

# MuGam homologue proteins - an analysis of function and viability as contributors to DNA repair

Theodore F Hewitt

MSc by research  
University of York  
Biology  
December 2020

## Abstract

Until recently it was thought that the only accurate DNA double strand break (DSB) repair mechanism used by bacteria was homologous recombination (HR), recent discoveries have found proteins that are orthologous to eukaryotic Ku heterodimeric proteins. These prokaryotic Ku proteins like their eukaryotic counterparts function by binding non-specifically to the ends of DSBs to recruit a DNA ligase to join the two ends of the DNA back together, this mechanism is known as non-homologous end joining (NHEJ). Bacteria that can utilize NHEJ to repair DNA DSBs are at a significant advantage to those which cannot but only a select few species contain a gene for the prokaryotic Ku protein. The protein Gam from bacteriophage Mu (MuGam) has been identified as possessing sequence homology with the eukaryotic Ku heterodimer which has led to research into the potential for this MuGam protein to contribute to the NHEJ repair mechanism as a functional orthologue to the eukaryotic Ku70/80 complex. One study found that introducing MuGam to *Escherichia coli* K-12 MG1655 cells lead to an increase in both survival rate and instances of accurate and inaccurate repair when the cells were subjected to a DSB inducing element. The gene for MuGam or similar slightly mutated genes (MuGam homologues) are commonly found in many different of bacteria. Here eight different MuGam homologues are shown to be capable of being overproduced in *Escherichia coli* K-12 MG1655 that are functional in binding linear DNA using electromobility shift assays (EMSA). The sequence similarity between MuGam and the eukaryotic Ku proteins is very low and secondary structure predictors do not support structural similarities between the eukaryotic Ku protein and the MuGam protein.

## Table of Contents

<b>Abstract</b> .....	<b>I</b>
<b>List of Contents</b> .....	<b>II</b>
<b>List of Figures</b> .....	<b>VII</b>
<b>List of Tables</b> .....	<b>XI</b>
<b>Author's Declaration</b> .....	<b>XII</b>
<b>1. Introduction</b> .....	<b>1</b>
1.1 Host-nuclease inhibitor protein Gam .....	2
1.2 The prominence of Mu prophage throughout the kingdom of bacteria.....	3
1.2.1 Incomplete prophage regions of Bacterial host genomes.....	3
1.2.2 The variations in Mu prophage DNA sequences .....	4
1.3 MuGam as a DNA binding protein with host-cell benefits.....	4
1.4 DNA mimic proteins.....	5
1.5. Aims and objectives.....	6
<b>2. Materials and Methods</b> .....	<b>8</b>
2.1 Buffers and media.....	9
2.2 Primers.....	10
2.3 Measuring concentrations of DNA, protein and cell density using Absorbance spectroscopy..	11
2.3.1 Estimating DNA concentration using absorbance spectroscopy.....	11
2.3.2 Estimating Protein concentration using absorbance spectroscopy.....	11
2.3.3 Estimating cell density using absorbance spectroscopy .....	12
2.4 Synthesised genes.....	12
2.5 PCR target gene amplification.....	12
2.6 Construction of expression vector with target gene.....	13
2.6.1 Cloning with Zero Blunt® TOPO® .....	14
2.6.2 Transformation to competent amplification strain <i>E. coli</i> DH5- $\alpha$ by heat shock.....	14
2.6.3 Plasmid digestion using restriction endonucleases.....	15
2.6.4 Ligation of gene-encoding DNA fragment into expression vector.....	15
2.7 Transformation to expression strain by high efficiency electroporation.....	16
2.8 Diagnostic confirmation of successful transformation.....	16
2.8.1 Diagnostic PCR using GoTaq® DNA Polymerase (M300).....	16

2.8.2 Diagnostic Digest.....	17
2.9 Target protein overproduction and purification.....	17
2.9.1 Tests to determine ideal bacterial growth and induction conditions for over- production of recombinant MuGam-homologue proteins.....	17
2.9.2 HI1450 protein overproduction.....	18
2.9.3 Diagnostic confirmation of protein production by HisPur purification.....	18
2.9.4 Purification of target protein.....	18
2.9.5 Large-scale over-production and extraction of recombinant protein using lysozyme and Triton X-100 for bacterial cell lysis”.....	20
2.9.6 Recombinant protein purification by Ni <sup>2+</sup> -affinity chromatography.....	20
2.9.7 Purification of target protein.....	21
2.9.7.1 Gam homologue proteins.....	21
2.9.7.2 Over-production and purification of HI1450 recombinant protein.....	22
2.10 Analyse DNA binding properties by EMSA.....	23
2.10.1 Standard EMSA for positive control with Im9HiGam.....	23
2.11 Competitive binding assay.....	24
2.12 Bioinformatic study of 3D structure comparison.....	25
2.12.1 Using HiGam protein sequence to predict secondary structure elements.....	25
2.12.2 Target template structure alignment.....	25
2.13 Mass spectrometry analysis of natural protein production in <i>H. influenzae</i> Rd KW20.....	26
2.13.1 <i>H. influenzae</i> Rd KW20 cell culture growth .....	26
2.13.2 Cell Lysis using Lysozyme.....	26
2.13.3 SDS-PAGE gel Digestion.....	26
2.13.4 LC-MS/MS.....	27
2.13.5 Database Searching.....	27
<b>3. Results – DNA Double strand break repair in Bacteria containing MuGam homologue genes..</b>	<b>28</b>
3.1. Introduction.....	29
3.2. Results.....	29
3.2.1 Investigating the widespread conservation of MuGam homologues in bacterial genomes and DNA ligases.....	29
3.2.2 The absence of LigA in <i>A. paragallinarum</i> JF4211.....	31

3.2.3 Domain analysis of LigA and fragmentation of LigA protein from <i>A. paragallinarum</i> JF4211.....	31
3.3. Discussion.....	34
<b>4. Results – 3D structure analysis of MuGam homologues.....</b>	<b>35</b>
4.1. Introduction.....	36
4.1.1 3-Dimensional structure for MuGam homologue from <i>Desulfovibrio vulgaris</i> .....	36
4.1.2 MuGam as an orthologue to Human Ku70/80.....	36
4.1.3 Secondary structure prediction principles.....	37
4.1.4 Secondary structure prediction.....	37
4.2 Results.....	38
4.2.1 SWISS-MODEL secondary structure predictions.....	38
4.2.2 HHpred secondary structure predictions.....	39
4.2.3 Phyre2 secondary structure predictions.....	40
4.2.4 JPred 4 secondary structure prediction.....	43
4.2.5 One-to-one threading HiGam sequence with Ku DNA binding site structure.....	44
4.3. Discussion.....	45
<b>5. Results - Production of recombinant MuGam homologue proteins.....</b>	<b>46</b>
5.1. Introduction.....	47
5.1.1 Gene transformation using pQE2 as a vector plasmid.....	47
5.1.2 Transforming <i>E. coli</i> DH5- $\alpha$ with pUC-57-Gam plasmids by heat shock.....	48
5.2. Results.....	48
5.2.1 Diagnostic PCR and double digest of recombinant pQE2-Im9-Gam homologue.....	48
5.2.2 Plasmid sequencing.....	50
5.2.3 Determining optimal growing and induction conditions .....	51
5.2.4 MuGam homologue protein purification by FPLC.....	57
5.2.5 Analysing non-target proteins eluted from FPLC purification of target Im9-MuGam homologue proteins .....	64
5.2.6 Producing pET15b-HI1450 recombinant expression plasmid.....	67
5.2.7 Overexpression of HI1450.....	72
5.3. Discussion.....	75
<b>6. Results – Characterisation of DNA Binding Properties for Gam Homologues .....</b>	<b>77</b>
6.1. Introduction.....	78

6.1.1 Action of HiGam in the presence of linear DNA.....	78
6.2. Results.....	79
6.2.1 Electrophoretic mobility shift assay (EMSA) to determine Gam homologue function.....	79
6.2.2 Relationship between molecular weight and point of saturation.....	86
6.2.3 Competitive inhibition analysis using HI1450 DNA mimic .....	88
6.3. Discussion.....	90
<b>7. Results – HiGam protein production in <i>Haemophilus influenzae</i> Rd KW20.....</b>	<b>92</b>
7.1. Introduction.....	93
7.1.1 Mu phage prophage on <i>Haemophilus influenzae</i> Rd KW20 ( <i>H. influenzae</i> Rd KW20) genome.....	93
7.1.2 Liquid chromatography-Mass spectrometry (LC-MS) to determine presence of HiGam protein in <i>H. influenzae</i> Rd KW20 cells.....	95
7.2 Results.....	95
7.2.1 Identification of key polypeptide fragments using LC-MS.....	95
7.3. Discussion.....	96
<b>8. Discussion.....</b>	<b>97</b>
8.1. The 3-Dimensional structure of MuGam.....	98
8.1.1 The structure of MuGam and the relationship between MuGam, DvGam and Eukaryotic Ku.....	99
8.1.2 How secondary structure translates to 3-Dimensional structure.....	99
8.2. The relationship between MuGam and Ligase for DNA repair.....	99
8.2.1 Proposed mechanism for DNA repair by NHEJ in bacteria.....	100
8.2.2 The presence of ligase in species with conserved MuGam homologue genes.....	100
8.2.3 How <i>Avibacterium paragallinarum</i> JF2411 may be different .....	100
8.3. The production of HiGam, HiA and HiB in <i>Haemophilus influenzae</i> Rd KW20.....	100
8.3.1 The presence of HiGam within the cell under natural conditions.....	100
8.3.2 The absence of HiA.....	101
8.4. The success of MuGam homologue and DNA mimic protein purification.....	101
8.5. Analysis of MuGam homologue function using Electrophoretic mobility shift assays (EMSA).....	102
8.5.1 How molecular weight (MW) relates to binding footprint.....	102
8.5.2 HI1450 as a competitive inhibitor to Im9-HiGam.....	103

8.6. The potential for MuGam homologues to function in a DNA repair mechanism.....	103
<b>9. References.....</b>	<b>105</b>
<b>10. Abbreviations.....</b>	<b>109</b>
10.1. Bacterial strains and proteins.....	110
<b>11. Appendix.....</b>	<b>111</b>
11.1. pQE2-Im9-HiGam plasmid sequence.....	111
11.2. 2 Gam homologue gene sequences optimised for transcription and translation in E. coli K-12 MG1655 as confirmed by sequencing by Eurofins Scientific Genomics, with translated amino acid sequences.....	113
11.3. Genetic and amino acid sequences for DNA Ligase protein fragments from A. paragallinarum JF4211.....	116
11.4. Amino acid sequences for key proteins, Ku70, Ku80, MuGam, HiGam and DvGam.....	116

## List of Figures

<b>Figure 1.</b> A schematic view of bacteriophage Mu injecting genomic DNA into a bacterial host cell. ...	2
<b>Figure 2.</b> Structural comparison between resolved structures for a MuGam homolog from <i>Desulfovibrio vulgaris</i> (DvGam), the Ku70/80 heterodimer and A theoretical model of MuGam based on a truncated version of the Ku70/80.....	3
<b>Figure 3.</b> The 3-Dimensional structure of the HI1450 protein .....	5
<b>Figure 4.</b> Schematic view of the potential mechanism for competitive inhibition of the HiGam protein by the DNA mimic HI1450 protein .....	6
<b>Figure 5.</b> Shows the conserved residues found in the adenylation domain of the LigA proteins of five different bacteria, <i>Citrobacter rodentium</i> (strain ICC168), <i>Staphylococcus aureus</i> , <i>Escherichia coli</i> O157:H7, <i>Avibacterium paragallinarum</i> JF4211 fragment 1 and <i>Haemophilus influenzae</i> Rd KW20..	31
<b>Figure 6.</b> Figure 1 from (Singleton <i>et al.</i> , 1999) showing the conserved residues of the NAD+-dependent DNA ligase (LigA) from different organisms ( <i>Enterococcus faecalis</i> (Efa), <i>Mycobacterium tuberculosis</i> (Mtu), <i>Bacillus stearothermophilus</i> (Bst), <i>Escherichia coli</i> (Eco) and <i>Tiedemannia filiformis</i> (Tfi). .....	33
<b>Figure 7.</b> 3-dimensional solved structure for putative host-nuclease inhibitor protein Gam from <i>Desulfovibrio vulgaris</i> .....	36
<b>Figure 8.</b> The Gam protein of bacteriophage Mu is an orthologue of eukaryotic Ku (d’Adda di Fagagna <i>et al.</i> , 2003).....	37
<b>Figure 9.</b> Schematic view of the top hits with DvGam as the to hit (red). The bar position and length correspond to the position and length of the region with sequence alignment.....	40
<b>Figure 10.</b> Shows the JPred 4 secondary structure prediction for Ku70 and Ku80 DNA binding domains .....	43
<b>Figure 11.</b> Shows the JPred 4 secondary structure prediction for DvGam and HiGam peptide sequences. ....	43
<b>Figure 12.</b> The predicted conformation of the HiGam protein sequence when forced into the full Ku80 3D structure (1JEY) when using the One-to-one threading mode on the Phyre2 tool. ....	44
<b>Figure 13.</b> Fitting the Ku70 DNA binding site sequence to the 3D structure of DvGam when using the One-to-one threading mode on the Phyre2 tool.....	45
<b>Figure 14.</b> Simplified schematic view of the methods described in 2.3 showing the restriction digest of the pUC-57-KpGam and pQE2-Im9-HiGam and subsequent ligation to form the final recombinant pQE2-Im9-KpGam. ....	48
<b>Figure 15.</b> 1% (w/v) agarose-TAE electrophoresis gel images presenting evidence of successful amplification of target genes from transformed <i>E. coli</i> DH5- $\alpha$ cells. ....	49
<b>Figure 16.</b> 1% (w/v) agarose-TAE gel electrophoresis showing the diagnostic double digest for each Gam homolog ligated with pQE2-Im9 and transformed to <i>E. coli</i> DH5- $\alpha$ . ....	50
<b>Figure 17.</b> SDS-PAGE analysis of <i>E. coli</i> K-12 MG1655 soluble and insoluble protein fractions after IPTG induction for overproduction of Im9-RcGam protein from the pQE2-Im9-RcGam recombinant plasmid incubated at 20°C post-induction. ....	52



