (2008) 'Discovering DNA: Friedrich Miescher and the early years of nucleic acid research', *Human Genetics*. Springer-Verlag, 122(6), pp. 565–581. doi: 10.1007/s00439-007-0433-0.

Darnell, R. B. (2010) 'HITS-CLIP: panoramic views of protein-RNA regulation in living cells', *Wiley Interdisciplinary Reviews: RNA*. Wiley-Blackwell, 1(2), pp. 266–286. doi: 10.1002/wrna.31.

Deamer, D., Akeson, M. and Branton, D. (2016) 'Three decades of nanopore sequencing', *Nature Biotechnology*. Nature Publishing Group, pp. 518–524. doi: 10.1038/nbt.3423.

Della, M. *et al.* (2004) 'Mycobacterial Ku and ligase proteins constitute a two-component NHEJ repair machine', *Science*, 306(5696), pp. 683–685. doi: 10.1126/science.1099824.

Doherty, A. J. (1999) 'Conversion of a DNA ligase into an RNA capping enzyme', *Nucleic Acids Research*, 27(16), pp. 3253–3258. doi: 10.1093/nar/27.16.3253.

Drabløs, F. *et al.* (2004) 'Alkylation damage in DNA and RNA - Repair mechanisms and medical significance', *DNA Repair*. Elsevier, pp. 1389–1407. doi: 10.1016/j.dnarep.2004.05.004.

Endy, D. (2005) 'Foundations for engineering biology', *Nature*, pp. 449–453. doi: 10.1038/nature04342.

Engler, C. *et al.* (2009) 'Golden gate shuffling: a one-pot DNA shuffling method based on type IIs restriction enzymes.', *PloS one*, 4(5), p. e5553. doi: 10.1371/journal.pone.0005553.

Engler, C., Kandzia, R. and Marillonnet, S. (2008) 'A one pot, one step, precision cloning method with high throughput capability.', *PloS one*, 3(11), p. e3647. doi: 10.1371/journal.pone.0003647.

Fiers, W. *et al.* (1976) 'Complete nucleotide sequence of bacteriophage MS2 RNA: Primary and secondary structure of the replicase gene', *Nature*. Nature Publishing Group, 260(5551), pp. 500–507. doi: 10.1038/260500a0.

Finn, R. D. *et al.* (2014) 'Pfam: The protein families database', *Nucleic Acids Research*. Oxford University Press, pp. D222-30. doi: 10.1093/nar/gkt1223.

Franklin, R. E. and Gosling, R. G. (1953) 'The structure of sodium thymonucleate fibres. I. The influence of water content', *Acta Crystallographica*, 6(8), pp. 673–677. doi: 10.1107/S0365110X53001939.

Galburt, E. A. *et al.* (2002) 'Structure of a tRNA repair enzyme and molecular biology workhorse: T4 polynucleotide kinase.', *Structure (London, England : 1993)*, 10(9), pp. 1249–60. Available at: http://www.ncbi.nlm.nih.gov/pubmed/12220496 (Accessed: 19 April 2018).

Gao, S. *et al.* (2015) 'A Simple and Convenient Sticky/Blunt-End Ligation Method for Fusion Gene Construction', *Biochemical Genetics*. Springer US, 53(1–3), pp. 42–48. doi: 10.1007/s10528-015-9669-x.

Gellert, M. (1967) 'Formation of covalent circles of lambda DNA by E. coli extracts', *Proceedings of the National Academy of Sciences*, 57(1).

George, R. A. and Heringa, J. (2003) *An analysis of protein domain linkers: their classification and role in protein folding*, *Protein Engineering*. Available at: http://mathbio.nimr.mrc.ac.uk. (Accessed: 15 February 2019).

Saint Girons, I. *et al.* (1987) 'Structural and catalytic characteristics of Escherichia coli adenylate kinase.', *The Journal of biological chemistry*, 262(2), pp. 622–9. Available at: http://www.ncbi.nlm.nih.gov/pubmed/3027060 (Accessed: 7 February 2018).

Git, A. *et al.* (2010) 'Systematic comparison of microarray profiling, real-time PCR, and next-generation sequencing technologies for measuring differential microRNA expression.', *RNA (New York, N.Y.)*, 16(5), pp. 991–1006. doi: 10.1261/rna.1947110.

Glaser, M., Nulty, W. and Vagelos, P. R. (1975) 'Role of adenylate kinase in the regulation of macromolecular biosynthesis in a putative mutant of Escherichia coli defective in membrane phospholipid biosynthesis.', *J. Bacteriol.* American Society for Microbiology, 123(1), pp. 128–136. Available at: http://www.ncbi.nlm.nih.gov/pubmed/166976 (Accessed: 11 May 2016).

Golovanov, A. P. *et al.* (2004) 'A simple method for improving protein solubility and long-term stability', *Journal of the American Chemical Society*. American Chemical Society, 126(29), pp. 8933–8939. doi: 10.1021/ja049297h.

Gong, C. *et al.* (2004) 'Biochemical and genetic analysis of the four DNA ligases of mycobacteria.', *The Journal of biological chemistry*. American Society for Biochemistry and Molecular Biology, 279(20), pp. 20594–606. doi: 10.1074/jbc.M401841200.

Gong, C. *et al.* (2005) 'Mechanism of nonhomologous end-joining in mycobacteria: A low-fidelity repair system driven by Ku, ligase D and ligase C', *Nature Structural and Molecular Biology*. Nature Publishing Group, 12(4), pp. 304–312. doi: 10.1038/nsmb915.

Goodsell, D. S. and Olson, A. J. (2000) 'Structural Symmetry and Protein Function', *Annual Review of Biophysics and Biomolecular Structure*. Annual Reviews 4139 El Camino Way, P.O. Box 10139, Palo Alto, CA 94303-0139, USA , 29(1), pp. 105–153. doi: 10.1146/annurev.biophys.29.1.105.

Grass, R. N. *et al.* (2015) 'Robust chemical preservation of digital information on DNA in silica with error-correcting codes', *Angewandte Chemie - International Edition*. Wiley-Blackwell, 54(8), pp. 2552–2555. doi: 10.1002/anie.201411378.

Gupta, R. *et al.* (2011) 'Mycobacteria exploit three genetically distinct DNA double-strand break repair pathways', *Molecular Microbiology*. Wiley/Blackwell (10.1111), 79(2), pp. 316–330. doi: 10.1111/j.1365-2958.2010.07463.x.

Hafner, M. *et al.* (2008) 'Identification of microRNAs and other small regulatory RNAs using cDNA library sequencing', *Methods*. NIH Public Access, 44(1), pp. 3–12. doi: 10.1016/j.ymeth.2007.09.009.

Hafner, M. *et al.* (2011) 'RNA-ligase-dependent biases in miRNA representation in deep-sequenced small RNA cDNA libraries.', *RNA (New York, N.Y.)*. Cold Spring Harbor Laboratory Press, 17(9), pp. 1697–712. doi: 10.1261/rna.2799511.

Hall, D. H. *et al.* (1980) 'Suppressors of mutations in the bacteriophage T4 gene coding for both RNA ligase and tail fiber attachment activities', *J Virol*, 36(1), pp. 103–108. Available at: http://www.ncbi.nlm.nih.gov/pubmed/7441817 (Accessed: 24 May 2018).

Harrison, B. and Zimmerman, S. B. (1986) 'Stabilization of T4 polynucleotide kinase by macromolecular crowding', *Nucleic Acids Research*. Oxford University Press, 14(4), pp. 1863–1870. doi: 10.1093/nar/14.4.1863.