<b>Figure 18.</b> SDS-PAGE analysis of <i>E. coli</i> K-12 MG1655 soluble and insoluble protein fractions after IPTG induction for overproduction of Im9-RcGam protein from the pQE2-Im9-RcGam recombinant plasmid incubated at 37°C post-induction. ....	53
<b>Figure 19.</b> SDS-PAGE analysis of <i>E. coli</i> K-12 MG1655 soluble and insoluble protein fractions after IPTG induction for overproduction of Im9-PaGam protein from the pQE2-Im9-PaGam recombinant plasmid incubated at 20°C post-induction. ....	54
<b>Figure 20.</b> . SDS-PAGE analysis of <i>E. coli</i> K-12 MG1655 soluble and insoluble protein fractions after IPTG induction for overproduction of Im9-PaGam protein from the pQE2-Im9-PaGam recombinant plasmid incubated at 20°C post-induction. ....	55
<b>Figure 21.</b> 15% (w/v) polyacrylamide-SDS gel images showing lysates for each transformed <i>E. coli</i> MG1655 culture containing each Gam homolog gene. ....	56
<b>Figure 22.</b> 10% (w/v) polyacrylamide-SDS gel image showing the protein fractions analysed to identify the fractions which contain the most target Im9-KpGam protein and the least amount of unwanted protein. ....	58
<b>Figure 23.</b> 10% (w/v) polyacrylamide-SDS gel image showing the protein fractions analysed to identify the fractions which contain the most target Im9-ApGam protein and the least amount of unwanted protein. ....	59
<b>Figure 24.</b> 10% (w/v) polyacrylamide-SDS gel iamge showing the protein fractions analysed to identify the fractions which contain the most target Im9-SsGam protein and the least amount of unwanted protein. ....	60
<b>Figure 25.</b> 10% (w/v) polyacrylamide-SDS gel image showing the protein fractions analysed to identify the fractions which contain the most target Im9-SeGam protein and the least amount of unwanted protein. ....	61
<b>Figure 26.</b> 10% (w/v) polyacrylamide-SDS gel image showing the protein fractions analysed to identify the fractions which contain the most target Im9-EcGam protein and the least amount of unwanted protein. ....	62
<b>Figure 27.</b> 10% (w/v) polyacrylamide-SDS gel image showing the protein fractions analysed to identify the fractions which contain the most target Im9-RcGam protein and the least amount of unwanted protein. ....	63
<b>Figure 28.</b> 10% (w/v) polyacrylamide-SDS gel image showing the protein fractions analysed to identify the fractions which contain the most target Im9-EcGam protein and the least amount of unwanted protein. ....	64
<b>Figure 29.</b> The MW calibration graph used to estimate the MW of the Im9-RcGam protein and the two protein fragments as seen in <b>Figure 26</b> . ....	65
<b>Figure 30.</b> 1.5% (w/v) agarose-TAE gel image showing Im9-RcGam fractions 6-9 and fractions 12-16 analysed by EMSA using 500bp linear DNA. ....	66
<b>Figure 31.</b> Products of the KOD PCR for two samples with lane 2 using 10µL of template DNA .....	67
<b>Figure 32.</b> Resulting bands from diagnostic Gotaq PCR from two colonies A and B confirming the presence of the HI1450 gene in colony B. ....	68
<b>Figure 33.</b> Diagnostic double restriction digestion of plasmids from colonies A and B .....	69
<b>Figure 34.</b> 1% agarose (w/v) + TAE gel image of the diagnostic Gotaq PCR product of the transformed <i>E. coli</i> DH5-alpha cell colony with HI1450-pET15b .....	70

<b>Figure 35.</b> 1% (w/v) agarose + TAE gel image of the diagnostic double restriction digest of the miniprep purified HI1450-pET15b plasmid.....	71
<b>Figure 36.</b> The sequencing data output received in the region with the unexpected anomalous sequence.....	72
<b>Figure 37.</b> 15% (w/v) polyacrylamide-SDS gel electrophoresis analysis of HI1450 over-production using IPTG induction and <i>E. coli</i> BL21(DE3) cells transformed with HI1450-pET15b plasmid.....	73
<b>Figure 38.</b> 15% (w/v) polyacrylamide-SDS gel electrophoresis was used to analyse the purified target 6x His tagged HI1450 protein.....	74
<b>Figure 39.</b> 10% (w/v) polyacrylamide-SDS gel image showing the protein fractions analysed to identify the fractions which contain the most target HI1450 protein and the least contaminant protein.....	75
<b>Figure 40.</b> Electrophoretic mobility shift assay of 0.3µM (6.67µg/mL) 500 base pair (bp) linear DNA on a 1.5% (w/v) agarose-TAE gel. Shows the shift in mobility of DNA in the presence of increasing Im9-HiGam protein concentration. ....	79
<b>Figure 41.</b> Electrophoretic mobility shift assay of 0.3µM (6.67µg/mL) 500bp linear DNA on a 1.5% (w/v) agarose-TAE gel. Shows the shift in mobility of DNA in the presence of increasing Im9-KpGam protein concentration. ....	80
<b>Figure 42.</b> ) Electrophoretic mobility shift assay of 0.3µM (6.67µg/mL) 500bp linear DNA on a 1.5% (w/v) agarose-TAE gel. Shows the shift in mobility of DNA in the presence of increasing Im9-ApGam protein concentration. ....	81
<b>Figure 43.</b> Electrophoretic mobility shift assay of 0.3µM (6.67µg/mL) 500bp linear DNA on a 1.5% (w/v) agarose-TAE gel. Shows the shift in mobility of DNA in the presence of increasing Im9-SsGam protein concentration. ....	82
<b>Figure 44.</b> Electrophoretic mobility shift assay of 0.3µM (6.67µg/mL) 500bp linear DNA on a 1.5% (w/v) agarose-TAE gel. Shows the shift in mobility of DNA in the presence of increasing Im9-SeGam protein concentration. ....	83
<b>Figure 45.</b> Electrophoretic mobility shift assay of 0.3µM (6.67µg/mL) 500bp linear DNA on a 1.5% (w/v) agarose-TAE gel. Shows the shift in mobility of DNA in the presence of increasing Im9-EcGam protein concentration. ....	84
<b>Figure 46.</b> Electrophoretic mobility shift assay of 0.3µM (6.67µg/mL) 500bp linear DNA on a 1.5% (w/v) agarose-TAE gel. Shows the shift in mobility of DNA in the presence of increasing Im9-RcGam protein concentration. ....	85
<b>Figure 47.</b> The result of electrophoresis of 0.3µM (6.67µg/mL) 500bp linear DNA in the presence and absence of Im9-PaGam. ....	86
<b>Figure 48.</b> The relationship between dimeric MW and the increase in MW of the 500bp linear DNA at saturation when Gam homologue proteins are ordered in ascending order of increase in MW on the gel at saturation. ....	87
<b>Figure 49. A:</b> Compares the MW of each Im9-MuGam homologue (blue) to the estimated binding footprint for each protein (orange). <b>B:</b> Shows the total increase in MW at saturation (blue) compared to the estimated DNA binding footprint for each of the MuGam homologue proteins. ....	88
<b>Figure 50.</b> .5% (w/v) agarose + TAE electrophoresis gel images showing the effects of competitive inhibition by HI1450 on Im9-HiGam binding to linear DNA .....	89

<b>Figure 51.</b> Evidence that the MuA, MuB and MuGam genes are transcribed onto a single polycistronic mRNA from purified samples of DNA amplified by PCR using selected primers to produce different lengths of DNA corresponding to different regions found on the Mu prophage after transcription (cDNA). .....	93
<b>Figure 52.</b> Schematic view of the <i>H. influenzae</i> Rd KW20 chromosome from Phage search tool (PHAST) search for <i>H. influenzae</i> Rd KW20 .....	94
<b>Figure 53.</b> 2PU2 solved crystal structure for putative host-nuclease inhibitor protein Gam from <i>Desulfovibrio vulgaris</i> (Bonano et. al 2007). .....	98
<b>Figure 54.</b> Schematic view of the pQE2-Im9-HiGam. ....	112

## List of Tables

<b>Table 1.</b> Composition of each Lysis buffers used throughout the methods in sections 2.8 and 2.9 .....	<b>9</b>
<b>Table 2.</b> Composition of the other buffers used in sections 2.8 and 2.9. ....	<b>9</b>
<b>Table 3.</b> Composition of media components for which are dissolved in ultra-pure H <sub>2</sub> O (resistivity $\approx 18 \text{ MOhm}\cdot\text{s}$ ). ....	<b>10</b>
<b>Table 4.</b> Extinction coefficients ( $\text{M}^{-1}\text{cm}^{-1}$ ) used to calculate the monomeric protein concentration. .	<b>12</b>
<b>Table 5.</b> Restriction enzymes and buffers used for each plasmid digestion reaction with the expected efficiency for the enzymes in the buffer used. ....	<b>15</b>
<b>Table 6.</b> Standardised reagent volumes used in EMSA experiments. ....	<b>24</b>
<b>Table 7.</b> DNA Ligases present in the 25 closest sequence matches to HiGam from HMMER. ....	<b>30</b>
<b>Table 8.</b> SWISS-MODEL output for HiGam query shows a list of the top 10 matches sorted by highest percentage of sequence alignment identities. ....	<b>39</b>
<b>Table 9.</b> The top 5 hits in order of model confidence from the Phyre2 bioinformatic tool with the HiGam amino acid sequence as the query. Results are consistent with what we have seen from the other prediction tools. ....	<b>41</b>
<b>Table 10.</b> Hits 14-18 ordered by confidence (%). Shows the secondary structure alignments for the C-terminal region. ....	<b>42</b>
<b>Table 11.</b> Overview of the MuGam homologue proteins that are being investigated. ....	<b>48</b>
<b>Table 12.</b> Peptide sequences identified using LC-MS. ....	<b>97</b>

## **Authors Declaration**

I declare that the work presented here is my own original work submitted, unless otherwise stated for my MSc Thesis and that I am the sole author. This work has not been presented for an award at this, or any other University. All sources and acknowledgements are listed as references.

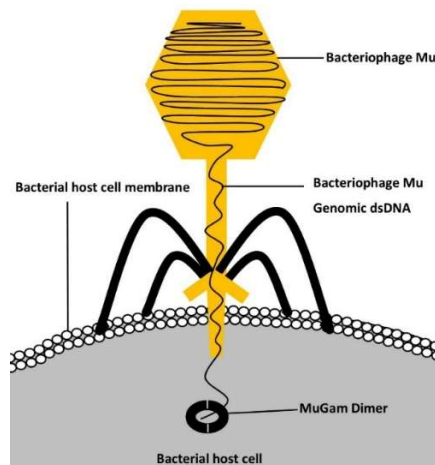
Theodore F Hewitt

# Introduction

## Chapter 1

### 1.1 Host-nuclease inhibitor protein Gam

Bacteriophage Mu injects its linear genomic DNA into its bacterial target as seen depicted in **Figure 1**. The DNA can be incorporated into the host genome at a random position and exists in its entirety as a transposable element (Fuller and Rice, 2017). As a transposon the bacteriophage Mu genome is capable of being transported to different locations at suitably identified sites on the genome of the infected host as well as between different host organisms (Haapa-Paananen, Rita and Savilahti, 2002). The Mu prophage can then lay dormant within the host genome or enter the lytic cycle to begin replicative transposition to rapidly produce copies of the virus genome as well as the viral proteins to produce many copies of the virus to infect further bacterial hosts (Clark, Pazdernik and McGehee, 2019).



**Figure 1.** A schematic view of bacteriophage Mu injecting genomic DNA into a bacterial host cell. The MuGam dimer binds to the end of the linear genomic DNA to prevent targeted DNA degradation by host bacterial defences. This likely occurs during packaging of the linear bacteriophage DNA, although it could also protect this DNA during the process of transposition.

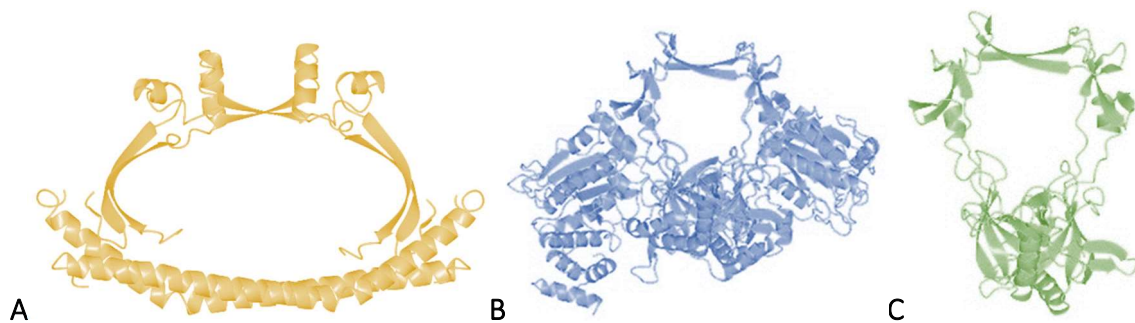
MuA and MuB are involved in transposing this viral DNA while the protein MuGam (Host-nuclease inhibitor protein) is used to protect the viral DNA by preventing exonuclease-mediated degradation by host defence pathways (Akroyd, Clayson and Higgins, 1986; Abraham and Symonds, 1990). MuGam is a homodimer which binds to linear DNA thanks to a large central channel capable of binding double-stranded DNA ends. This structure has not been confirmed experimentally, but we have theoretical structures based on sequence analysis and the full 3D structure of a similar homologous protein from *Desulfovibrio vulgaris* (DvGam) by x-ray crystallography (PDB code: 2P2U) (Bhattacharyya et al., 2018). The gene for Gam is consistently conserved within the genomes of a range of Gram-negative bacteria (Bhattacharyya et al., 2018; d’Adda di Fagagna et al., 2003). Here, we focus predominantly on *Haemophilus influenzae* strain Rd KW20 which contains a single intact prophage from bacteriophage Mu including a homologue of MuGam (termed “HiGam”). It has been suggested that the consistent conservation of the Gam protein is due to an improvement in survival, though this has not been explicitly proven. The MuGam homologue gene can be conserved in incomplete Mu prophage regions (Mu bacteriophage genome found on the bacterial host genome). Suggesting conservation of the Gam gene due to advantage to the host or as potential advantage to other phage and pathogenesis due to helping other phage to be transposed to the host genome (Bhattacharyya et al., 2018; Canchaya et al., 2003; Lawrence and Roth, 2014).

A eukaryotic heterodimeric protein Ku70/80 binds to linear DNA ends resulting from a double-strand break (DSB) and recruits a DNA ligase (Walker, Corpina and Goldberg, 2001), which joins the

## Introduction

two ends by non-homologous end-joining (NHEJ). NHEJ is the other major DNA double strand break (DSB) repair mechanism alongside homologous recombination (HR) in eukaryotes, the Ku70/80 heterodimers will bind with high affinity to each end of the DNA following a DSB and recruits DNA ligase to the site where the two ends will be joined. The mechanism is non-specific and is possible for imperfect repair to occur leading to mutations (Pastwa and Błasiak, 2003).

Ku70/80 shows sequence homology to the MuGam protein in its DNA binding domain (d'Adda di Fagagna et al., 2003). The resolved 3D structure for Ku70/80 (PDB code: 1JEQ) (Walker, Corpina and Goldberg, 2001) shows the same central channel for binding DNA ends. In humans, the Ku heterodimer will interact with ATP-dependant ligase (Lig IV) as well as DNA-PKcs, Artemis XRCC4 and XLF to form the active holoenzyme for NHEJ (Downs and Jackson, 2004; Fattah et al., 2010).



**Figure 2.** Structural comparison between **A** (Yellow): A homodimeric structure modelled from the fully resolved 3D structure of monomeric units of a MuGam homolog from *Desulfovibrio vulgaris* (DvGam) (PDB code: 2P2U) (Bonano et. al 2007). **B** (Blue): Resolved 3D structure of the Ku70/80 heterodimer (PDB code: 1JEQ) (Walker, Corpina and Goldberg, 2001). **C** (Green): A theoretical model of MuGam based on a truncated version of the Ku70/80 3D structure due to similarity in the sequence of the DNA binding site between Ku70/80 and MuGam (d'Adda di Fagagna et al., 2003).

We currently do not have experimental evidence for the structure in Figure 2C, it is purely theoretical. We are sceptical of the assumption that MuGam would conform to a structure like that in Figure 2A as the primary sequences of DvGam and MuGam only share 26% sequence identities (same amino acids in the same positions) and 50% similarity (similarly charged and/or structured amino acids in the same positions). As a comparison, MuGam and HiGam (a MuGam homologue identified in *Haemophilus influenzae* Rd KW20) share 61% identity and 75% similarity at the sequence level. These percentages were obtained through BLASTp multiple sequence alignment. Bhattacharyya et al. also point out that this structure (**Fig2 A**) is large enough (approx. 50Å) to potentially accommodate two DNA duplexes (23Å each) within its central channel. This structural feature is totally different to the structure of the Ku70/80 which exclusively binds to one duplex with a channel spanning approx. 27Å.

### 1.2 The prominence of Mu prophage throughout the kingdom of bacteria

Bacteriophage Mu was first discovered due to the invasion of *Escherichia coli* (*E. coli*) genomes though we now know that it is present on a wide variety of host bacterial species (Harshey, 2012; Akroyd and Symonds, 1986). The bacteriophage Mu is only able to directly inject its genomic DNA to a relatively small number of Enterobacteria. Due to the viral genome existing as a transposon, it is common for the genome to be passed to other bacterial strains through conjugative transfer (horizontal gene transfer) between adjacent bacterial cells (Ferrières et al., 2010). This means the intact Mu prophage can be found in a wide variety of bacteria. The prophage region can be passively



transcribed by the host to produce the Mu phage proteins without the viral DNA being triggered to kick start the lytic cycle.

### **1.2.1 Incomplete prophage regions of Bacterial host genomes**

It is very common for bacteria to remove non-essential DNA and genes from their genome, maintaining only genes that provide tangible benefit to the cell to relieve unnecessary demand on the resources for protein production and regulation in the cell (Koskiniemi *et al.*, 2012). Prophage regions of the host genome can be a target for gene deletion which often leaves genomes with a partial prophage region containing just a few of the viral genes which have not been targeted for deletion (Koskiniemi *et al.*, 2012). This also prevents the viral DNA from being triggered to enter the lytic cycle which would result in the destruction of the cell, and the infection and potential destruction of other bacterial cells. Partial Mu prophage regions are found in bacterial genomes where Mu phage genes which the host-cell does not derive any benefit from have been deleted. In these cases, the gene for MuGam is consistently conserved (Brissett and Doherty, 2009). The conservation of the gene on incomplete prophage regions suggests the gene is retained due to providing a benefit to the cell. The gene product would therefore be produced in the cell, correctly folded and functional to improve bacterial host survival or efficiency in some way.

### **1.2.2 The variations in Mu prophage DNA sequences**

The bacteriophage Mu was given its namesake due to the frequency of the mutations of the viral genes. In each of the bacterial host genomes no two sequences for the MuGam gene were the same (**Results Chapter 1: 2.1**). The different DNA sequences of the MuGam gene produce different protein sequences of which some are more similar than others, but these different proteins are homologues of each other. The differences in these protein sequences may lead to variations in structure and may give rise to function that may deviate from the function originally identified for the protein.

### **1.3 MuGam as a DNA binding protein with host-cell benefits**

The MuGam protein has been used to identify and quantify sites of spontaneous double strand breaks in living *E. coli* cells and cultured mammalian cell lines due to its high binding affinity to DNA ends. This was achieved by fusing green fluorescent protein (GFP) to the C-terminus of the MuGam protein (Shee *et al.*, 2013). The study also presented evidence that the identification of DNA double strand breaks using MuGam was inhibited by the presence of the Ku heterodimer in HeLa cells suggesting that they are in direct competition for DNA end binding.

MuGam is known to bind at linear DNA ends, and it has been shown to have sequence homology to a eukaryotic NHEJ DNA repair protein (Ku70/80) (d'Adda di Fagagna *et al.*, 2003). If the gene is consistently conserved in the bacterial host due to the protein providing a tangible benefit to the host cell, DNA repair is one possible mechanism. MuGam contributes to the recruitment of the MuA and MuB proteins that are responsible for incorporating the Mu prophage into the host genome. This involves DNA ligation in the form of the transposition mechanism (Harshey, 2012). This has led to the theory that MuGam contributes to a DNA repair mechanism resembling the NHEJ pathway found in eukaryotic cells.

Bhattacharyya *et al.* (2018) found that by using an expression vector for MuGam and NAD<sup>+</sup>-dependent ligase (LigA) in *E. coli* K-12 MG1655 instances of DNA repair increased in the presence of induced DSBs by phleomycin. They also found that by knocking out RecA, a protein essential for homologous recombination (HR), that the presence of Gam still improved DNA repair. They observed both exact accurate repair, where the resulting sequence is the same as the original, and

## Introduction

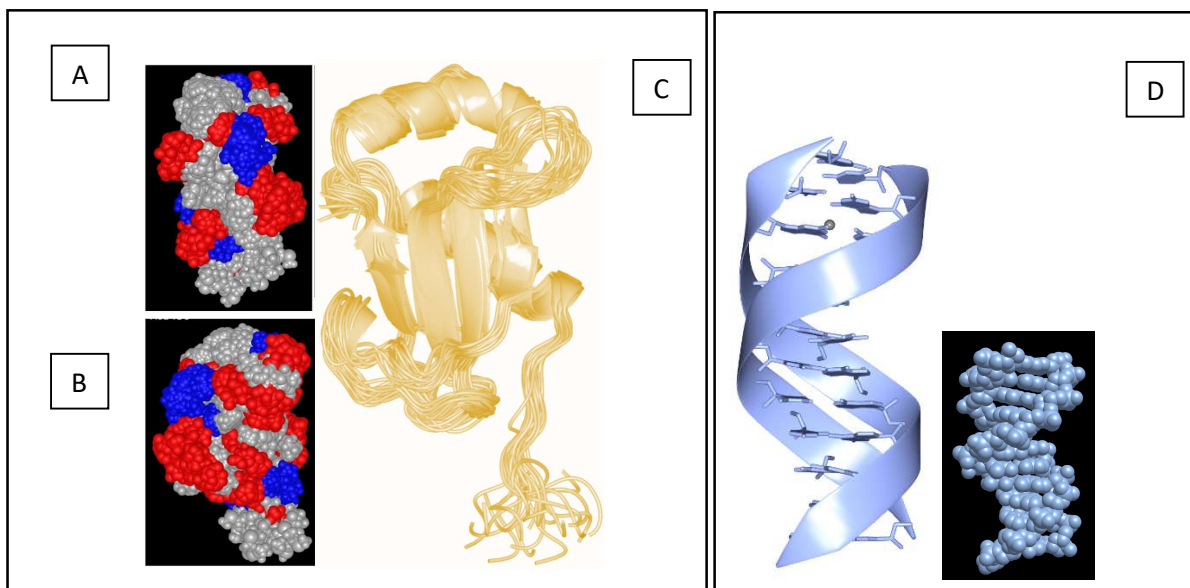
approximate repair where the sequence differs from the original sequence by 1-13 nucleotides. This heavily suggests the mechanism is unrelated to HR and instead relates to NHEJ. The difference in the ligase used between the eukaryotic and prokaryotic mechanisms for NHEJ is interesting and something that we wish to explore. We want to perform a group of growth experiments to investigate the effect of Gam on growth in the presence of DSBs using different Gam homologues in the presence and absence of specific ligases.

Unpublished data from our lab showed that the HiGam protein is produced in *H. influenzae* Rd KW20. Using a Liquid Chromatography-Mass Spectrometry (LC-MS) system with improved sensitivity, we looked for expression of the HiA and HiB genes (MuA and MuB homologues, respectively) by RT-PCR experiments which we think are co-transcribed with HiGam. MuGam has been proven to have a positive effect on cell growth (Bhattacharyya et al., 2018) – we aim to prove that the homologous HiGam protein is produced and functional within *H. influenzae*.

### 1.4 DNA mimic proteins

There are multiple DNA mimic proteins that resemble both the shape and charge of the double-helix structure of double stranded DNA, often to act as a competitive inhibitor for DNA-binding proteins. The HI1450 gene is present on the genome of *Haemophilus influenzae* Rd KW20 and encodes a protein designed to mimic the 3-dimensional structure and charge of approx. 10 bp of double stranded DNA (Parsons, Yeh and Orban, 2004). HI1450 is known to interact with the histone-like protein Hu- $\alpha$  and other homologous and non-homologous DNA mimic proteins have been shown to interact with different DNA binding proteins in an inhibitory fashion (Parsons, Liu and Orban, 2009; Putnam and Tainer, 2005). We are interested to see if DNA mimic proteins interact with Gam proteins to inhibit protein-DNA interactions.

The surface of the HI1450 protein (Fig 3A and B) shows the charge of the residues at each position around the surface of the protein. It is easy to identify bands of negative charge (in red) which imitate the structure of the negatively charged backbone of the dsDNA double helix.

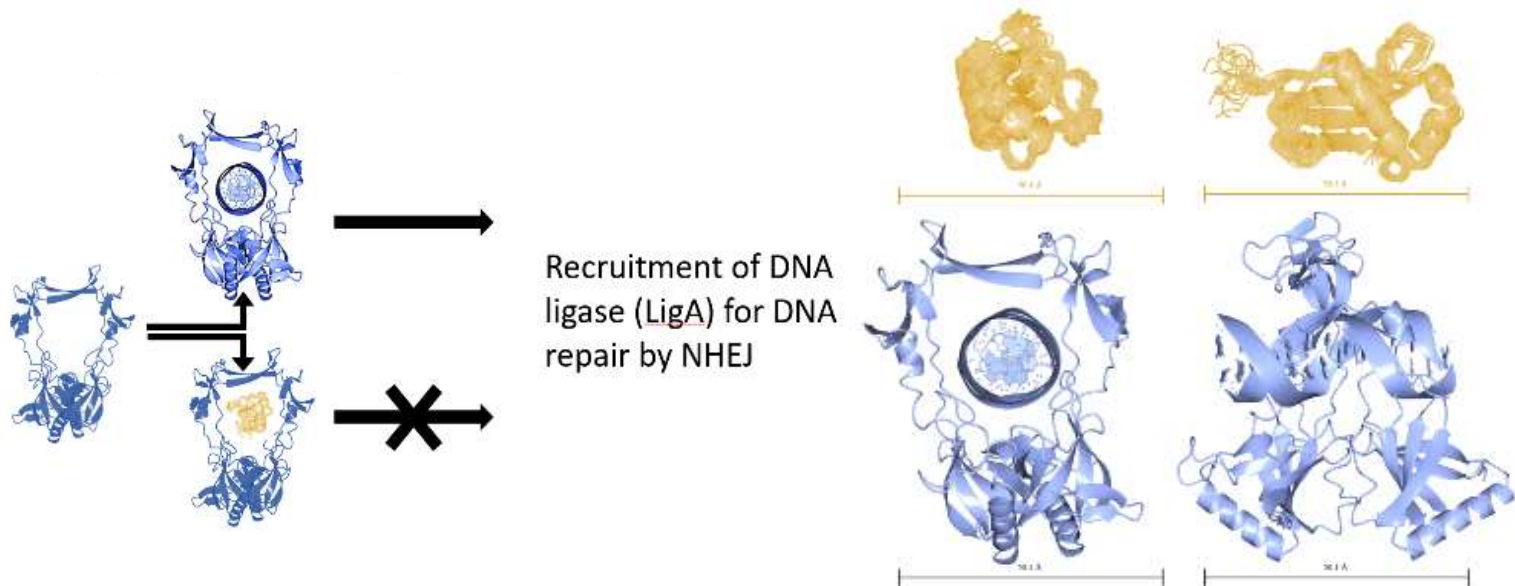


**Figure 3.** **A:** The 3-Dimensional structure of the HI1450 protein, the globular surface structure and charge (Red indicates a negative charge, blue indicates positive charge and grey indicates neutral charge) imaged using iCn3D 3D-structure viewer. **B:** same as **A**, rotated 180 degrees on the x axis. **C:** The ribbon structure imaged using CCP4mg molecular graphic viewer (approx. 50Å). The 1NNV 3D-

## Introduction

structure was solved by solution NMR (Parsons, Yeh and Orban, 2004). **D**: Double stranded DNA fragment of 12 base pairs (4C64) (approx. 40Å) (Lercher *et al.*, 2014).

The DNA mimic known to be produced in the same cells as the HiGam protein may therefore interact with the DNA binding protein resulting in the inhibition of the potential function of the HiGam protein produced in the *H. influenzae* Rd KW20 cells. **Figure 4** describes the proposed mechanism for the HI1450 inhibition of HiGam by competitive inhibition.



**Figure 4.** LHS: Schematic view of the potential mechanism for competitive inhibition of the HiGam protein (blue) by the DNA mimic HI1450 protein (yellow), the HI1450 protein would bind to the DNA binding cavity of the HiGam protein (3D structure of HiGam is based on the truncated version of the eukaryotic Ku70/80 heterodimer described in **Fig 2**). RHS: Size comparison of the two proteins with equal relative scale representation (the scale bars beneath each protein image represent 50.1 Å).

### 1.5 Aims and objectives

Here the potential for a positive survival benefit to the bacterial host by the production of the MuGam homologue proteins are investigated. This thesis describes investigations of the following: confirmation of HiGam protein production in the host cell (**7. Results– HiGam protein production in *Haemophilus influenzae* Rd KW20**), ubiquity of the pairing of an intact MuGam homologue gene with a LigA homologue gene as evidence of potential interactions with the DNA ligase protein (**3. Results – DNA Double strand break repair in Bacteria containing MuGam homologue genes**), and probing the function of a wide variety of different MuGam homologues from different bacterial strains (**6. Results – Characterisation of DNA Binding Properties for Gam Homologues**). The 3D structure of a MuGam homologue protein is not known, though the models for the dimeric structure shown in **Figure 2** are investigated further and their limitations discussed (**4. Results – 3D structure analysis of MuGam homologues**). Finally, the potential interaction between the MuGam homologue protein HiGam and the DNA mimic protein HI1450 is tested experimentally to determine whether HI1450 binds to HiGam to act as a competitive inhibitor (Chapter 2).

## Introduction

Secondary structure predicting algorithms are used to produce a clearer picture of the 3D structure of the MuGam homologue proteins. The results are used to discuss the viability of the two models for MuGam protein dimer 3D structures (based on the known crystallographic structure of the monomeric units of DvGam **Figure 2A** and the truncated Ku70/80-based structural homology model for MuGam **Figure 2C**).