Hayward, S. (2004) 'Identification of Specific Interactions that Drive Ligand-induced Closure in Five Enzymes with Classic Domain Movements', *Journal of Molecular Biology*, 339(4), pp. 1001–1021. doi: 10.1016/j.jmb.2004.04.004.

Head, S. R. *et al.* (2014) 'Library construction for next-generation sequencing: overviews and challenges.', *BioTechniques*. NIH Public Access, 56(2), p. 61–4, 66, 68, passim. doi:

10.2144/000114133.

Heather, J. M. and Chain, B. (2016) 'The sequence of sequencers: The history of sequencing DNA', *Genomics*. Academic Press, pp. 1–8. doi: 10.1016/j.ygeno.2015.11.003.

Hefferin, M. L. and Tomkinson, A. E. (2005) 'Mechanism of DNA double-strand break repair by non-homologous end joining', *DNA Repair*. Elsevier, 4(6), pp. 639–648. doi: 10.1016/J.DNAREP.2004.12.005.

Hershey, A. D. (1952) 'Independent Functions of Viral Protein and Nucleic Acid in Growth of Bacteriophage', *The Journal of General Physiology*, 36(1), pp. 39–56. doi: 10.1085/jgp.36.1.39.

Hill, J. M. and Lukiw, W. J. (2016) 'microRNA (miRNA)-Mediated Pathogenetic Signaling in Alzheimer's Disease (AD)', *Neurochemical Research*. Springer US, 41(1–2), pp. 96–100. doi: 10.1007/s11064-015-1734-7.

Ho, C. K. *et al.* (2004) 'Structure and mechanism of RNA ligase', *Structure*, 12(2), pp. 327–339. doi: 10.1016/j.str.2004.01.011.

Ho, C. K. and Shuman, S. (2002) 'Bacteriophage T4 RNA ligase 2 (gp24.1) exemplifies a family of RNA ligases found in all phylogenetic domains.', *Proceedings of the National Academy of Sciences of the United States of America*, 99, pp. 12709–12714. doi: 10.1073/pnas.192184699.

Hochuli, E. *et al.* (1988) 'Genetic Approach to Facilitate Purification of Recombinant Proteins with a Novel Metal Chelate Adsorbent', *Nature Biotechnology*. Nature Publishing Group, 6(11), pp. 1321–1325. doi: 10.1038/nbt1188-1321.

Holley, R. W. *et al.* (1961) 'Nucleotide and oligonucleotide compositions of the alanine-, valine-, and tyrosine-acceptor "soluble" ribonucleic acids of yeast', *Journal of the American Chemical Society*. American Chemical Society, pp. 4861–4862. doi: 10.1021/ja01484a040.

Holliger, P. and Hudson, P. J. (2005) 'Engineered antibody fragments and the rise of single domains', *Nature Biotechnology*, 23(9), pp. 1126–1136. doi: 10.1038/nbt1142.

Houtgraaf, J. H., Versmissen, J. and van der Giessen, W. J. (2006) 'A concise review of DNA damage checkpoints and repair in mammalian cells', *Cardiovascular Revascularization Medicine*. Elsevier, 7(3), pp. 165–172. doi: 10.1016/j.carrev.2006.02.002.

Huang, Y. *et al.* (2013) 'A gold nanoparticle-enhanced fluorescence polarization biosensor for amplified detection of T4 polynucleotide kinase activity and inhibition', *Journal of Materials Chemistry B*. The Royal Society of Chemistry, 1(15), p. 2018. doi: 10.1039/c3tb00025g.

Hulett, H. R. (1970) 'Non-enzymatic Hydrolysis of Adenosine Phosphates', *Nature*. Nature

Publishing Group, 225(5239), pp. 1248–1249. doi: 10.1038/2251248a0.

Hustedt, N. and Durocher, D. (2017) 'The control of DNA repair by the cell cycle', *Nature Cell Biology*. Nature Publishing Group, 19(1), pp. 1–9. doi: 10.1038/ncb3452.

Hutchison, C. A. (2007) 'DNA sequencing: Bench to bedside and beyond', *Nucleic Acids Research*. Oxford University Press, 35(18), pp. 6227–6237. doi: 10.1093/nar/gkm688.

Hyman, E. D. (1988) 'A new method of sequencing DNA', *Analytical Biochemistry*. Academic Press, 174(2), pp. 423–436. doi: 10.1016/0003-2697(88)90041-3.

Jackson, S. P. and Bartek, J. (2009) 'The DNA-damage response in human biology and disease', *Nature*. Nature Publishing Group, pp. 1071–1078. doi: 10.1038/nature08467.

Jayaprakash, A. D. *et al.* (2011) 'Identification and remediation of biases in the activity of RNA ligases in small-RNA deep sequencing', *Nucleic Acids Res*, 39(21), p. e141. doi: 10.1093/nar/gkr693.

Jou, W. M. *et al.* (1972) 'Nucleotide sequence of the gene coding for the bacteriophage MS2 coat protein', *Nature*. Nature Publishing Group, 237(5350), pp. 82–88. doi: 10.1038/237082a0.

Karam, J. D. and Barker, B. (1971) 'Properties of bacteriophage T4 mutants defective in gene 30 (deoxyribonucleic acid ligase) and the rII gene.', *Journal of virology*, 7(2), pp. 260–6. Available at: http://www.ncbi.nlm.nih.gov/pubmed/4939059 (Accessed: 24 May 2018).

Kelly, T. J. and Smith, H. O. (1970) 'A restriction enzyme from Hemophilus influenzae: II. Base sequence of the recognition site', *Journal of Molecular Biology*. Academic Press, 51(2), pp. 393–409. doi: 10.1016/0022-2836(70)90150-6.

Kendrew, J. C. *et al.* (1958) 'A three-dimensional model of the myoglobin molecule obtained by x-ray analysis', *Nature*, 181, pp. 662–666.

Khalil, A. S. and Collins, J. J. (2010) 'Synthetic biology: Applications come of age', *Nature Reviews Genetics*. NIH Public Access, pp. 367–379. doi: 10.1038/nrg2775.

Kimple, M. E. and Sondek, J. (2004) 'Overview of Affinity Tags for Protein Purification', in *Current Protocols in Protein Science*. NIH Public Access, p. 9.9.1-9.9.19. doi: 10.1002/0471140864.ps0909s36.

Kimpton, J. and Emerman, M. (1992) 'Detection of replication-competent and pseudotyped human immunodeficiency virus with a sensitive cell line on the basis of activation of an integrated beta-galactosidase gene.', *Journal of virology*. American Society for Microbiology, 66(4), pp. 2232–9. Available at: http://www.ncbi.nlm.nih.gov/pubmed/1548759 (Accessed: 28

March 2018).

Klein, J. S. *et al.* (2014) 'Design and characterization of structured protein linkers with differing flexibilities.', *Protein engineering, design & selection : PEDS*. Oxford University Press, 27(10), pp. 325–30. doi: 10.1093/protein/gzu043.

Klug, A. (1968) 'Rosalind Franklin and the discovery of the structure of DNA', *Nature*. Nature Publishing Group, 219(5156), pp. 808–844. doi: 10.1038/219808a0.

Knight, T. (2003) 'Idempotent Vector Design for Standard Assembly of Biobricks'. Available at: http://oai.dtic.mil/oai/oai?verb=getRecord&metadataPrefix=html&identifier=ADA457791 (Accessed: 3 March 2015).

Kodama, K., Barnes, D. E. and Lindahl, T. (1991) 'In vitro mutagenesis and functional expression in Escherichia coli of a cDNA encoding the catalytic domain of human DNA ligase I.', *Nucleic acids research*. Oxford University Press, 19(22), pp. 6093–6099. doi: 10.1093/nar/19.22.6093.

Kohler, R. E. (1973) 'The Enzyme Theory and the Origin of Biochemistry', *Isis*. History of Science Society, 64(2), pp. 181–196. doi: 10.1086/351080.

Korycka-Machala, M. *et al.* (2006) 'Distinct DNA repair pathways involving RecA and nonhomologous end joining in Mycobacterium smegmatis', *FEMS Microbiology Letters*. Oxford University Press, 258(1), pp. 83–91. doi: 10.1111/j.1574-6968.2006.00199.x.

Korycka-Machala, M. *et al.* (2007) 'Evaluation of NAD(+) -dependent DNA ligase of mycobacteria as a potential target for antibiotics.', *Antimicrobial agents and chemotherapy*. American Society for Microbiology, 51(8), pp. 2888–97. doi: 10.1128/AAC.00254-07.

Kossel, A. (1910) 'Über das Agmatin', *Hoppe-Seyler's Zeitschrift fur Physiologische Chemie*. Walter de Gruyter, Berlin / New York, 66(3), pp. 257–261. doi: 10.1515/bchm2.1910.66.3.257.

Krainer, F. W. and Glieder, A. (2015) 'An updated view on horseradish peroxidases: recombinant production and biotechnological applications.', *Applied microbiology and biotechnology*. Springer, 99(4), pp. 1611–25. doi: 10.1007/s00253-014-6346-7.

Kumar, M. S. *et al.* (2007) 'Impaired microRNA processing enhances cellular transformation and tumorigenesis', *Nature Genetics*. Nature Publishing Group, 39(5), pp. 673–677. doi: 10.1038/ng2003.

Kunkel, T. A. (1985) 'Rapid and efficient site-specific mutagenesis without phenotypic selection.', *Proceedings of the National Academy of Sciences*, 82(2), pp. 488–492. doi: 10.1073/pnas.82.2.488.

Leiman, P. G. *et al.* (2003) 'Structure and morphogenesis of bacteriophage T4', *Cellular and Molecular Life Sciences*, pp. 2356–2370. doi: 10.1007/s00018-003-3072-1.

Leiman, P. G. *et al.* (2010) 'Morphogenesis of the T4 tail and tail fibers', *Virology Journal*. BioMed Central, 7(1), p. 355. doi: 10.1186/1743-422X-7-355.

Lillehaug, J. R. and Kleppe, K. (1975) 'Effect of salts and polyamines on T $_4$ polynucleotide kinase', *Biochemistry*. American Chemical Society, 14(6), pp. 1225–1229. doi: 10.1021/bi00677a021.

Lin, L. *et al.* (2011) 'Sensitive and rapid screening of T4 polynucleotide kinase activity and inhibition based on coupled exonuclease reaction and graphene oxide platform', *Analytical Chemistry*. American Chemical Society, 83(22), pp. 8396–8402. doi: 10.1021/ac200593g.

Linnaeus, C. (1758) *Systema Naturae, Vol. 1*, *Systema naturae, Vol. 1*. Available at: https://www.cabdirect.org/cabdirect/abstract/20057000018 (Accessed: 22 May 2018).

Liu, X. *et al.* (2015) 'Use of adenylate kinase as a solubility tag for high level expression of T4 DNA ligase in Escherichia coli.', *Protein expression and purification*, 109, pp. 79–84. doi: 10.1016/j.pep.2015.02.010.

Lowder, M. A. *et al.* (2011) 'Visualizing protein partnerships in living cells and organisms.', *Current opinion in chemical biology*. NIH Public Access, 15(6), pp. 781–8. doi: 10.1016/j.cbpa.2011.10.024.

Lu, M. *et al.* (2008) 'An Analysis of Human MicroRNA and Disease Associations', *PLoS ONE*. Edited by M. Isalan. Public Library of Science, 3(10), p. e3420. doi: 10.1371/journal.pone.0003420.

Luo, J. and Barany, F. (1996) 'Identification of essential residues in Thermus thermophilus DNA ligase', *Nucleic Acids Research*. Oxford University Press, 24(15), pp. 3079–85. doi: 10.1093/nar/24.15.3079.

Luria, S. E. and Human, M. L. (1952) 'A nonhereditary, host-induced variation of bacterial viruses.', *Journal of bacteriology*. American Society for Microbiology (ASM), 64(4), pp. 557–569. doi: 10.1016/S0022-3476(76)80394-0.

Mao, Z. *et al.* (2008) 'Comparison of nonhomologous end joining and homologous recombination in human cells.', *DNA repair*. NIH Public Access, 7(10), pp. 1765–71. doi: 10.1016/j.dnarep.2008.06.018.

Martin, H., Wani, S. and Steptoe, A. (2014) 'Imperfect centered miRNA binding sites are common and can mediate repression of target mRNAs', *Genome biology*, 13(3).

Mason, P. A. and Cox, L. S. (2012) 'The role of DNA exonucleases in protecting genome stability

and their impact on ageing.', *Age (Dordrecht, Netherlands)*. Springer, 34(6), pp. 1317–40. doi: 10.1007/s11357-011-9306-5.

Matthews, L. A. and Simmons, L. A. (2014) 'Bacterial nonhomologous end joining requires teamwork', *Journal of Bacteriology*. American Society for Microbiology, pp. 3363–3365. doi: 10.1128/JB.02042-14.

Maxam, A. M. and Gilbert, W. (1977) 'A new method for sequencing DNA.', *Proceedings of the National Academy of Sciences*, 74(2), pp. 560–564. doi: 10.1073/pnas.74.2.560.

McCarty, M. and Avery, O. T. (1946) 'Studies on the chemical nature of the substance inducing transformation of pneumococcal types. III an improved method for the isolation of the transforming substance and its application to pneumococcus types II, III, and VI', *The Journal of experimental medicine*. Rockefeller University Press, 83(2), pp. 97–104. doi: 10.1084/jem.83.2.97.

Mendes, N. D., Freitas, A. T. and Sagot, M.-F. (2009) 'Current tools for the identification of miRNA genes and their targets', *Nucleic Acids Research*. Oxford University Press, 37(8), pp. 2419–2433. doi: 10.1093/nar/gkp145.

Merriman, B., R&D Team, I. T. and Rothberg, J. M. (2012) 'Progress in Ion Torrent semiconductor chip based sequencing', *ELECTROPHORESIS*. Wiley-Blackwell, 33(23), pp. 3397–3417. doi: 10.1002/elps.201200424.

Metzenberg, S. (2002) *Site-Directed Mutagenesis*. Available at: http://www.escience.ws/b572/L4/L4.htm (Accessed: 21 April 2015).

Michelson-Horowitz, D. J. *et al.* (2005) 'Comparison of affinity tags for protein purification', *Protein Expression and Purification*. Academic Press, 41(1), pp. 98–105. doi: 10.1016/j.pep.2005.01.019.

Miller *et al.* (2003) 'Bacteriophage T4 Genome', *Microbiology and Molecular Biology Reviews*, 67(1), pp. 86–156. doi: 10.1128/MMBR.67.1.86-156.2003.

Mitchell, P. S. *et al.* (2008) 'Circulating microRNAs as stable blood-based markers for cancer detection.', *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, 105(30), pp. 10513–8. doi: 10.1073/pnas.0804549105.

Motameny, S. *et al.* (2010) 'Next Generation Sequencing of miRNAs – Strategies, Resources and Methods', *Genes*. Molecular Diversity Preservation International, 1(1), pp. 70–84. doi: 10.3390/genes1010070.