Liquid chromatography-mass spectrometry was used to identify the proteins present in the total cell lysate of the *H. influenzae* Rd KW20 cells to search for the MuGam homologue protein. HiGam was found to be preferentially produced over other viral proteins from the Mu prophage such as HiA and HiB. In previous research the HiGam protein was identified but it was not possible to identify any other proteins from the Mu prophage genes. Improvements in the LC-MS method have made it possible to identify proteins at much lower concentrations more reliably. This provides clear evidence that HiGam is produced, while the other proteins (HiA and HiB) involved in DNA transposition are produced at much lower levels or were not detected.

The genomes of a variety of species identified to contain a gene for a MuGam homologue were searched to identify different DNA ligases. This work focussed on LigA (identified as functioning with MuGam to increase the DNA repair potential in bacterial host cells by Bhattacharyya et al. (2018)) as well as LigD (a bacterial protein with sequence similarity to the Lig IV protein which human Ku70/80 recruits to sites of double strand breaks in DNA to promote DNA repair by NHEJ (Brissett and Doherty, 2009; Fattah *et al.*, 2010; Pannunzio, Watanabe and Lieber, 2018) If the MuGam protein is an orthologue to the Ku70/80 protein and acts in a similar mechanism to repair DNA by NHEJ then LigD may be a more appropriate ligase to use to investigate the function of MuGam to promote DNA repair.

The function of HiGam as a linear double stranded DNA binding protein has been shown previously. Eight new MuGam homologue proteins are investigated in order to show that the function of the MuGam homologues conserved across the wide variety of bacterial species function in a similar way despite variations in their primary sequences. This could be used as evidence of the widespread conservation of the gene being linked to protein function and therefore to a role in providing a survival benefit to the host bacterial cells. Variations in the binding characteristics of the different MuGam homologues are discussed and linked to differences in their primary sequences.

The interaction between HiGam and HI1450 would be a significant discovery, as it would suggest that DNA mimic proteins could act as competitive inhibitors to the MuGam homologue protein. This would have implications for the mechanism suggested for the MuGam homologue role in NHEJ DNA repair mechanism proposed by Bhattacharyya et al. (2018). If an interaction is found, the protein can be used to target MuGam homologues to act as an inhibitor to DNA repair by NHEJ provided to the host by MuGam homologues. The interaction could also be utilised to produce a structure capable of being solved through X-ray crystallography. To this point the dimeric 3D structure of a MuGam homologue has not been solved and crystallisation using purified HiGam alone and HiGam with a short sequence of double stranded linear DNA has not proven successful.

# Materials and Methods

## Chapter 2

## 2.1 Buffers and media

Lysis Buffer 1	Lysis Buffer 2	Lysis Buffer 3	Lysis Buffer 4
0.8mM HEPES pH 7.5, 20mM NaP pH7.7, 500mM NaCl, 0.5% Triton X-100, 1mg/mL Lysozyme.	50mM HEPES pH 7.5, 100mM KCl, 15mM NaCl, 10mM MgCl <sub>2</sub> , 10% (v/v) Glycerol, 0.5% (v/v) Triton X-100, 1ug/mL Leupeptin, 1ug/mL Peptidin, 1mg/mL Lysozyme, 20ug/mL DNase I, 20ug/mL RNase A.	20mM Sodium Phosphate Buffer pH7.7, 500mM NaCl, 10mM imidazole, 1% (v-v) Triton X-100.	20mM Sodium Phosphate Buffer pH7.7, 500mM NaCl, 10mM imidazole, 5mM β-mercaptoethanol.

**Table 1.** Composition of each Lysis buffers used throughout the methods in sections 2.8 and 2.9 Lysis Buffer 1, Lysis Buffer 2 and Lysis Buffer 3 are used to isolate soluble protein from E. coli K-12 MG1655 by cell lysis and Lysis Buffer 4 to prepare soluble cell lysate for Ni<sup>2+</sup> affinity selection.

Elution Buffer	Wash Buffer	Dialysis Buffer	Storage Buffer
20mM Sodium Phosphate Buffer pH 7.7, 500mM NaCl, 500mM imidazole, 5mM β-mercaptoethanol.	20mM Sodium Phosphate Buffer pH7.7, 300mM NaCl, 20mM imidazole.	50 mM Tris-HCl, pH 7.4 (at 25°C) (pH 8 at 4 °C), 200 mM NaCl, 0.5 mM EDTA, 0.1% (w/v) β-mercaptoethanol.	50 mM Tris-HCl, pH 7.4 (at 25°C) (pH 8 at 4 °C), 200 mM NaCl, 0.5 mM EDTA, 1 mM DTT.

**Table 2.** Composition of the other buffers used in sections 2.8 and 2.9. Elution Buffer is used to elute the bound proteins from the Ni<sup>2+</sup> affinity column, Wash Buffer is used prior to the Elution buffer to wash weakly bound proteins from the Ni<sup>2+</sup> column. The Dialysis Buffer and Storage Buffer are used to reduce the NaCl concentration of the purified soluble protein for long term storage.

Lysogeny broth - Miller (LB)	Brain Heart Infusion Broth (BHI)
10 g/L Tryptone	5 g/L Beef heart (infusion from 250g)
5 g/L Yeast extract	12.5 g/L Calf brains (infusion from 200g)
10 g/L NaCl	2.5 g/L Na <sub>2</sub> HPO <sub>4</sub>

**Table 3.** Composition of media components for which are dissolved in ultra-pure H<sub>2</sub>O (resistivity ≈ 18 MOhm•s). The media could be used to create agar for bacterial growth on plates by adding a further 15g/L agar.

## 2.2 Primers

### Putative Double stranded DNA-mimic from *Haemophilus influenzae* - HI1450 (324bp)

**Forward primer JA001** – 5'GCCGC**CATATG**ACAACAGAAATTA~~AAAAA~~ACTTGATCCCG<sup>3'</sup>

**Reverse primer JA002** – 5'GCGC**GGATC**CTTACTTTTTCAAATAACGTGAAATTCGC<sup>3'</sup>

The restriction sites for NdeI (JA001) and BamHI-HF (JA002) are highlighted in bold were utilised for digestion and ligation.

### Host-nuclease inhibitor Gam family protein from *Rhodobacter capsulatus* SB 1003 – RcGam (543bp)

**Forward primer RKG9** - 5'GGGG**CATATG**ACGGTGTCTAACAGGTCC<sup>3'</sup>

**Reverse primer RKG10** - 5'GGGG**CTAAGC**TTACACCCCTCCGGCAG<sup>3'</sup>

Restriction sites for NdeI and BlnI are highlighted in bold.

### *Pseudomonas aeruginosa* LESB58 – PaGam (378bp)

**Forward primer RKG11** - 5'GGGG**CATATG**GCCGAGGAAGTCAGCA<sup>3'</sup>

**Reverse primer RKG10** - 5'GGGG**AAGC**TTTACGCCTCCTGCTTGGTGAGC<sup>3'</sup>

Restriction sites for NdeI and HindIII are highlighted in bold.

## 2.3 Measuring concentrations of DNA, protein and cell density using Absorbance spectroscopy

### 2.3.1 Estimating DNA concentration using absorbance spectroscopy

DNA absorbs light at wavelength of 260nm ( $A_{260}$ ). An estimate of DNA concentration ([DNA]) is measured using UV-visible absorbance at wavelengths 230, 260, 280 and 320nm with a NanoDrop® ND-1000 UV visible spectrophotometer using 2µL of the DNA sample based on the determined relationship of  $A_{260} = 1$ , [pure dsDNA] = 50µg/mL. The range of accuracy for the NanoDrop sensor is an absorbance between 0.1-1.0 and therefore dilutions may need to be made for the DNA sample before measuring to ensure accurate estimation of [DNA]. The [DNA] can then be calculated from the equation:

$$[DNA] (\mu g/ml) = (A_{260} - A_{320}) \times dilutionfactor \times 50\mu g/ml$$

The  $A_{260}/A_{280}$  ratio should be approximately 1.6 or greater, a lower value suggests a level of contaminants that may impact the results of experiments using the DNA sample. Values of absorbance at  $A_{230}$  and  $A_{320}$  are measured to detect other impurities, the  $A_{260}/A_{230}$  ratio should be greater than 1.5 and the  $A_{320}$  measures solution turbidity and should be close to 0.

Using the appropriate a NanoDrop® ND-1000 software UV-Vis setting was selected, 2µL of Nuclease-Free water (Qiagen) was applied directly to the NanoDrop sensor, the arm was lowered carefully and the wash step confirmed. Once complete the arm was lifted and both it and the sensor were wiped gently using a tissue to remove the Nuclease-Free water. 2µL of the blank solution (Solution used to elute DNA during purification) was applied to the sensor, arm lowered and absorbance measured using the “Blank” setting to measure the background absorbance of the solution the DNA was present in. The Arm and sensor were wiped again and finally 2µL of the DNA sample is applied to the sensor, arm lowered and absorbance measured using the “Measure” setting.

### 2.3.3 Estimating Protein concentration using absorbance spectroscopy

Estimation of purified protein concentration ([protein]) is determined by measuring  $A_{260}$ ,  $A_{280}$  and  $A_{320}$  on a NanoDrop® ND-1000 UV visible spectrophotometer. Protein Absorbs UV-visible light at a wavelength of 280nm and [protein] was calculated using the equation:

$$[Protein](M) = \frac{(A_{280} - A_{320})}{Protein\ extinction\ coefficient\ (M^{-1}cm^{-1}) \times path\ length(cm)}$$

The ratio of  $A_{280}/A_{260}$  should be approximately 1.6 or greater as an indication of sufficient sample purity and  $A_{320}$  should be close to 0 to indicate level of solution turbidity. The method for measuring absorbance is the same for measuring [DNA], the blank solution should be a sample taken from the **Storage Buffer** taken post-dialysis of the purified protein sample being measured.



Protein Name	Extinction coefficient ( $M^{-1}cm^{-1}$ ) for monomeric protein sequence
Im9-HiGam	18450
Im9-ApGam	23950
Im9-KpGam	22460
Im9-SsGam	23950
Im9-SeGam	23950
Im9-EcGam	23950
Im9-RcGam	16960
Im9-PaGam	26470
HI1450	19630

**Table 4.** Extinction coefficients ( $M^{-1}cm^{-1}$ ) used to calculate the monomeric protein concentration using the equation above.

### 2.3.3 Estimating cell density using absorbance spectroscopy

The optical density (OD) of a bacterial cell culture is measured using a Jenway® 6300 spectrophotometer at a wavelength of 600nm ( $OD_{600}$ ). 1mL of the sterile media (same media used to grow cell culture) was transferred to a 1mL cuvette and measured using the spectrophotometer using the calibration setting to measure the background absorption of the sterile media. 1mL of the cell culture was transferred to a separate 1mL cuvette and  $OD_{600}$  was recorded once the value remained constant.

### 2.4 Synthesised genes

The Gam homologue genes for KpGam, ApGam, SsGam, SeGam, NmGam and EcGam were synthesised by Genewiz® (GENEWIZ Germany GmbH [www.genewiz.com/en-GB/](http://www.genewiz.com/en-GB/)) from nucleic acid code obtained from NCBI Genome database ([www.ncbi.nlm.nih.gov/genome](http://www.ncbi.nlm.nih.gov/genome)). These gene sequences were altered for optimised for expression in *Escherichia coli* (*E. coli*) K-12 using codon optimising tool from Novo Pro® (NovoPro Bioscience Inc. [www.novoprolabs.com/tools/codon-optimization](http://www.novoprolabs.com/tools/codon-optimization)) as well as with restriction sites for NdeI at the beginning and BamHI-HF at the end of the gene sequence. These synthesised genes were then integrated into pUC-57-Amp vector and shipped to our lab. Therefore, the recombinant plasmid could be used to transform *E. coli* DH5- $\alpha$  and be digested using the NdeI and BamHI-HF restriction endonucleases for each DNA sequence that encodes a MuGam homologue, effectively skipping the protocols detailed in 2.5 and 2.6.1.

### 2.5 PCR target gene amplification

Using genomic DNA isolated from *Haemophilus influenzae* Rd KW20 the HI1450 gene fragment was amplified using primers designed to anneal to either end of the gene on the forward and reverse strands (2.2). The KOD Hot-start DNA polymerase originates from *Thermococcus kodakaraensis* KOD1 DNA polymerase expressed in *E. coli*. The KOD PCR system is designed to amplify DNA from crude sources of DNA with minimal processing such as with chromosomal DNA in this case.

The exact conditions of the PCR needed to be investigated and altered through testing. The optimal conditions were:

**5 $\mu$ L** 10x KOD PCR buffer (Sigma-Aldrich inc.)

**1 $\mu$ L** 10mM dNTPs

**2 $\mu$ L** 25mM MgCl<sub>2</sub>

**1.5 $\mu$ L** 10 $\mu$ M Forward primer (JA001)

## Materials and Methods

- 1.5µL** 10µM Reverse primer (JA002)
- 5µL** 60ng/µL Template DNA (*H. influenzae* Rd KW20 gDNA)
- 1µL** KOD PCR DNA Polymerase (1U/ µL)
- 33µL** ultra-pure water (resistivity ≈ 18 MOhm•s)
- 50µL** Total volume

Using the following thermocycler conditions (Bio-Rad T100™ Thermal Cycler):

- Stage 1:** Initial melting at 94°C for 2 min,
  - Stage 2:** melting at 94°C for 15s,
  - Stage 3:** annealing and replication at 72°C for 15s.
- Repeat stages 2 and 3 through 40 cycles
- Stage 4:** Final polymerisation step at 72°C for 1 min
  - Stage 5:** Hold at 4°C.

Once the PCR was finished, DNA polymerase was inhibited by the addition of 1µL 0.5M EDTA. A 5µL sample of this PCR product is mixed well with 1µL 6x purple DNA loading dye and analysed by 1% (w/v) agarose gel electrophoresis. Using the PCR conditions above a band appeared at the expected 324bp so the experiment was repeated with five reaction mixes of 50µL producing a large quantity of the amplified gene. These PCR-product mixes were pooled together and cleaned up using the Wizard® SV PCR clean up system (Promega Corporation) to produce a concentrated stock of the linear amplified HI1450 gene.

### 2.6 Construction of expression vector with target gene

For each MuGam homologue the gene expression system used Isopropyl β- D-1-thiogalactopyranoside (IPTG) induction of the pQE2 expression plasmid in *E. coli* K-12 MG1655. The expression vector uses a lac operon/expression system. The T5-phage promoter is recognized by host (*E. coli* K-12 MG1655) RNA polymerase, binding is inhibited by lac repressor protein binding at lac operator sequence (adjacent upstream to T5 promoter), IPTG binds to the lac repressor protein resulting in dissociation from the lac operator and allowing the host RNA polymerase to bind to the T5 promoter. The T5 promoter is very strong providing a very high rate of transcription and resulting in overexpression of the target gene. (The QIAexpressionist™ handbook 2002;Rosano and Ceccarelli, 2014).