Mukherjee, D. *et al.* (2004) 'Analysis of RNA exonucleolytic activities in cellular extracts.', *Methods in molecular biology (Clifton, N.J.)*. New Jersey: Humana Press, 257, pp. 193–212. doi: 10.1385/1-59259-750-5:193.

Müller, C. W. *et al.* (1996) 'Adenylate kinase motions during catalysis: An energetic counterweight balancing substrate binding', *Structure*. Cell Press, 4(2), pp. 147–156. doi: 10.1016/S0969-2126(96)00018-4.

Müller, C. W. and Schulz, G. E. (1992) 'Structure of the complex between adenylate kinase from Escherichia coli and the inhibitor Ap5A refined at 1.9 Å resolution. A model for a catalytic transition state', *Journal of Molecular Biology*. Academic Press, 224(1), pp. 159–177. doi: 10.1016/0022-2836(92)90582-5.

Nabavi, S. and Nazar, R. N. (2005) 'Simplified one-tube "megaprimer" polymerase chain reaction mutagenesis', *Analytical Biochemistry*, 345(2), pp. 346–348. doi: 10.1016/j.ab.2005.06.033.

Nandakumar, J. *et al.* (2004) 'RNA substrate specificity and structure-guided mutational analysis of bacteriophage T4 RNA ligase 2', *Journal of Biological Chemistry*, 279, pp. 31337–31347. doi: 10.1074/jbc.M402394200.

Nandakumar, J. and Shuman, S. (2004) 'How an RNA ligase discriminates RNA versus DNA damage', *Molecular Cell*, 16(2), pp. 211–221. doi: 10.1016/j.molcel.2004.09.022.

Nandakumar, J., Shuman, S. and Lima, C. D. (2006) 'RNA Ligase Structures Reveal the Basis for RNA Specificity and Conformational Changes that Drive Ligation Forward', *Cell*, 127(1), pp. 71–84. doi: 10.1016/j.cell.2006.08.038.

Nordlund, P. *et al.* (2008) 'Protein production and purification', *Nature Methods*. Nature Publishing Group, 5(2), pp. 135–146. doi: 10.1038/nmeth.f.202 [doi].

Olby, R. C. (Robert C. (1994) *The path to the double helix : the discovery of DNA*. Available at: https://books.google.co.uk/books?hl=en&lr=&id=s_UmoMXRTIYC&oi=fnd&pg=PP1&dq=dna+discovery+1943&ots=XO-eKR6Md3&sig=oYiulfl_t4UVpsXUBknJN0XkFUY#v=onepage&q=dna discovery 1943&f=false (Accessed: 22 May 2018).

El Omari, K. *et al.* (2006) 'Molecular architecture and ligand recognition determinants for T4 RNA ligase', *Journal of Biological Chemistry*. American Society for Biochemistry and Molecular Biology, 281(3), pp. 1573–1579. doi: 10.1074/jbc.M509658200.

Padmanabhan, S., Banerjee, S. and Mandi, N. (2011) 'Screening of Bacterial Recombinants: Strategies and Preventing False Positives', in *Molecular Cloning - Selected Applications in Medicine and Biology*. InTech. doi: 10.5772/22140.

Pascal, J. M. *et al.* (2004) 'Human DNA ligase I completely encircles and partially unwinds nicked DNA.', *Nature*, 432, pp. 473–478. doi: 10.1038/nature03082.

Pascal, J. M. (2008) 'DNA and RNA ligases: structural variations and shared mechanisms', *Current Opinion in Structural Biology*. Elsevier Current Trends, 18(1), pp. 96–105. doi: 10.1016/J.SBI.2007.12.008.

Patron, N. J. *et al.* (2015) 'Standards for plant synthetic biology: A common syntax for exchange of DNA parts', *New Phytologist*. Wiley/Blackwell (10.1111), 208(1), pp. 13–19. doi: 10.1111/nph.13532.

Penner, M. *et al.* (1995) 'Phage T4-coded Stp: Double-edged effector of coupled DNA and tRNA-restriction systems', *Journal of Molecular Biology*. Academic Press, 249(5), pp. 857–868. doi: 10.1006/jmbi.1995.0343.

Pergolizzi, G., Wagner, G. K. and Bowater, R. P. (2016) 'Biochemical and Structural Characterisation of DNA Ligases from Bacteria and Archaea.', *Bioscience reports*. Portland Press Ltd, 36(5). doi: 10.1042/BSR20160003.

Perutz, M. (1995) 'Hoppe-Seyler, Stokes and haemoglobin.', *Biological chemistry Hoppe-Seyler*, 376(8), pp. 449–50. Available at: http://www.ncbi.nlm.nih.gov/pubmed/7576244 (Accessed: 22 May 2018).

Petsev, D. N. *et al.* (2000) 'Interactions and aggregation of apoferritin molecules in solution: Effects of added electrolytes', *Biophysical Journal*. Cell Press, 78(4), pp. 2060–2069. doi: 10.1016/S0006-3495(00)76753-1.

Pitcher, R. S. *et al.* (2007) 'Structure and Function of a Mycobacterial NHEJ DNA Repair Polymerase', *Journal of Molecular Biology*. Academic Press, 366(2), pp. 391–405. doi: 10.1016/j.jmb.2006.10.046.

Pitcher, R. S., Brissett, N. C. and Doherty, A. J. (2007) 'Nonhomologous End-Joining in Bacteria: A Microbial Perspective', *Annual Review of Microbiology*. Annual Reviews, 61(1), pp. 259–282. doi: 10.1146/annurev.micro.61.080706.093354.

Płociński, P. *et al.* (2017) 'DNA Ligase C and Prim-PolC participate in base excision repair in mycobacteria', *Nature Communications*. Nature Publishing Group, 8(1), p. 1251. doi: 10.1038/s41467-017-01365-y.

Pogue, A. I. *et al.* (2014) 'Evolution of microRNA (miRNA) Structure and Function in Plants and Animals: Relevance to Aging and Disease.', *Journal of aging science*. NIH Public Access, 2(2). doi: 10.4172/2329-8847.1000119.

Poldolsky, R. and Morales, M. (1956) 'The Enthalpy Change of Adenosine Triphospate Hydrolysis', *J. Biol. Chem.*, 218, pp. 945–959.

Pollister, A. W. and Mirsky, A. E. (1943) 'Terminology of Nucleic Acids', *Nature*. Nature Publishing Group, 152(3867), pp. 692–693. doi: 10.1038/152692b0.

Popow, J., Schleiffer, A. and Martinez, J. (2012) 'Diversity and roles of (t)RNA ligases', *Cell Mol Life Sci*, 69(16), pp. 2657–2670. doi: 10.1007/s00018-012-0944-2.

Porath, J. *et al.* (1975) 'Metal chelate affinity chromatography, a new approach to protein fractionation', *Nature*. Nature Publishing Group, 258(5536), pp. 598–599. doi: 10.1038/258598a0.

Pray, L. (2008) 'Discovery of DNA Double Helix: Watson and Crick | Learn Science at Scitable', *Nature Education*. Available at: https://www.nature.com/scitable/topicpage/discovery-of-dna-structure-and-function-watson-397 (Accessed: 15 February 2019).

Qing, T. *et al.* (2017) 'Dumbbell DNA-templated CuNPs as a nano-fluorescent probe for detection of enzymes involved in ligase-mediated DNA repair', *Biosensors and Bioelectronics*, 94, pp. 456–463. doi: 10.1016/j.bios.2017.03.035.

Reva, B., Antipin, Y. and Sander, C. (2011) 'Predicting the functional impact of protein mutations: Application to cancer genomics', *Nucleic Acids Research*. Oxford University Press, 39(17), pp. e118–e118. doi: 10.1093/nar/gkr407.

Richardson, C. C. (1965) 'Phosphorylation of nucleic acid by an enzyme from T4 bacteriophage-infected Escherichia coli', *Proceedings of the National Academy of Sciences*, 54(1).

Roberts, R. J. *et al.* (2003) 'A nomenclature for restriction enzymes, DNA methyltransferases, homing endonucleases and their genes', *Nucleic Acids Research*. Oxford University Press, 31(7), pp. 1805–1812. doi: 10.1093/nar/gkg274.

Rötzschke, O. *et al.* (2002) 'A pH-sensitive histidine residue as control element for ligand release from HLA-DR molecules.', *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, 99(26), pp. 16946–50. doi: 10.1073/pnas.212643999.

Sanger, F. *et al.* (1977) 'Nucleotide sequence of bacteriophage phi X174 DNA.', *Nature*. Nature Publishing Group, 265(5596), pp. 687–695. doi: 5,386 base pairs...

Sanger, F., Brownlee, G. G. and Barrell, B. G. (1965) 'A two-dimensional fractionation procedure for radioactive nucleotides', *Journal of Molecular Biology*. Academic Press, 13(2), pp. IN1-IN4. doi: 10.1016/S0022-2836(65)80104-8.

Sanger, F. and Coulson, A. R. (1975) 'A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase', *Journal of Molecular Biology*. Academic Press, 94(3), pp. 441–448. doi: 10.1016/0022-2836(75)90213-2.

Sanger, F., Nicklen, S. and Coulson, A. R. (1977) 'DNA sequencing with chain-terminating inhibitors', *Proceedings of the National Academy of Sciences*. National Academy of Sciences, 74(12), pp. 5463–5467. doi: 10.1073/pnas.74.12.5463.

Schlauderer, G. J., Proba, K. and Schulz, G. E. (1996) 'Structure of a Mutant Adenylate Kinase Ligated with an ATP-analogue Showing Domain Closure Over ATP', *Journal of Molecular Biology*. Academic Press, 256(2), pp. 223–227. doi: 10.1006/JMBI.1996.0080.

Schwer, B. *et al.* (2004) 'Portability and fidelity of RNA-repair systems', *Proceedings of the National Academy of Sciences*, 101(9), pp. 2788–2793. doi: 10.1073/pnas.0305859101.

Serrano, L. (2007) 'Synthetic biology: promises and challenges.', *Molecular systems biology*, 3, p. 158. doi: 10.1038/msb4100202.

Shermer, M. (2002) *In Darwin's shadow : the life and science of Alfred Russel Wallace : a biographical study on the psychology of history*. Oxford University Press. Available at: https://books.google.co.uk/books?hl=en&lr=&id=7GnnCwAAQBAJ&oi=fnd&pg=PR7&dq=alfred+wallace+evolution+darwin&ots=3W41q4dT_8&sig=Dyw31800u5zUPnfM6wSXKIfLzsE#v=onepage&q=alfred wallace evolution darwin&f=false (Accessed: 22 May 2018).

Shetty, R. P., Endy, D. and Knight, T. F. (2008) 'Engineering BioBrick vectors from BioBrick parts.', *Journal of biological engineering*, 2, p. 5. doi: 10.1186/1754-1611-2-5.

Shi, K. *et al.* (2018) 'T4 DNA ligase structure reveals a prototypical ATP-dependent ligase with a unique mode of sliding clamp interaction', *Nucleic Acids Research*, 46(19), pp. 10474–10488. doi: 10.1093/nar/gky776.

Shuman, S. (2009) 'DNA ligases: Progress and prospects', *Journal of Biological Chemistry*, pp. 17365–17369. doi: 10.1074/jbc.R900017200.

Shuman, S. and Glickman, M. S. (2007) 'Bacterial DNA repair by non-homologous end joining', *Nature Reviews Microbiology*. Nature Publishing Group, 5(11), pp. 852–861. doi: 10.1038/nrmicro1768.

Shuman, S. and Lima, C. D. (2004) 'The polynucleotide ligase and RNA capping enzyme superfamily of covalent nucleotidyltransferases', *Curr Opin Struct Biol*, 14(6), pp. 757–764. doi: 10.1016/j.sbi.2004.10.006.

Shuman, S. and Schwer, B. (1995) *RNA capping enzyme and DNA ligase: a superfamily of covalent nucleotidyl transferases*, *Molecular Microbiology*. Wiley/Blackwell (10.1111). doi: 10.1111/j.1365-2958.1995.mmi_17030405.x.

Silber, R., Malathi, V. G. and Hurwitz, J. (1972) 'Purification and Properties of Bacteriophage T4-Induced RNA Ligase', *Proceedings of the National Academy of Sciences*, 69(10). Available at: http://www.pnas.org/content/69/10/3009.short (Accessed: 29 March 2018).

Sinden, R. R. (1994) *DNA structure and function*. Academic Press.

Sirotkin, K. *et al.* (1978) 'A role in true-late gene expression for the T4 bacteriophage 5' polynucleotide kinase 3' phosphatase', *Journal of Molecular Biology*. Academic Press, 123(2), pp. 221–233. doi: 10.1016/0022-2836(78)90322-4.

Smith, D. B. (2000) 'Generating fusions to glutathione S-transferase for protein studies', in *Methods in Enzymology*. Academic Press, pp. 254–270. doi: 10.1016/S0076-6879(00)26059-X.

Smith, D. B. and Johnson, K. S. (1988) 'Single-step purification of polypeptides expressed in Escherichia coli as fusions with glutathione S-transferase', *Gene*. Elsevier, 67(1), pp. 31–40. doi: 10.1016/0378-1119(88)90005-4.

Smith, H. O. and Welcox, K. W. (1970) 'A Restriction enzyme from Hemophilus influenzae. I. Purification and general properties', *Journal of Molecular Biology*. Academic Press, 51(2), pp. 379–391. doi: 10.1016/0022-2836(70)90149-X.

Snopek, T. J. *et al.* (1977) 'Bacteriophage T4 RNA ligase is gene 63 product, the protein that promotes tail fiber attachment to the baseplate.', *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, 74(8), pp. 3355–3359. doi: 10.1073/pnas.74.8.3355.

Snow, C., Qi, G. and Hayward, S. (2007) 'Essential dynamics sampling study of adenylate kinase: comparison to citrate synthase and implication for the hinge and shear mechanisms of domain motions.', *Proteins*, 67(2), pp. 325–37. doi: 10.1002/prot.21280.

Song, C. *et al.* (2015) 'A sensitive detection of T4 polynucleotide kinase activity based on beta-cyclodextrin polymer enhanced fluorescence combined with an exonuclease reaction', *Chem Commun (Camb)*. Royal Society of Chemistry, 51(10), pp. 1815–1818. doi: 10.1039/c4cc08991j.

Song, C. and Zhao, M. (2009) 'Real-Time Monitoring of the Activity and Kinetics of T4 Polynucleotide Kinase by a Singly Labeled DNA-Hairpin Smart Probe Coupled with λ Exonuclease Cleavage', *Analytical Chemistry*. American Chemical Society, 81(4), pp. 1383–1388. doi: 10.1021/ac802107w.

Song, Y., Liu, K. J. and Wang, T.-H. (2014) 'Elimination of Ligation Dependent Artifacts in T4 RNA Ligase to Achieve High Efficiency and Low Bias MicroRNA Capture', *PLoS ONE*. Edited by B. Lin. Public Library of Science, 9(4), p. e94619. doi: 10.1371/journal.pone.0094619.