A previously prepared expression vector (pQE2-Im9-HiGam) was used to overproduce recombinant HiGam with an N-terminal Im9 solubility tag (His<sub>7</sub>-Im9-HiGam). The recombinant His<sub>7</sub>-Im9-HiGam (Im9-HiGam) protein was purified using the same methods as described below.

When cloning to overproduce the HI1450 protein the pET15b expression system in *E. coli* BL21(DE3) was used to reproduce the successful recombinant protein production observed previously in (Parsons, Liu and Orban, 2009). This system also uses IPTG to induce gene expression and overproduction of the target gene though it uses the T7 expression system.

The pET15b vector utilizes IPTG activated lac repressor dissociation from lacUV5 promoter (modified version of the classic lac promoter) to result in transcription of a T7 RNA polymerase gene found in the chromosomal DNA of the *E. coli* BL21(DE3) host strain. The

resulting overexpression of the T7 polymerase leads to transcription of the target gene on the pET15b plasmid which is flanked by a T7 promoter. (pET System Manual 1999; Rosano and Ceccarelli, 2014).

### **2.6.1 Cloning with Zero Blunt® TOPO®**

The NdeI enzyme cuts at only 0-20% efficiency when the restriction site is 0-2 base pairs from the end of a linear fragment of double stranded DNA (New England BioLabs Inc.). To avoid inefficient DNA cutting when producing linear fragments of the target gene from the amplified PCR products (**2.6**), the PCR product of the gene was transformed into the Zero Blunt® TOPO® (Thermo Fisher Scientific) plasmid. The target gene insert can then be cut from the recombinant HI1450-Topo plasmid after being transformed to a competent strain for amplification to ligate with the expression vector.

The raw PCR product is cleaned up using the Wizard® SV PCR clean up system (Promega Corporation) to produce a single DNA band when analysed by gel electrophoresis at the expected length for the 324bp. This should remove some of the unwanted DNA including free dNTPs from the PCR reaction. The concentration of the resulting solution was estimated from UV-vis absorption at 260nm ( $A_{260}$ ) and 320nm ( $A_{320}$ ) wavelengths measured from 2 $\mu$ L of the PCR product in a Nanodrop machine (**2.3.1**). This was then diluted to approx. 22 $\mu$ g/mL and ligated with Zero Blunt Topo plasmid, using the raw PCR product did not produce the expected recombinant plasmid.

**1 $\mu$ L** PCR product 22 $\mu$ g/mL

**1 $\mu$ L** Salt Solution (1.2 M NaCl, 0.06 M MgCl<sub>2</sub>)

**3 $\mu$ L** Ultra-pure water (resistivity  $\approx$  18 MOhm $\cdot$ s)

**1 $\mu$ L** pCR II-Blunt Topo 10ng/ $\mu$ L

**(6 $\mu$ L** Total volume)

The solution was mixed by vortexing, briefly centrifuged at 17,000xg and left to incubate at room temperature (20-22°C) for 30 min after which was stored at -20°C.

### **2.6.2 Transformation to competent amplification strain *E. coli* DH5- $\alpha$ by heat shock**

Glycerol stocks of *E. coli* DH5- $\alpha$  were thawed without allowing them to warm to greater than 4°C. 5 $\mu$ L of the recombinant Zero Blunt® TOPO® plasmid and 50 $\mu$ L of the *E. coli* DH5 - $\alpha$  stock was transferred to a chilled 2mL tube for each target gene. Without mixing, the tube was placed on ice for 20 min, then transferred to a heating block at 42°C for 45s and finally back to ice for a further 2 min. 450 $\mu$ L of pre-warmed Lysogeny broth (LB) was added to the heat shocked cell suspension and incubated at 37°C for 2h shaking at 220rpm.

After incubation the cell suspension was plated onto an LB agar plate with 50 $\mu$ g/mL kanamycin added by spreading 100 $\mu$ L of the cell suspension. The remaining cell suspension was centrifuged at 10,062xg for 2 min, without disturbing the pellet 300 $\mu$ L of the supernatant was removed and discarded. The cell pellet was resuspended in the remaining

supernatant then spread onto a second LB agar plate with 50µg/mL kanamycin. The plates were incubated upside down at 37°C overnight to produce colonies of transformed cells.

### 2.6.3 Plasmid digestion using restriction endonucleases

A single transformant bacterial colony generated using procedure in section 2.6.2 was used to inoculate a separate 10mL volume of LB media supplemented with 50µg/mL kanamycin and incubated overnight at 37°C shaking at 220rpm in order to grow up a culture for small-scale plasmid DNA purification. After incubation the cultures were centrifuged at 6,000xg for 15 min at 4°C, the supernatant was discarded without disturbing the pellet and the plasmids were purified using the QIAprep Spin Miniprep Kit (QIAGEN).

Target gene name	Restriction endonuclease 1	Restriction endonuclease 2	Restriction Digest Buffer	Enzyme cutting efficiency in chosen buffer
PaGam	NdeI	HindIII	10x NEB 2.1	100%
RcGam	NdeI	BlnI	10x NEB 2.1	100%
MuGam homologues Kp, Ap, Ss, Se, Nm, Ec	NdeI	BamHI-HF <sup>®</sup>	10x CutSmart <sup>®</sup>	100%
HI1450	NdeI	BamHI-HF <sup>®</sup>	10x CutSmart <sup>®</sup>	100%

**Table 5.** Restriction enzymes and buffers used for each plasmid digestion reaction with the expected efficiency for the enzymes in the buffer used.

Both the recombinant amplification plasmid and the expression plasmid are digested using the corresponding restriction endonucleases (RE) to create complementary sticky ends for efficient DNA ligation.

Appropriate volumes of the purified plasmids were used such that at least 2µg of plasmid DNA is present in the mixture alongside the appropriate cutting buffer and each enzyme (20U) with nuclease-free water added to a final reaction volume of 50µL. The mixture is vortexed, centrifuged at 17,000xg for approx. 10s and incubated overnight at 37 °C.

### 2.6.4 Ligation of gene-encoding DNA fragment into expression vector

The resulting mixture from the double RE digest was separated on ultrapure 1% w/v agarose-TAE gel and electrophoresis done at 50mA (constant current) for 90 min. After this the expected bands were excised from the agarose gel using a scalpel and isolated into separate weighed 2ml microcentrifuge tubes. These excised gel slices were processed using the DNA using the Wizard<sup>®</sup> SV PCR clean up system resulting in linear plasmid DNA and gene encoding DNA with complementary sticky ends.

Appropriate volumes of the linear pQE2-Im9 plasmid and Gam gene insert were mixed such that there was a 2:1 molar ratio of insert DNA over the vector plasmid and a final DNA concentration of 5µM. This was mixed with 1µL T4 DNA Ligase 400,000 U/mL (400 units) and 2.5µL 10x Ligation buffer with PCR water to bring the total volume to 20µL. The mixture could then be vortexed, spun down and incubated at room temperature overnight.

## **2.7 Transformation to expression strain by high efficiency electroporation or heat shock**

The MuGam homologue constructs (pQE2-Im9-Gam would be used to transform the E. coli K-12 MG1655 cultures by electroporation as they are not naturally competent. The pET15b-HI1450 construct was used to transform E. coli BL21 (DE3) cultures by heat shock.

Liquid overnight cultures were inoculated with a single colony selected from a LB agar plate which had been streaked using glycerol stocks of the bacterial strains stored at  $-80^{\circ}\text{C}$ . A single colony was used to inoculate 15mL of pre-warmed LB which was incubated at  $37^{\circ}\text{C}$  with shaking at 220rpm for 2 hours or until the OD600 reached 0.5-0.7. The culture was placed immediately on ice for 25 minutes then centrifuged at 3,400xg for 20 minutes, the supernatant was carefully removed under sterile conditions without disturbing the pellet. All subsequent steps were done under sterile conditions unless stated otherwise. 15mL of sterile ice-cold ultra-pure water was added and used to resuspend the pellet, this was then centrifuged again at 3,400xg and supernatant carefully removed using a pipette. This wash procedure was repeated an additional two times. The supernatant from the final wash was removed and enough water was left to transfer the cell pellet to a chilled 2ml microcentrifuge tube. The cell suspension was centrifuged at 12,000xg for 10 minutes at  $4^{\circ}\text{C}$ , and the supernatant was removed using a pipette. The pellet was resuspended in 700 $\mu\text{L}$  of the sterile ice-cold ultra-pure water and split into two 350 $\mu\text{L}$  samples which were then mixed with 1-5 $\mu\text{L}$  of DNA ligation reaction mix.

1-5 $\mu\text{L}$  of DNA ligation reaction mix ( $\sim 0.07\mu\text{M}$  final DNA molecular concentration) was used to transform 350 $\mu\text{L}$  of electro-competent cells using a MicroPulser (BioRad) electroporator and pre-set parameters for Escherichia coli (2.49kV and time constant = 5.4-5.7 ms). The cell suspension was transferred to a fresh 2mL microcentrifuge tube and incubated at  $37^{\circ}\text{C}$  for 2 hours shaking at 220rpm. 100 $\mu\text{L}$  of the resulting cell suspension was spread on a LB agar plate with 100 $\mu\text{g}/\text{mL}$  ampicillin added for selection of transformants. The remainder of the electroporated cell suspension were concentrated by centrifuging as in **2.6.2** and 100 $\mu\text{L}$  of the resulting concentrated cell suspension was spread onto a separate plate. Isolated colonies should contain the plasmid with the cloned gene as linearised plasmid DNA has non-complementary ends. Transformation should only occur if the gene insert has been successfully ligated into the plasmid.

## **2.8 Diagnostic confirmation of successful transformation**

To determine if the expected gene was correctly integrated into the expression vector and correctly transformed into the bacterial expression strain two separate assays were used: diagnostic PCR and double-digestion of plasmid purified from transformed cell culture with appropriate restriction endonucleases. The colonies isolated from the ampicillin selection plates were used to inoculate cultures grown up overnight in 10mL LB media supplemented with 100 $\mu\text{g}/\text{mL}$  ampicillin. Plasmid DNA was isolated and purified from these cultures as described in section **2.6.3**. Diagnostic PCR was not possible for the synthesised MuGam homologue genes as we did not have appropriate primers for these genes, so only a diagnostic double digest was used to determine if the transformation was successful.

### **2.8.1 Diagnostic PCR using GoTaq<sup>®</sup> DNA Polymerase (M300)**

To determine if the expected gene was correctly integrated into the expression vector and correctly transformed into the bacterial expression strain two separate assays were used: diagnostic PCR and double-digestion of plasmid purified from transformed cell culture with appropriate restriction endonucleases. The colonies isolated from the ampicillin selection plates were used to inoculate cultures grown up overnight in 10mL LB media supplemented with 100µg/mL ampicillin. Plasmid DNA was isolated and purified from these cultures as described in section 2.6.3. Diagnostic PCR was not possible for the synthesised MuGam homologue genes as we did not have appropriate primers for these genes, so only a diagnostic double digest was used to determine if the transformation was successful.

**10µL** 5U/µL GoTaq buffer (Promega)  
**1µL** 10mM dNTPs  
**5µL** 25mM MgCl<sub>2</sub>  
**5µL** 10mM of the forward PCR primer  
**5µL** 10mM of the reverse PCR primer  
**22.5µL** ultra-pure water (resistivity ≈ 18 MOhm•s)  
**0.5µL** GoTaq-G2 DNA polymerase (Promega)  
(**50µL** total reaction volume).

Using the following thermocycler conditions (Bio-Rad T100™ Thermal Cycler):

**Stage 1:** Initial melting at 94°C for 2 min,  
**Stage 2:** melting at 94°C for 30s,  
**Stage 3:** annealing and replication at 72°C for 75s.  
Repeat stages 2 and 3 through 30 cycles  
**Stage 4:** Final polymerisation step at 72°C for 5min  
**Stage 5:** Hold at 4°C.

Once complete, 1µL of 0.5M EDTA was added to the samples to chelate the Mg<sup>2+</sup> and they were analysed by 1% (w/v) agarose-TAE gel electrophoresis using a HyperLadder 1kb Plus DNA ladder (New England Biolabs Inc.®) as a molecular weight marker.

### 2.8.2 Diagnostic Digest

The purified recombinant plasmids were enzymatically digested overnight at 37°C using the restriction endonucleases and buffers appropriate for each cloned gene (**Table 5**).

The existing recombinant pQE2-Im9-HiGam plasmid was used with a volume such that approximately 200ng of DNA was present. Each sample was cut with the enzymes that correspond to the gene insert that will be used in ligation.

## 2.9 Target protein overproduction and purification

### 2.9.1 Tests to determine ideal bacterial growth and induction conditions for over-production of recombinant MuGam-homologue proteins

In order to optimise conditions for over-production of recombinant Im9-RcGam and Im9-PaGam proteins in *E. coli* MG1655, the concentration of IPTG, length of post-induction growth period and post-induction growth temperature were varied while monitoring total

protein and total soluble protein yields. The results from these tests would determine the conditions used for subsequent MuGam homologue protein overproduction experiments to produce optimal yield.

Four 250mL conical flasks (labelled A, B, C and D) with 50mL sterile LB media supplemented with 0.4% (w/v) glucose and 100µg/mL ampicillin were inoculated with enough pQE2-Im9-Gam transformed MG1655 cells for a starting OD600 = ~0.03 from 10mL overnight culture grown up under the same conditions as described in section 2.5.3. These were then incubated at 37°C with shaking at 180 rpm until an OD600 = ~0.6 was obtained. A 1 mL pre-induction control was removed and centrifuged at 6,000x g for 5 minutes at 4°C, the supernatant removed and cell pellet stored at 4°C for later analysis.

Each bacterial culture was then induced to over-produce the MuGam homologue protein by adding IPTG to a 0.1mM (flasks A and C) or 5µM final concentration (flasks B and D), then incubated with shaking (180 rpm) at either 20°C (A and C) or 37°C (B and D). The media was again supplemented to give final concentrations of 0.4% (w/v) glucose and 100µg/mL ampicillin to ensure continued antibiotic selection and rich bacterial growth.

After 2 hours the OD600 was measured by removing a 1mL sample, which was centrifuged at 6,000x g for 5 minutes at 4°C, followed by removal of the supernatant and storage of the cell pellet as before. This was repeated at 4 hours and 20 hours post-induction.

To analyse the collected 1mL samples by SDS-PAGE, the cells were lysed and fractionated to assess the content of both the soluble and insoluble protein fractions. Two separate lysis buffers were used in this procedure (**Table 1: Lysis buffer 1** and **Lysis buffer 2**) as the chromosomal DNA had to be intact (not sheared) when separating the soluble and insoluble cell fractions to more accurately assess the amount of soluble recombinant MuGam homologue protein present in the cells. If the chromosomal DNA was fragmented, it is possible for soluble recombinant MuGam protein to bind to the resulting linear DNA fragments. If this occurs, the recombinant protein will reside in the cell debris pellet after centrifugation, rather than the supernatant, due to co-pelleting with the chromosomal DNA.