Sorefan, K. *et al.* (2012) 'Reducing ligation bias of small RNAs in libraries for next generation sequencing.', *Silence*, 3(1), p. 4. doi: 10.1186/1758-907X-3-4.

Sriskanda, V. (2002) 'Role of nucleotidyltransferase motifs I, III and IV in the catalysis of phosphodiester bond formation by Chlorella virus DNA ligase', *Nucleic Acids Research*, 30(4), pp. 903–911. doi: 10.1093/nar/30.4.903.

Sriskanda, V. and Shuman, S. (1998) 'Mutational analysis of Chlorella virus DNA ligase: catalytic roles of domain I and motif VI', *Nucleic Acids Research*. Oxford University Press, 26(20), pp. 4618–4625. doi: 10.1093/nar/26.20.4618.

Stanford, P. K. (2005) 'August Weismann's Theory of the Germ-Plasm and the Problem of Unconceived Alternatives', *History and philosophy of the life sciences*. Stazione Zoologica Anton Dohrn - Napoli, 27(2), pp. 163–199. doi: 10.2307/23333865.

Studier, F. W. and Moffatt, B. A. (1986) 'Use of bacteriophage T7 RNA polymerase to direct selective high-level expression of cloned genes', *Journal of Molecular Biology*. Academic Press, 189(1), pp. 113–130. doi: 10.1016/0022-2836(86)90385-2.

Swerdlow, H. and Gesteland, R. (1990) 'Capillary gel electrophoresis for rapid, high resolution DNA sequencing', *Nucleic Acids Research*. Oxford University Press, 18(6), pp. 1415–1419. doi: 10.1093/nar/18.6.1415.

Tanford, C. and Reynolds, J. A. (Jacqueline A. (2003) *Nature's robots : a history of proteins*. OUP Oxford. Available at: https://books.google.co.uk/books?hl=en&lr=&id=Vg5H7InfABEC&oi=fnd&pg=PT4&dq=1789:+Antoine+Fourcroy+protein+&ots=tC5pTsoaqB&sig=lv3bPxfA5srbpnUoX8QBP3UHqn8#v=onepage&q=1789%3A Antoine Fourcroy protein&f=false (Accessed: 22 May 2018).

Taylor, D. D. and Gercel-Taylor, C. (2008) 'MicroRNA signatures of tumor-derived exosomes as diagnostic biomarkers of ovarian cancer.', *Gynecologic oncology*. Elsevier, 110(1), pp. 13–21. doi: 10.1016/j.ygyno.2008.04.033.

Terpe, K. (2003) 'Overview of tag protein fusions: from molecular and biochemical fundamentals to commercial systems', *Applied Microbiology and Biotechnology*. Springer-Verlag, 60(5), pp. 523–533. doi: 10.1007/s00253-002-1158-6.

Thomson, D. W., Bracken, C. P. and Goodall, G. J. (2011) 'Experimental strategies for microRNA

target identification.', *Nucleic acids research*. Oxford University Press, 39(16), pp. 6845–53. doi: 10.1093/nar/gkr330.

Timson, D. J., Singleton, M. R. and Wigley, D. B. (2000) 'DNA ligases in the repair and replication of DNA', *Mutation Research - DNA Repair*. Elsevier, pp. 301–318. doi: 10.1016/S0921-8777(00)00033-1.

Tomkinson, A. E. and Levin, D. S. (1997) 'Mammalian DNA ligases', *BioEssays*. Wiley Subscription Services, Inc., A Wiley Company, 19(10), pp. 893–901. doi: 10.1002/bies.950191009.

Travers, A. and Muskhelishvili, G. (2015) 'DNA structure and function', *FEBS Journal*. John Wiley & Sons, Ltd (10.1111), 282(12), pp. 2279–2295. doi: 10.1111/febs.13307.

Tulay, P. and Sengupta, S. B. (2016) 'MicroRNA expression and its association with DNA repair in preimplantation embryos.', *The Journal of reproduction and development*. Japanese Society of Animal Reproduction, 62(3), pp. 225–34. doi: 10.1262/jrd.2015-167.

Unciuleac, M. C. and Shuman, S. (2013) 'Distinctive effects of domain deletions on the manganese-dependent DNA polymerase and DNA phosphorylase activities of Mycobacterium smegmatis polynucleotide phosphorylase', *Biochemistry*. American Chemical Society, 52(17), pp. 2967–2981. doi: 10.1021/bi400281w.

Uzan, M. and Miller, E. S. (2010) 'Post-transcriptional control by bacteriophage T4: MRNA decay and inhibition of translation initiation', *Virology Journal*. BioMed Central, p. 360. doi: 10.1186/1743-422X-7-360.

Vinnikov, I. A. and Domanskyi, A. (2017) 'Can we treat neurodegenerative diseases by preventing an age-related decline in microRNA expression?', *Neural regeneration research*. Wolters Kluwer -- Medknow Publications, 12(10), pp. 1602–1604. doi: 10.4103/1673-5374.217328.

Viollet, S. *et al.* (2011) 'T4 RNA ligase 2 truncated active site mutants: improved tools for RNA analysis', *BMC Biotechnol*, 11, p. 72. doi: 10.1186/1472-6750-11-72.

Wang, L. K. (2002) 'Mutational analysis defines the 5'-kinase and 3'-phosphatase active sites of T4 polynucleotide kinase', *Nucleic Acids Research*, 30(4), pp. 1073–1080. doi: 10.1093/nar/30.4.1073.

Wang, L. K. *et al.* (2003) 'Mutational analysis of bacteriophage T4 RNA ligase 1: Different functional groups are required for the nucleotidyl transfer and phosphodiester bond formation steps of the ligation reaction', *Journal of Biological Chemistry*. American Society for Biochemistry and Molecular Biology, 278(32), pp. 29454–29462. doi: 10.1074/jbc.M304320200.

Wang, L. K., Lima, C. D. and Shuman, S. (2002) 'Structure and mechanism of T4 polynucleotide kinase: An RNA repair enzyme', *EMBO Journal*. European Molecular Biology Organization, 21(14), pp. 3873–3880. doi: 10.1093/emboj/cdf397.

Wang, L. K., Schwer, B. and Shuman, S. (2006) 'Structure-guided mutational analysis of T4 RNA ligase 1', *RNA*. Cold Spring Harbor Laboratory Press, 12(12), pp. 2126–2134. doi: 10.1261/rna.271706.

Wang and Shuman (2001) 'Domain Structure and Mutational Analysis of T4 Polynucleotide Kinase', *Journal of Biological Chemistry*, 276(29), pp. 26868–26874. doi: 10.1074/jbc.M103663200.

Wang, Y. *et al.* (2004) 'A novel strategy to engineer DNA polymerases for enhanced processivity and improved performance in vitro', *Nucleic Acids Research*. Oxford University Press, 32(3), pp. 1197–1207. doi: 10.1093/nar/gkh271.

Watson, J. D. *et al.* (1953) 'Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid.', *Nature*, 171(4356), pp. 737–8. doi: 10.1126/science.aaf5508.

Watson, J. D. and Crick, F. H. (1953) 'The structure of DNA.', *Cold Spring Harbor symposia on quantitative biology*, 18, pp. 123–131. doi: 10.1101/SQB.1953.018.01.020.

Weinfeld, M. *et al.* (2011) 'Tidying up loose ends: The role of polynucleotide kinase/phosphatase in DNA strand break repair', *Trends in Biochemical Sciences*. Elsevier Current Trends, pp. 262–271. doi: 10.1016/j.tibs.2011.01.006.

Weiss, B. and Richardson, C. C. (1967) 'Enzymatic breakage and joining of deoxyribonucleic acid, I. Repair of single-strand breaks in DNA by an enzyme system from Escherichia coli infected with T4 bacteriophage.', *Proceedings of the National Academy of Sciences of the United States of America*, 57(4), pp. 1021–1028. doi: 10.1073/pnas.57.4.1021.

Weller, G. R. *et al.* (2002) 'Identification of a DNA nonhomologous end-joining complex in bacteria.', *Science (New York, N.Y.)*. American Association for the Advancement of Science, 297(5587), pp. 1686–9. doi: 10.1126/science.1074584.

Werle, E. *et al.* (1994) 'Convenient single-step, one tube purification of PCR products for direct sequencing.', *Nucleic acids research*. Oxford University Press, 22(20), pp. 4354–5. Available at: http://www.ncbi.nlm.nih.gov/pubmed/7937169 (Accessed: 15 May 2018).

Wilkins, M. H. (1956) 'Physical studies of the molecular structure of deoxyribose nucleic acid and nucleoprotein.', *Cold Spring Harbor symposia on quantitative biology*. Cold Spring Harbor Laboratory Press, 21, pp. 75–90. doi: 10.1101/SQB.1956.021.01.007.

Wilkinson, A., Day, J. and Bowater, R. (2001) 'Bacterial DNA ligases', *Molecular Microbiology*. Wiley/Blackwell (10.1111), 40(6), pp. 1241–1248. doi: 10.1046/j.1365-2958.2001.02479.x.

Williamson, A., Hjerde, E. and Kahlke, T. (2016) 'Analysis of the distribution and evolution of the ATP-dependent DNA ligases of bacteria delineates a distinct phylogenetic group "Lig E"', *Molecular Microbiology*, 99(2), pp. 274–290. doi: 10.1111/mmi.13229.

Wilson, G. G. and Murray, N. E. (1979) 'Molecular cloning of the DNA ligase gene from bacteriophage T4', *Journal of Molecular Biology*, 132(3), pp. 471–491. doi: 10.1016/0022-2836(79)90270-5.

Wilson, R. H. *et al.* (2013) 'Engineered DNA ligases with improved activities in vitro', *Protein engineering, design & selection : PEDS*, 26(7), pp. 471–8. doi: 10.1093/protein/gzt024.

Wood, B. *et al.* (1978) 'Attachment of Tail Fibers in Bacteriophage T4 Assembly', *Journal of Biological Chemistry*, 253(7), pp. 2437–2445. Available at: http://www.jbc.org/ (Accessed: 10 January 2019).

Wu, X. *et al.* (2018) 'Flexible Peptide Linkers Enhance the Antimicrobial Activity of Surface-Immobilized Bacteriolytic Enzymes', *ACS Applied Materials & Interfaces*, 10(43), pp. 36746–36756. doi: 10.1021/acsami.8b14411.

Xu, P. *et al.* (2015) 'An improved protocol for small RNA library construction using High Definition adapters', *Methods in Next Generation Sequencing*. De Gruyter Open, 2(1). doi: 10.1515/mngs-2015-0001.

Yakovleva, L. and Shuman, S. (2006) 'Nucleotide misincorporation, 3'-mismatch extension, and responses to abasic sites and DNA adducts by the polymerase component of bacterial DNA ligase D', *Journal of Biological Chemistry*. American Society for Biochemistry and Molecular Biology, 281(35), pp. 25026–25040. doi: 10.1074/jbc.M603302200.

Yan, H. G., Shi, Z. T. and Tsai, M. D. (1990) 'Mechanism of adenylate kinase. Structural and functional demonstration of arginine-138 as a key catalytic residue that cannot be replaced by lysine.', *Biochemistry*, 29(27), pp. 6385–92. Available at: http://www.ncbi.nlm.nih.gov/pubmed/2119801 (Accessed: 14 February 2019).

Yin, S. *et al.* (2004) 'Characterization of bacteriophage KVP40 and T4 RNA ligase 2', *Virology*, 319, pp. 141–151. doi: 10.1016/j.virol.2003.10.037.

Yin, S., Ho, C. K. and Shuman, S. (2003) 'Structure-function analysis of T4 RNA ligase 2.', *Journal of Biological Chemistry*, 278(20), pp. 17601–8. doi: 10.1074/jbc.M300817200.

Zamore, P. D. and Haley, B. (2005) 'Ribo-gnome: the big world of small RNAs.', *Science (New York, N.Y.)*, 309(5740), pp. 1519–24. doi: 10.1126/science.1111444.

Zhang, C. (2009) 'Novel functions for small RNA molecules.', *Current opinion in molecular therapeutics*, 11(6), pp. 641–51. Available at: http://www.ncbi.nlm.nih.gov/pubmed/20072941 (Accessed: 9 January 2019).

Zhang, W., Dahlberg, J. E. and Tam, W. (2007) 'MicroRNAs in tumorigenesis: a primer.', *The American journal of pathology*. American Society for Investigative Pathology, 171(3), pp. 728–38. doi: 10.2353/ajpath.2007.070070.

Zhao, H. *et al.* (2010) 'A Pilot Study of Circulating miRNAs as Potential Biomarkers of Early Stage Breast Cancer', *PLoS ONE*. Edited by C. Creighton. Public Library of Science, 5(10), p. e13735. doi: 10.1371/journal.pone.0013735.

Zheng, S. *et al.* (2010) 'Label-free optical diagnosis of hepatitis B virus with genetically engineered fusion proteins', *Talanta*. Elsevier, 82(2), pp. 803–809. doi: 10.1016/J.TALANTA.2010.05.059.

Zhu, H. *et al.* (2006) 'Atomic structure and nonhomologous end-joining function of the polymerase component of bacterial DNA ligase D.', *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, 103(6), pp. 1711–6. doi: 10.1073/pnas.0509083103.

Zhu, H. and 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), pp. 13873–13881. doi: 10.1074/jbc.M600055200.

Zhu, H. and Shuman, S. (2007) 'Characterization of Agrobacterium tumefaciens DNA ligases C and D', *Nucleic Acids Research*. Oxford University Press, 35(11), pp. 3631–3645. doi: 10.1093/nar/gkm145.

Zhu, H. and Shuman, S. (2008) 'Bacterial Nonhomologous End Joining Ligases Preferentially Seal Breaks with a 3-OH Monoribonucleotide *'. doi: 10.1074/jbc.M705476200.

Zhu, H., Wang, L. K. and Shuman, S. (2005) 'Essential constituents of the 3'-phosphoesterase domain of bacterial DNA ligase D, a nonhomologous end-joining enzyme', *Journal of Biological Chemistry*, 280(40), pp. 33707–33715. doi: 10.1074/jbc.M506838200.

Zhuang, F. *et al.* (2012) 'Structural bias in T4 RNA ligase-mediated 3'-adapter ligation', *Nucleic Acids Research*, 40. doi: 10.1093/nar/gkr1263.

Chapter 9

Appendices

Chapter 9: Appendices

9.1. Appendix 1

Primer design for the addition of the *BsaI* cut sites to the T4 RNA ligases and the T4Dnl OB domain.