The pellets were resuspended in **Lysis buffer 1**, mixed well and incubated at 30°C for 20 minutes with periodic mixing every 5 minutes. The amount of **Lysis buffer 1** added to each cell pellet was calculated using the following equation:

$$\text{Volume of Lysis Buffer 1 } (\mu\text{L}) = \frac{(\text{Volume of cell culture (mL)} \times \text{OD600})}{20} \times 1000$$

### 2.9.2 Large-scale over-production and extraction of recombinant protein using lysozyme and Triton X-100 for bacterial cell lysis

The cell pellets from an induced, overnight 500mL bacterial culture were resuspended in 35mL of chilled Lysis Buffer 3 and transferred to a 50mL screw-cap plastic tube, followed by addition of 350µL 100 µM phenylmethylsulfonyl fluoride (PMSF – serine protease inhibitor) PMSF, 350µL 10mM Benzamidine and 350µL 1mg/mL Bacitracin (protease inhibitors) and

20mg Lysozyme (hen egg white, Sigma). The cell suspension was mixed well by pipetting while trying to avoid excessive bubble formation. The cell suspension was then incubated at 37°C for 20 min with periodic mixing every 5 min.

The lysed cell suspension was transferred to a chilled plastic screw-cap centrifuge tube (Nalgene) and centrifuged at 20,000xg for 30 minutes at 4°C. The supernatant was carefully removed with a pipette without disturbing the pellet or gelatinous layer of chromosomal DNA above the pellet. The supernatant was filtered through a 0.45µm pore size syringe filter (manufacturer) and transferred to a clean 50mL screw-cap plastic tube and stored on ice prior to loading onto Ni<sup>2+</sup>-affinity column.

### 2.9.3 HI1450 protein overproduction

The overproduction of HI1450 from pET15b-HI1450 transformed *E. coli* BL21(DE3) was performed the same as in 2.8.1 though only temperature was investigated as a variable for production optimization and Lysis buffer 2 was used throughout the lysis protocol as concern for DNA damage and protein-DNA interactions were not a concern.

### 2.9.4 Diagnostic confirmation of protein production by HisPur purification

During test overproductions it was unclear whether the expected HI1450 protein was being overproduced as the protein band produced did not appear to show at the correct molecular weight. To investigate whether the protein was produced correctly the HisPur purification system was used as a relatively quick method to determine if the protein that had been overproduced contained the 6xHis tag and therefore corresponded to the HI1450 protein.

A 100mL culture of *E. coli* BL12(DE3) with HI1450-pET15b in LB media with 0.4% (w/v) final concentration D-Glucose and 100ng/mL ampicillin was grown to an OD<sub>600</sub> of 0.6 and production of the HI1450 protein was induced using a 1mM final concentration of (IPTG) and allowed to incubate on a 180rpm shaker at 37°C for 4h. The culture was centrifuged at 6,000xg for 15 min at 4°C to harvest cells prior to purification of the recombinant protein.

The pellet is resuspended in 12mL of **Lysis Buffer 3** then 120µL of 100mM PMSF is added to the mixture and allowed to incubate on ice with mixing throughout to avoid clumping and forming bubbles. Add lysozyme to a final concentration of 1mg/mL to the sample suspension as well as another 120µL PMSF.

The cell suspension was sonicated at 40W 10s on and 30s off for a total of 5 times or until the suspension appears to be lysed. Add a few flakes of solid DNase I to the cell suspension and another 120µL 100mM PMSF and incubate for a further 30 min on ice.

The lysed cell suspension was centrifuged at 15,000xg for 15 min at 4°C, once complete carefully transfer the supernatant into a fresh sterile 50mL falcon tube through a 0.2µm syringe filter.

Two Ni-NTA spin columns were transferred to a 50mL falcon screw-cap plastic tube and the stopper is removed, the storage solution must be removed by centrifugation at 700xg for 3 min at 4°C. The spin column is then equilibrated with 6mL lysis buffer, the solution is centrifuged through the columns as previously into a collection tube.



## Materials and Methods

The approximately 12mL of supernatant/cell lysate is split equally between two columns with the stoppers on and transferred to a rotating wheel for 45 min to homogenise the Ni-NTA resin with the supernatant from the induced cells.

Once finished the spin columns are removed and transferred to a fresh collection tube labelled "Flow through" ensuring that the stopper is removed. Centrifuge the solution through the spin columns at 700xg for 3 min at 4°C to remove most proteins from the resin leaving only the 6xHis tagged target protein and perhaps any others that have formed weak interactions with the resin and contain consecutive and/or neighbouring His residues.

These proteins that have formed weak non-specific bonds with the resin should be removed using **Wash Buffer (Table 2)** as it contains a greater concentration of Imidazole but not enough to elute the target protein, 6mL of the wash buffer is added to each spin column and centrifuged as previously into a fresh collection tube labelled "Wash".

Finally, 3mL of the **Elution Buffer (Table 2)** is added to the spin column and the column is transferred to a fresh sterile 50mL screw-cap plastic tube labelled as "Elution 1". This is centrifuged at 700xg for 3 min at 4°C to elute the target protein into the collection tube. This should then be repeated twice more for a total of three elution samples.

This protocol was repeated without the use of Lysozyme during cell lysis. Since Lysozyme has a similar molecular weight to HI1450 we could isolate which bands belong to which protein.

### **2.9.5 Large-scale over-production and extraction of recombinant protein using lysozyme and Triton X-100 for bacterial cell lysis".]**

The cell pellets from an induced, overnight 500mL bacterial culture were resuspended in 35mL of chilled Lysis Buffer 3 and transferred to a 50mL screw-cap plastic tube, followed by addition of 350µL PMSF 100 µM, 350µL 10mM Benzamidine and 350µL 1mg/mL Bacitracin (protease inhibitors) and 20mg Lysozyme (hen egg white, Sigma). The cell suspension was mixed well by pipetting while trying to avoid excessive bubble formation. The cell suspension was then incubated at 37°C for 20 min with periodic mixing every 5 min.

The lysed cell suspension was transferred to a chilled plastic screw-cap centrifuge tube (Nalgene) and centrifuged at 20,000xg for 30 minutes at 4°C. The supernatant was carefully removed with a pipette without disturbing the pellet or gelatinous layer of chromosomal DNA above the pellet. The supernatant was filtered through a 0.45µm pore size syringe filter (manufacturer) and transferred to a clean 50mL screw-cap plastic tube, and stored on ice prior to loading onto Ni<sup>2+</sup>-affinity column.

### **2.9.6 Recombinant protein purification by Ni<sup>2+</sup>-affinity chromatography**

Using a peristaltic pump, a HisTrap™ Ni<sup>2+</sup> affinity column was washed (through a 0.45 µm pore syringe filter) using the following sequence of solutions: 10 ml of ultra-pure water, 5 ml of 100 mM NH<sub>4</sub>Acetate pH 4, 10 ml of ultra-pure water, 2.5 ml of 100 mM NiSO<sub>4</sub> and 10 ml of ultra-pure water. The HisTrap™ Ni<sup>2+</sup> affinity column was connected to a Fast Protein Liquid Chromatography (FPLC) instrument (Biologic DuoFlow, Bio-Rad)) and washed with

100% **Lysis Buffer 4** (Table 1) at a flow-rate of 1 ml/min (back pressure was 9-10 psi) until the absorbance ( $\lambda = 280$  nm) of the column effluent reached a constant baseline value. The protein sample was loaded onto the FPLC and the protein-containing flow-through collected and stored at 4°C. The absorbance ( $\lambda = 280$  nm) of the column effluent was measured. Once all protein was loaded, the 100% Lysis buffer was used to wash the column until the baseline absorbance was constant. The column was subsequently washed with 92% **Lysis Buffer 4**/8% **Elution Buffer** (Table 2) to remove any non-specifically bound protein.

Load the filtered supernatant from the cell lysis (**2.9.1**) at 1mL/min onto the column and collect the flow through once the absorbance (Abs280) starts to climb, keep this labelled "Flow through Load" once the supernatant has all been loaded wash the column through with Lysis Buffer 4 until the Abs280 reaches baseline or ceases to continue to decrease and stop collecting the flow through. Wash the column through with 92% Lysis Buffer 4 and 8% Elution Buffer (50mM Imidazole) and collect "Flow through Wash" once the Abs280 increases again. As Before wash through until Abs280 reaches baseline or no longer decreases, stop collecting the flow through and wash through with 100% Lysis Buffer 4.

Elute the trapped protein from the column into 40 1mL fractions using increasing Imidazole concentration by steadily increasing the percentage of Elution Buffer from 0-100% at constant rate. The Abs280 will peak at the point at which the His-tagged Im9-HiGam is eluted. Gather the 1mL fractions that fall within this peak and analyse them by SDS-PAGE. Based on the SDS-PAGE analysis, the fractions containing the most recombinant protein were pooled and transferred to 3.5kDa molecular weight cut off (MWCO) dialysis tubing (manufacturer) and dialysed against 1L of Dialysis buffer (Appendix) while stirring the buffer with a magnetic stirrer. The pooled fractions were left to dialyse in the Dialysis buffer overnight (18 hours), then transferred to 1L of Storage buffer (Appendix) and left stirring overnight on a magnetic stirrer.

After this period the solution is transferred from the dialysis tubing to a 10mL screw-cap plastic centrifuge tube and stored at 4°C.

## **2.9.7 Purification of target protein**

### **2.9.7.1 Gam homologue proteins**

The grow ups of each *E. coli* MG1655 culture containing each Gam homologue gene were scaled up to 500mL and induced to overproduce the Im9-Gam homologue using the same protocol as described in **2.9.1**. The optimised conditions for protein production were decided to be 0.1mM IPTG induction concentration at a starting  $OD_{600} = 0.60 - 0.70$  and incubation overnight at 20°C on 180 rpm shaker. After this incubation period the cultures were centrifuged at 6,000xg for 15 min at 4°C to form cell pellets which were then stored at -20°C. At pre-induction and after 2h, 4h and overnight incubation post-induction the  $OD_{600}$  was measured and a 1mL sample was collected for analysis by SDS-PAGE as in **2.9.2 - 2.9.3**.

The pellets from 500mL culture are resuspended in 35mL Lysis Buffer 3 and transferred to a 50mL centrifuge tube, add 350 $\mu$ L 100mM PMSF, 350 $\mu$ L 10mM Benzamidine and 350 $\mu$ L 1mg/mL Bacitracin

## Materials and Methods

and 20mg Lysozyme. Mix well by pipetting avoiding forming excessive bubbles. Incubate the mixture in 37°C for 20 min mixing every 5 min.

Centrifuge the mixture at 20,000xg for 30 min at 4°C, without disturbing the pellet or gelatinous layer of chromosomal DNA above remove the supernatant with a pipette, filter the supernatant through 0.45µM syringe filter and transfer to a clean 50mL centrifuge tube, store on ice.

Wash the column through at 1mL/min with 5mL deionised water (dH<sub>2</sub>O) then with 5mL 100mM Ammonium acetate pH4, again with 5mL dH<sub>2</sub>O, with 2mL 100mM Nickel Sulphate in water (until the column turns noticeably blue in colour and finally with 5mL dH<sub>2</sub>O. Avoid introducing bubbles to the column throughout and once complete load the column to the FPLC to be washed through with water then 10mL Lysis Buffer 4.

Load the filtered supernatant from the cell lysis (2.9.1) at 1mL/min onto the column and collect the flow through once the absorbance (Abs280) starts to climb, keep this labelled "Flow through Load" once the supernatant has all been loaded wash the column through with Lysis Buffer 4 until the Abs280 reaches baseline or ceases to continue to decrease and stop collecting the flow through. Wash the column through with 92% Lysis Buffer 4 and 8% Elution Buffer (50mM Imidazole) and collect "Flow through Wash" once the Abs280 increases again. As Before wash through until Abs280 reaches baseline or no longer decreases, stop collecting the flow through and wash through with 100% Lysis Buffer 4.

Elute the trapped protein from the column into 40 1mL fractions using increasing Imidazole concentration by steadily increasing the percentage of Elution Buffer from 0-100% at constant rate. The Abs280 will peak at the point at which the His-tagged Im9-HiGam is eluted. Gather the 1mL fractions that fall within this peak and analyse them by SDS-PAGE.

The fractions containing the desired protein are pooled and transferred to 3.5kDa molecular weight cut off (MWCO) dialysis tubing, seal off the edges with clips and submerge the tubing in 1L of Dialysis buffer (Appendix) while the buffer mixes with magnetic stirrer. The tubing is left in the Dialysis buffer overnight (18h) then transferred to 1L of Storage buffer (Appendix) and left overnight mixing on a magnetic stirrer.

After this period the solution is transferred from the dialysis tubing to a 10mL centrifuge tube and stored at 4°C.

### 2.9.7.2 Over-production and purification of HI1450 recombinant protein

A 500mL liquid culture of *E. coli* BL21(DE3) with HI1450-pET15b in LB media with 0.4% final concentration D-Glucose and 100ng/mL ampicillin was grown to an OD<sub>600</sub> of 0.6 and induced production of the HI1450 protein using a final concentration of 1mM and incubated at 37°C on a 180rpm shaker for 4h. The culture was centrifuged at 6,000xg for 15 min and the pellet resuspended in 35mL of **Lysis Buffer 3**. Add 350µL of 100mM PMSF and incubate on ice for 30 min, add another 350µL of 100mM PMSF and sonicate (without adding lysozyme) at 102W, 7.0 amplitude for 6x 15s with 30s cool-down periods between sonication. Once the cells are lysed add a few flakes of DNase I and 350µL of 100mM PMSF and incubate for 30 min on ice, finally centrifuge at 20,000xg for 30 min at 4°C, filter and collect the supernatant through a 0.45µm syringe filter.

Load the filtered supernatant from the cell lysis (2.8.1) at 1mL/min onto the column and collect the flow through once the absorbance (Abs280) starts to climb, keep this labelled "Flow through Load" once the supernatant has all been loaded wash the column through with Lysis Buffer 4 until the Abs280 reaches baseline or ceases to continue to decrease and stop collecting the flow through.

## Materials and Methods

Wash the column through with 92% Lysis Buffer 4 and 8% Elution Buffer (50mM Imidazole) and collect "Flow through Wash" once the Abs280 increases again. As Before wash through until Abs280 reaches baseline or no longer decreases, stop collecting the flow through and wash through with 100% Lysis Buffer 4.

Elute the trapped protein from the column into 40 1mL fractions using increasing Imidazole concentration by steadily increasing the percentage of Elution Buffer from 0-100% at constant rate. The Abs280 will peak at the point at which the His-tagged Im9-HiGam is eluted. Gather the 1mL fractions that fall within this peak and analyse them by SDS-PAGE.