T4Dnl Gene:

catatgattcttaaaattctgaacgaaatagcatctattggttcaactaaacagaagca
agcaattcttgaaaagaataaagataatgaattgcttaaacgagtatatcgtctgactt
attctcgtgggttacagtattatatcaagaaatggcctaaacctggtattgctacccag
agttttggaatgttgactcttaccgatatgcttgacttcattgaattcacattagctac
tcggaaattgactggaaatgcagcaattgaggaattaactggatatatcaccgatggta
aaaaagatgatgttgaagttttgcgtcgagtgatgatgcgagaccttgaatgtggtgct
tcagtatctattgcaaacaaagtttggccaggtttaattcctgaacaacctcaaatgct
cgcaagttcttatgatgaaaaaggcattaataagaatatcaaatttccagcctttgctc
agttaaaagctgatggagctcggtgttttgctgaagttagaggtgatgaattagatgat
gttcgtcttttatcacgagctggtaatgaatatctaggattagatcttcttaaggaaga
gttaattaaaatgaccgctgaagcccgccagattcatccagaaggtgtgttgattgatg
gcgaattggtataccatgagcaagttaaaaaggagccagaaggcctagattttcttttt
gatgcttatcctgaaaacagtaaagctaaagaattcgccgaagtagctgaatcacgtac
tgcttctaatggaatcgccaataaatctttaaagggaaccatttctgaaaaagaagcac
aatgcatgaagtttcaggtctgggattatgtcccgttggtagaaatatacagtcttcct
gcatttcgtttgaaatatgatgtacgtttttctaaactagaacaaatgacatctggata
tgataaagtaattttaattgaaaaccaggtagtaaataacctagatgaagctaaggtaa
tttataaaaagtatattgaccaaggtcttgaaggtattattctcaaaaatatcgatgga
ttatgggaaaatgctcgttcaaaaaatctttataaatttaaagaagtaattgatgttga
tttaaaaattgtaggaatttatcctcaccgtaaagaccctactaaagcgggtggattta
ttcttgagtcagagtgtggaaaaattaaggtaaatgctggttcaggcttaaaagataaa
gccggtgtaaaatcgcatgaacttgaccgtactcgcattatggaaaaccaaaattatta
tattggaaaaattctagagtgcgaatgcaacggttggttaaaatctgatggccgcactg
attacgttaaattatttcttccgattgcgattcgtttacgtgaagataaaactaaagct
aatacattcgaagatgtatttggtgattttcatgaggtaactggtctatgaggatcc

BsaI site

5′...GGTCTC(N)$_1$▼...3′
3′...CCAGAG(N)$_5$▲...5′

TGGTCTCT *BsaI* recognition site plus extra nucleotides

CATC First 4 nucleotides of his tag

ACCC 4 nucleotides of RBS

Forward Primer:

TGGTCTCT**gaagtaattgatgttgatttaaaaattgtaggaatttatcc**

Tm = 63.4 °C

Reverse Primer: atgaggtaactggtctatgaggatccCATC

Reverse Compliment:TGGTCTCT GATGggatcctcatagaccagttacctcat

Tm = 63 °C

T4 Rnl1 gene:

catatgcaagaacttttaacaatttaatggaactatgtaaggattcgcagcgtaagtt
tttttactcagatgatgtaagtgcatctggaagaacttacagaattttctcatataatt
atgcatcttattctgattggttacttccagatgcactagaatgtcgtggaattatgttt
gaaatggatggagaaaaaccagtaagaattgcttctcgtcctatggaaaagtttttaa
cttgaatgaaaatccgttcacgatgaatatcgatttaaacgatgttgattatattctaa
caaaagaagacgggtctttggtatcaacttatttagacggtgatgaaattctgttcaaa
tcaaagggttcaatcaaatctgagcaggctttaatggctaatggaattttgatgaatat
taatcaccatcggttgcgtgatagacttaaagaattagctgaagatggatttactgcta
acttcgaattcgttgccccgacgaatagaatcgttcttgcttatcaagagatgaaaatt
attttactgaatgttcgtgaaaacgaaacgggtgaatacatttcatacgatgatattta
taaagatgctactcttcgtccgtatctagttgaacgatacgaaatcgatagccctaaat
ggatagaagaagctaaaaatgcagaaaacatcgaaggctatgttgctgtgatgaaagat
ggttctcattttaaaattaagtctgactggtacgtgtctcttcatagtacaaaaagttc
attagataatccagaaaaattgtttaagactattattgatggtgcatcagatgatctta
aagcaatgtatgctgacgatgaatattcatacagaaaaattgaagcatttgaaacgact
tatctgaagtacttagaccgagctctgtttttagttcttgactgtcataataagcattg
cggtaaggatagaaagacttatgcaatggaagcacaaggtgttgctaaaggtgctggaa
tggatcacctgttcggtatcatcatgagcttataccaggggtacgatagtcaagaaaag
gtcatgtgtgaaatcgaacagaatttttgaaaaattataaaaaatttatcccagaagg
atactaaggatcc

Forward Primer: TGGTCTCT
ACCCatgcaagaacttttaacaatttaatggaactatg

Tm = 61 °C

Reverse Primer: ttgctgtgatgaaagatggttctcattttaaaattaaggaag

Reverse Compliment:
TGGTCTCTcttccttaattttaaaatgagaaccatctttcatcacag

Tm = 62 °C

T4Rnl2 gene:

catatgtttaaaaagtatagcagtcttgaaaatcattacaactctaaatttattgaaaa
actttatagcttgggattgactggtggggagtgggtagctcgtgaaaagattcacggca
caaatttctcattgattattgagcgtgataaagtgacttgcgctaaacgcactggaccg
attcttcctgctgaagatttctttgggtatgaaattattttgaagaattatgctgattc

cattaaagctgtacaagatattatggaaacctcagcggttgtatcttatcaagtctttg
gcgaattcgctggacctggcattcagaagaatgttgattatGgtgataaagatttttat
gtatttgacattattgttactacagaaagcggtgatgtgacttatgtagatgattatat
gatggaatcattctgtaatacatttaaatttaaaatggctccacttttaggtcgcggta
aatttgaagagcttattaaattgccaaatgatttagattctgtcgtccaagattataat
tttacagtagaccatgctggattagttgatgcaaataaatgcgtttggaatgccgaagc
aaaaggcgaagtatttactgctgaaggatatgtattgaaaccttgttatccttcttggc
ttcgtaatggaaatcgtgtagcgattaaatgcaagaactctaaatttagtgaaaagaaa
aagtctgataagcctattaaagctaaagttgagctatcagaagctgataacaaattggt
gggaattttagcttgttacgttacactgaaccgcgtaaataacgttatttctaaaattg
gcgaaattggtccaaaggattttggaaaggtgatggggctaactgttcaagatattttg
gaagaaacttctcgtgaaggtattactctaactcaagcagataatccttctttgattaa
aaaggaattagttaaaatggtacaagatgtacttcgtccagcttggattgagttggtaa
gttaaggatcc

Forward Primer:
TGGTCTCTACCCatgtttaaaaagtatagcagtcttgaaaatcattac

Tm = 61.5 °C

Reverse Primer: ggaaatcgtgtagcgattaaatgcaaggaag

Reverse Compliment: TGGTCTCTcttccttgcatttaatcgctacacgatttcc

Tm = 61.9 °C

9.2. Appendix 2

A predicted structure forT4Dnl generated using Phyre2 software, available at http://www.sbg.bio.ic.ac.uk/phyre2/html/page.cgi?id=index and illustrated using CCP4MG software. The amino acid sequence for T4Dnl was uploaded to the Phyre2 website, an image is generated from this sequence based on its similarity to the sequences of other published structures. When this image was created, the structure for T4Dnl was not solved and the most similar sequence to T4Dnl was human ligase I (PDB 1X9N) (Pascal *et al.*, 2004) which this model is based on. The generated image file was edited in the CCP4MG software. The polypeptide motifs are highlighted: motif I (red), motif III (orange), motif IIIb (pink), motif IV (blue), motif V (green) and the OB domain (purple).