The fractions containing the greatest majority of HI1450 proteins based the SDS-PAGE analysis are pooled together and cleaned up by Dialysis with 3.5kD MWCO film into Dialysis Buffer (Appendix) and mixed well overnight at 4°C. The dialysis film containing pooled fractions is then transferred to the Storage Buffer (Appendix) and mixed overnight. The pooled solution is then filtered through a 0.22µm filter and stored at 4°C.

### 2.10 Analyse DNA binding properties by EMSA

#### 2.10.1 Standard EMSA for positive control with Im9-HiGam

Samples for the EMSA were prepared by adding appropriate volumes of 100mM Sodium Phosphate buffer pH7 to separate 0.2 mL PCR tubes according to **Table 6** (Samples 1-7) followed by 2µL of 50µg/mL 500bp DNA, and finally the appropriate volume of 30µM Im9-HiGam (with DTT added to final concentration 1mM DTT). The samples were then mixed well and incubated at 20°C for 20 minutes. After the incubation step, 5µL of 6x Purple DNA loading dye without added SDS (manufacturer) was added to each tube and the samples were mixed. The samples are then loaded onto a 1.5% (w/v) agarose-TAE gel (without SYBR Safe added) and electrophoresed at 50 mA (constant current) for 90 minutes.

The agarose gel was then removed from the electrophoresis tank and placed into a plastic lunch box containing SYBRsafe staining solution (prepared in TAE) for 20 minutes in a dark environment. The agarose gel was destained in ultra-pure water for ~20 min prior to visualising the SYBRsafe-stained DNA using a Gene Genius imaging system (SynGene) and a UV-transilluminator. SYBR-safe prevents Im9-HiGam from binding to the DNA, thus the agarose gel was stained after the electrophoresis step.

#### 2.11 Competitive binding assay

Using the same procedure as in section **2.9.1**, the EMSA analysis was performed to compare the DNA-binding properties of Im9-HiGam-linear in the presence and absence of the HI1450 DNA mimic protein. This experiment was done to determine if the Im9-HiGam and HI1450 proteins form a protein-protein interaction which competes with linear DNA binding by Im9-HiGam.

The samples listed in **Table 6** were prepared by adding the specified volumes of 100mM Sodium phosphate buffer pH7 to separate 0.2 mL PCR tubes, followed by 2µL linear 500bp DNA (50µg/mL) and the appropriate volumes of 30µM Im9-HiGam dimer and 255 µM

## Materials and Methods

HI1450. The samples were then mixed well by vortexing, centrifuged briefly at 17,000xg and incubated at 20°C for 20 minutes.

Samples 8-21 were prepared by adding the Sodium phosphate buffer pH7 to separate 0.2 mL PCR tubes, followed by aliquots of the HI1450 and Im9-HiGam protein. The samples were then incubated at 20°C for 20 minutes to enable any protein-protein interactions to form. After this incubation step, the DNA was added and the mixture was incubated for another 20 minutes at 20°C. Agarose gel electrophoresis, staining and imaging were done as per section **2.10.1**.

Sample Label	100mM Sodium Phosphate Buffer pH 7	50µg/mL 500bp DNA (µL)	30µM Im9-HiGam (+ 1mM DTT) vol. (µL)	255µM HI1450 vol. (µL)
1	13.0	2	0.0	0.0
2	12.5	2	0.5	0.0
3	12.0	2	1.0	0.0
4	11.5	2	1.5	0.0
5	11.0	2	2.0	0.0
6	10.0	2	3.0	0.0
7	9.0	2	4.0	0.0
8	11.5	2	0.0	1.5
9	11.0	2	0.5	1.5
10	10.5	2	1.0	1.5
11	10.0	2	1.5	1.5
12	9.5	2	2.0	1.5
13	8.5	2	3.0	1.5
14	7.5	2	4.0	1.5
15	9.0	2	4.0	0.0
16	8.5	2	4.0	0.5
17	8.0	2	4.0	1.0
18	7.5	2	4.0	1.5
19	7.0	2	4.0	2.0
20	6.5	2	4.0	2.5
21	6.0	2	4.0	3.0

**Table 6.** Standardised reagent volumes used in EMSA experiments.

## 2.12 Bioinformatic study of 3D structure comparison

The 3-Dimensional structure of the MuGam homologue from *Desulfovibrio vulgaris* (DvGam) has been determined by X-Ray diffraction at 2.57Å. The unpublished work “Crystal structure of putative host-nuclease inhibitor protein Gam from *Desulfovibrio vulgaris*” Bonano et. al (2007) shows the solved structure for the monomeric protein and models a possible dimeric structure. A model for the MuGam protein of a truncated version of the Eukaryotic Ku heterodimer (Ku70/80), based on suggested sequence similarity (d’Adda di Fagagna *et al.*, 2003), looks very different from the solved structure of DvGam. Ku70/80 is involved in DNA repair by the Non-homologous end joining (NHEJ) pathway in Eukaryotes such as Humans (Aravind and Koonin, 2001). Determining the predicted secondary structure elements of the HiGam protein sequence would provide information on the potential validity of the two models. Bioinformatic tools including Phyre2 (Kelley *et al.*, 2015), SWISS-MODEL (Schwede *et al.*, 2003), HHpred (Söding, Biegert and Lupas, 2005) and Jpred 4 (Drozdetskiy *et al.*, 2015) were used to predict secondary structure elements based on primary sequence alignments with proteins with existing solved 3-D structures on the RCSB Protein Data Bank (PDB).

### 2.12.1 Using HiGam protein sequence to predict secondary structure elements

The FASTA format amino acid (aa) sequence of the HiGam protein was obtained from the NCBI Genome database for *Haemophilus influenzae* Rd KW20 – Mu-like host-nuclease inhibitor protein Gam. This was used as the search query in Phyre2 with the modelling mode set at normal, in SWISS-MODEL with the default settings using the “Build Model” search, in Jpred 4 with default settings. In HHpred parameters used were as follows: MSA generation method: HHblits=>UniRef30, Maximal no. of MSA generation steps: 3, E-value incl. threshold for MSA generation:  $1e^{-3}$ , Min. seq. identity of MSA hits with query (%): 0, Min. coverage of MSA hits (%): 20, Secondary structure scoring: during\_alignment, Alignment Mode:Realign with MAC: local:norealign, MAC realignment threshold: 0.3, No. of target sequences (up to 10000): 250, Min. probability in hit list (> 10%): 20.

This process was also repeated with the amino acid sequences of DvGam and the truncated sequences of Human Ku70 and Ku80 proteins using the same search parameters as listed above.

### 2.12.2 Target template structure alignment

Using the FASTA amino acid sequence of HiGam as in 2.12.1 as the query in the “Expert Mode” of Phyre2 with alignment method: local, Secondary structure scoring: yes, Secondary structure weight: 0.1. With the uploaded 3D structure of DvGam (PDB - 2P2U) as the target template structure, the tool would align the sequences of the two proteins and attempt to produce the secondary structure elements from the target template in the query sequence.

This was again repeated for the Human Ku70 and Ku80 protein target template structures from PDB – 1JEY as well as the truncated structures for both Ku70 and Ku80. The truncated 3D structures of the Ku70 and Ku80 proteins containing only the DNA binding domains were produced by using the truncated primary amino acid sequence for each (Aravind and Koonin, 2001), Ku70 aa: 262-446 and Ku80 aa: 257-443 as the query sequences for basic Phyre2 search, this produced a unique 3D structure identical to the original 1JEY for the truncated sequence only.

### 2.13 Mass spectrometry analysis of natural protein production in *H. influenzae* Rd KW20

To determine whether HiGam, HiA and HiB proteins are produced naturally within *H. influenzae* Rd KW20 cells at detectable levels Liquid Chromatography-Mass Spectrometry was used to identify the three target proteins in whole cell lysates.

#### 2.13.1 *H. influenzae* Rd KW20 cell culture growth

*H. influenzae* Rd KW20 colonies were streaked from glycerol stocks onto plates of BHI + Bacto medium (37 g/L BHI and 15 g/L Bacto) with 0.01mg/mL hemin and 0.2 ng/mL Nicotinamide adenine dinucleotide (NAD) and incubated at 37°C overnight or until colonies started to form (typically at least 24 hours). A single colony was picked and used to inoculate a 5 mL volume of BHI media with 0.01 mg/mL hemin and 0.2 ng/mL NAD in a 50mL screw cap centrifuge tube and incubated overnight at 37°C on 220 rpm shaker. The OD<sub>600</sub> was measured and appropriate volumes were used to inoculate four 10mL volumes of BHI media with 0.01 mg/mL hemin and 0.2 ng/mL NAD in 50mL screw cap centrifuge tubes to give a starting OD<sub>600</sub> ≈ 0.02 and incubated overnight at 37°C. The OD<sub>600</sub> of the cultures are measured and recorded after incubation (aiming for actual OD<sub>600</sub> ≈ 1). The four tubes were balanced and centrifuged at 6,000xg at 4°C for 15 min, supernatant was removed and discarded carefully and dry cell pellets stored at -20°C for cell lysis.

#### 2.13.2 Cell Lysis using Lysozyme

One pellet was chosen, thawed and resuspended in 300µL Phosphate-Buffered Saline (PBS buffer) pH 7.4 with EDTA to a final concentration 0.5 mg/mL. If the cells are not fully resuspended add a further 700µL with 0.5 mg/mL EDTA. Once the cells were fully resuspended lysozyme is added to a final concentration 0.5 mg/mL. The cell suspension was incubated at 37°C for 30 min mixing gently every 5 min by inverting the tube. The tube was transferred to ice for 5 min and Sodium dodecyl sulfate (SDS) is added to a final concentration of 1% (w/v). The cell suspension was centrifuged at 20,000xg at 4°C for 10 min, supernatant was removed by pipetting and transferred to a fresh 1.5 mL microcentrifuge tube and centrifuged again at 20,000xg at 4°C for 10 min. The supernatant was removed by pipetting and transferred to a fresh 1.5 mL microcentrifuge tube representing the soluble fraction of the cell lysate. Aliquots of the soluble cell lysate were analysed by SDS-PAGE on 15% (w/v) polyacrylamide gel. The gel electrophoresis was performed at 180V through the stacking gel and 120V through the resolving gel until the dye front runs out of the gel. The gel was stained using Coomassie-brilliant blue stain for 10 min and destained overnight in water.

#### 2.13.3 SDS-PAGE gel Digestion

One lane from the Coomassie-stained gel excised and split into 20 evenly sized fractions for proteomic analysis. In-gel tryptic digestion was performed after reduction with dithioerythritol and S-carbamidomethylation with iodoacetamide. Gel pieces were washed two times with aqueous 50% (v:v) acetonitrile containing 25mM ammonium bicarbonate, then once with acetonitrile and dried in a vacuum concentrator for 20min. Sequencing-grade, modified porcine trypsin (Promega) was dissolved in 50mM acetic acid, then diluted 5-fold with 25mM ammonium bicarbonate to give a final trypsin concentration of 0.02µg/µL. Gel pieces were rehydrated by adding 25µL of trypsin solution, and after 10 min enough 25mM ammonium bicarbonate solution was added to cover the gel pieces. Digests were incubated overnight at 37°C. Peptides were extracted by washing three times with aqueous 50% (v:v) acetonitrile containing 0.1% (v:v) trifluoroacetic acid, before drying in a vacuum concentrator and reconstituting in 50µL of aqueous 0.1% (v:v) trifluoroacetic acid.

#### 2.13.4 LC-MS/MS

## Materials and Methods

Peptides were loaded onto an mClass nanoflow UPLC system (Waters) equipped with a nanoEase M/Z Symmetry 100Å C<sub>18</sub>, 5µm trap column (180µm x 20mm, Waters) and a PepMap, 2µm, 100Å, C<sub>18</sub> EasyNano nanocapillary column (75 µm x 500 mm, Thermo). The trap wash solvent was aqueous 0.05% (v:v) trifluoroacetic acid and the trapping flow rate was 15µL/min. The trap was washed for 5 min before switching flow to the capillary column. Separation used gradient elution of two solvents: solvent A, aqueous 1% (v:v) formic acid; solvent B, acetonitrile containing 1% (v:v) formic acid. The flow rate for the capillary column was 300nL/min and the column temperature was 40°C. The linear multi-step gradient profile was: 3-10% B over 7 min, 10-35% B over 30 min, 35-99% B over 5 min and then proceeded to wash with 99% solvent B for 4 min. The nanoLC system was interfaced with an Orbitrap Fusion hybrid mass spectrometer (Thermo) with an EasyNano ionisation source (Thermo). Positive ESI-MS and MS<sup>2</sup> spectra were acquired using Xcalibur software (version 4.0, Thermo). Instrument source settings were: ion spray voltage, 1,900V; sweep gas, 0 Arb; ion transfer tube temperature; 275°C. MS<sup>1</sup> spectra were acquired in the Orbitrap with: 120,000 resolution, scan range: m/z 375-1,500; AGC target, 4e<sup>5</sup>; max fill time, 100ms. Data dependant acquisition was performed in top speed mode using a 1s cycle, selecting the most intense precursors with charge states >1. Easy-IC was used for internal calibration. Dynamic exclusion was performed for 50s post precursor selection and a minimum threshold for fragmentation was set at 5e<sup>3</sup>. MS<sup>2</sup> spectra were acquired in the linear ion trap with: scan rate, turbo; quadrupole isolation, 1.6m/z; activation type, HCD; activation energy: 32%; AGC target, 5e<sup>3</sup>; first mass, 110m/z; max fill time, 100ms. Acquisitions were arranged by Xcalibur to inject ions for all available parallelizable time.

### 2.13.5 Database Searching

Thermo .raw files were imported into PEAKSX Studio (Build 20181106, Bioinformatics Solutions Inc.) for peak picking, combining and database searching. Spectra were searched against the Haemophilus influenzae subset of the UniProt database (3,094 sequences; 961,643 residues), appended with expected target protein sequences. The following search parameters were specified: Parent Mass Error Tolerance, 3.0 ppm; Fragment Mass Error Tolerance, 0.5 Da; Precursor Mass Search Type, monoisotopic; Enzyme, Trypsin; Digest Mode, Specific; Fixed Modifications, Carbamidomethylation (C); Variable modifications, Oxidation (M), Deamidation (N,Q), Pyroglutamate (E,Q). Resulting peptide matches were filtered to 2.1% false discovery rate as determined in PEAKSX against a decoy database, with protein identifications further filtered to require a minimum of two unique peptide matches. Relative inter-protein abundance was estimated using extracted precursor ion areas from identified peptides.



## Results

# Chapter 3 – DNA Double strand break repair in Bacteria containing MuGam homologue genes

### 3.1. Introduction

MuGam has been shown to have sequence homology with Eukaryotic Ku proteins particularly in the DNA binding domain. Bhattacharyya et al. (2018) found that when MuGam and NAD<sup>+</sup>-dependent ligase (LigA) proteins were produced in *E. coli* K-12 MG1655 instances of DNA repair increased in the presence of phleomycin-induced DSBs. In the study they found that knocking out RecA, a protein essential for homologous recombination (HR), the presence of the MuGam protein continued to improve instances of DNA repair. They propose that the MuGam protein works with LigA to repair DNA by NHEJ. In Eukaryotes the Ku heterodimer recruits an ATP-dependent ligase to the site of a DSB to repair DNA by NHEJ, if MuGam is indeed an orthologue to human Ku it may induce a repair response with an orthologue to human ATP-dependent ligase (Lig IV) such as LigD found in many bacterial genomes. The ATP-dependent ligases of Eukaryotes and those present in bacteria are not necessarily homologous outside of the ATP-binding and adenylation domains though the similarities of these regions may in some way be significant for recruitment in NHEJ.

### 3.2. Results

#### 3.2.1 Investigating the widespread conservation of MuGam homologues in bacterial genomes and DNA ligases

The sequence alignment tool HMMER (EMBL-EBI) was used to find appropriate organisms containing a Gam-like protein in their genome. HMMER was used as it does not suffer from the same issues with redundancy as BLASTp (NCIB). Entering the amino acid sequence of HiGam to HMMER resulted in 153 hits (individual instances of significantly similar sequences present on genomes in the database) for bacterial genomes.

The top 25 genome hits (ascending order of E value) were analysed to find an example of an organism which contains a Gam-like protein but without the LigA. The genomes were searched using the Genome database (NCBI) search cross-referenced with a list of prokaryotic organisms containing a LigA homologue from UniPortKB of 49,744 organisms. An example of an organism containing a conserved homologue of Gam without LigA suggests that the MuGam homologue gene is conserved despite the lack of potential LigA gene. This would suggest a survival benefit in the absence of LigA, so MuGam homologue may interact with LigD or act independently as suggested in Bhattacharyya et al. (2018).

This search found an example of a micro-organism with a homologue to MuGam, but without a copy of the LigA gene and instead, its only functional DNA ligase is the ATP-dependent ligase (LigD). This micro-organism is *Avibacterium paragallinarum* JF4211 (*A. paragallinarum* JF4211) and its existence could suggest that LigA is not the only Ligase that is able to work with MuGam to facilitate NHEJ. It could also strengthen the connection between Gam and Ku as it may give evidence that they interact with similar Ligase proteins and may help infer a functional role. Bhattacharyya et al., 2018 found that MuGam, when overproduced in *E. coli* K-12 MG1655, would increase the rate of DSB repair without the presence of overproduced LigA, this may suggest that the MuGam is working with LigA produced from normal expression and other Ligases within the cell such as LigB as they suggest (Bhattacharyya et al., 2018).

## DNA double strand break repair in Bacteria containing MuGam homologue genes

Name of Organism	E value/similarity to HiGam	NAD+-dependent Ligase LigA	ATP-dependent Ligase LigD
<i>Haemophilus influenzae</i> Rd Kw20	0.0	✓	
<i>Escherichia coli</i> O157:H7	3.1e-110	✓	
<i>Citrobacter rodentium</i> (strain ICC168)	1.0e-108	✓	
<i>Nitrosomonas communis</i>	7.8e-53	✓	
<i>Sulfuriferula sp.</i> AH1	9.2e-53	✓	
<i>Brenneria sp.</i> EniD312	4.5e-52	✓	
<i>Nitrosomonas communis</i>	3.5e-50	✓	
Bibersteinia trehalosi USDA-ARS-USMARC-192	8.7e-49	✓	✓
Pseudomonas phage D3112	6.5e-48		
<i>Nitrospira lacus</i>	8.0e-48	✓	✓
<i>Nitrosomonas halophila</i>	9.1e-48	✓	✓
Proteobacteria bacterium CG1_02_64_396	9.7e-48	✓	
<i>Haemophilus influenzae</i> (strain ATCC 51907 / DSM 11121 / KW20 / Rd)	6.4e-47	✓	
Comamonadaceae bacterium NML00-0135	2.2e-46	✓	
<i>Nitrosomonas sp.</i> Nm166	6.7e-46	✓	
<i>Actinobacillus minor</i> NM305	8.8e-46	✓	✓
<i>Haemophilus parasuis</i> serovar 5 (strain SH0165)	2.0e-45	✓	
<i>Haemophilus sputorum</i> HK 2154	9.2e-45	✓	✓
<i>Acidovorax caeni</i>	1.6e-44	✓	✓
<i>Pasteurella bettyae</i> CCUG 2042	2.3e-43	✓	✓
<i>Conservatibacter flavescens</i>	6.4e-43	✓	
<i>Hydrogenophaga sp.</i> A37	9.0e-43	✓	
<i>Cardiobacterium hominis</i> (strain ATCC 15826 / DSM 8339 / NCTC 10426 / 6573)	9.4e-43	✓	✓
<i>Avibacterium paragallinarum</i> JF4211	3.6e-42		✓

**Table 7.** DNA Ligases present in the 25 closest sequence matches to HiGam from HMMER. This table shows that while there were multiple examples of Organisms that contained a protein homologous to HiGam, LigA and LigD in their genome, of those investigated only *Avibacterium paragallinarum* JF4211 contained both Gam and LigD homologues.

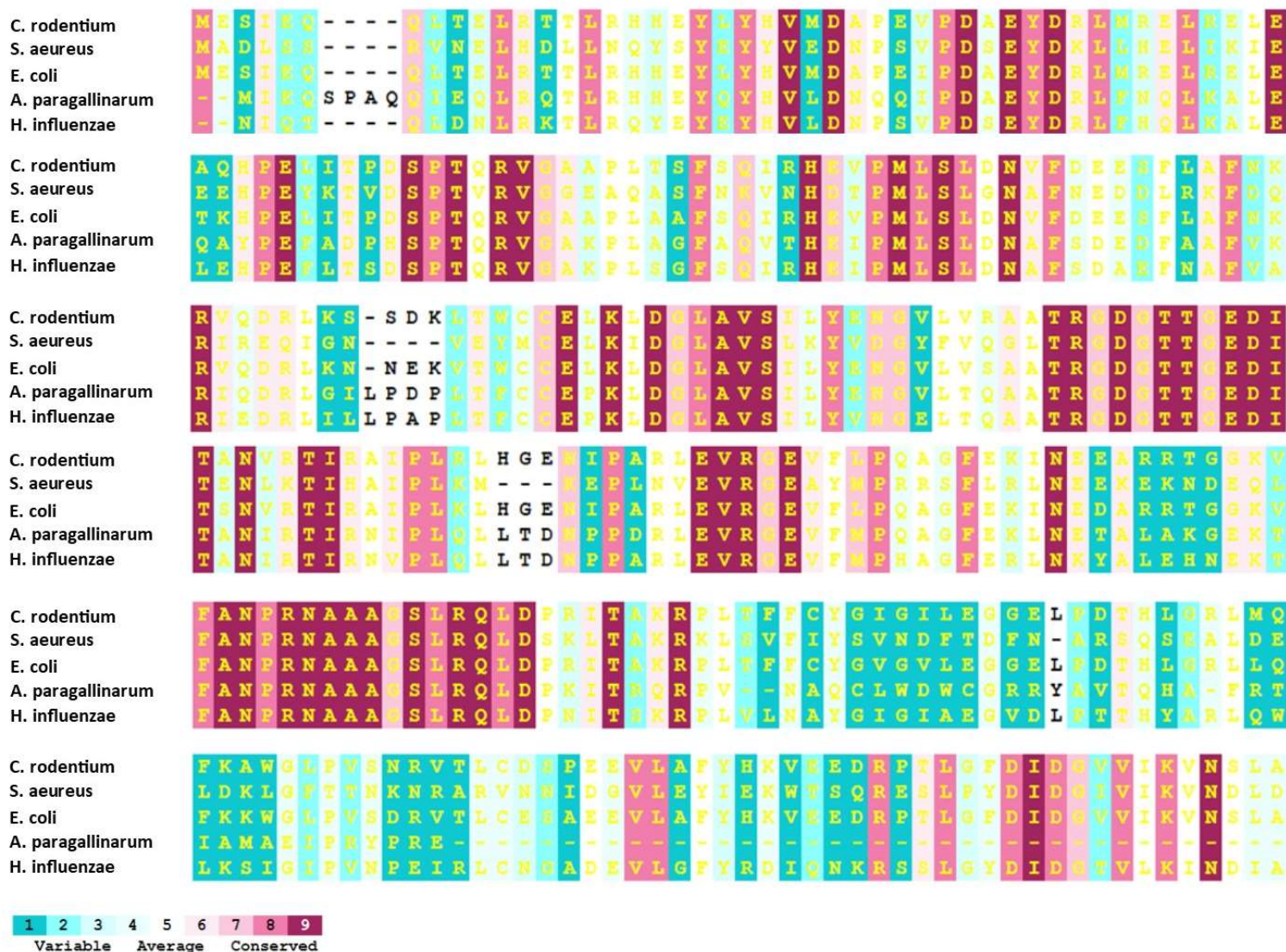
### **3.2.2 The absence of LigA in *A. paragallinarum* JF4211**

The absence of a functional LigA protein in *A. paragallinarum* JF4211 does not provide evidence that MuGam homologue proteins provide benefits for the host in the absence of LigA, it is also possible that the mutation in the LigA gene has happened recently and the Gam gene has not been purified out of the genome. However, the MuGam protein has been shown to induce NHEJ repair using other ligase proteins and further investigation into the Gam protein function in this host may provide evidence for NHEJ repair using other DNA ligases such as LigD.

Further investigation into the genome of *A. paragallinarum* JF4211 confirms that it does not contain a complete gene for LigA though the gene is present. It contains an insertion mutation 666bp into the gene that results in a premature stop codon 777bp into the gene (intact gene consists of 2025bp). Though the gene will likely produce a truncated protein it is still possible for truncated proteins to function, perhaps at a lesser capacity, which would undermine the idea that MuGam homologues can function to induce DSB repair in the absence of LigA.

### 3.2.3 Domain analysis of LigA and fragmentation of LigA protein from *A. paragallinarum* JF4211

To determine if either of the protein fragments potentially produced in *A. paragallinarum* JF4211 are functional the conserved domains of the protein were identified and mapped onto the sequences for both truncated LigA protein fragments. The bacterial LigA protein has two domains the adenylation and NAD+ binding domain

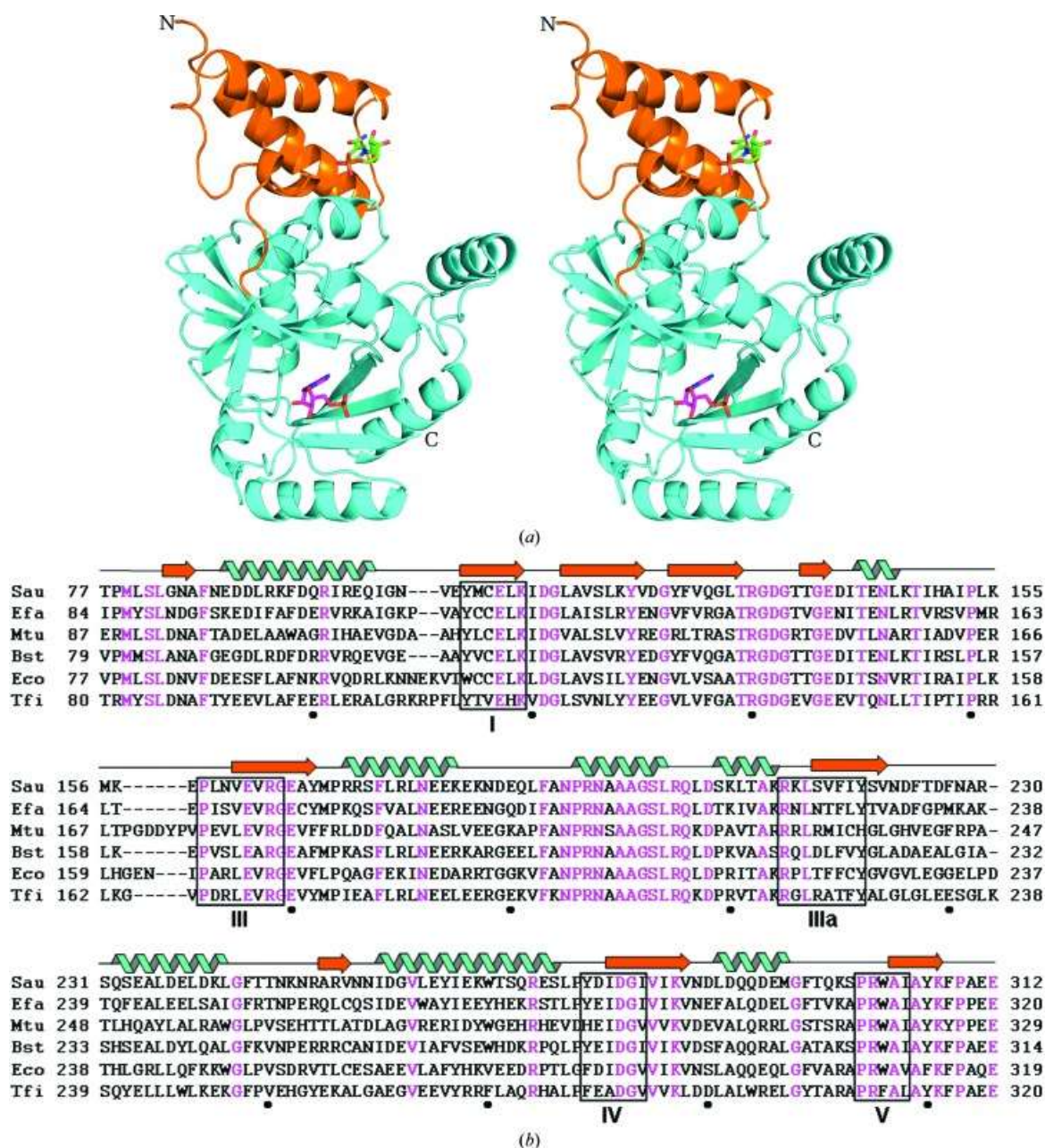


**Figure 5.** Shows the conserved residues found in the adenylation domain of the LigA proteins of five different bacteria, *Citrobacter rodentium* (strain ICC168), *Staphylococcus aureus*, *Escherichia coli* O157:H7, *Avibacterium paragallinarum* JF4211 fragment 1 and *Haemophilus influenzae* Rd KW20. It shows the conserved residues of the proteins are consistent and provides evidence that fragment 1 of the *A. paragallinarum* JF4211 LigA protein contains the majority but incomplete adenylation domain, suggesting the protein fragment would likely not be functional in its truncated state.

The shorter fragment contains a majority of the N-terminal NAD+ binding domain but it is missing many of the key amino acids pointed out in **Figure 5** that bind NAD+. It is also missing specific amino acids key to forming the binding pocket where NAD+ binds. The Lysine involved with binding NAD+ and eventually covalently binding AMP (Lys112) is present on this short fragment (Unciuleac, Goldgur and Shuman, 2017; Han, Chang and Griffor, 2009). This suggests that this truncated domain could bind NAD+ but more importantly it means that the second fragment (Fragment 2) does not

## DNA double strand break repair in Bacteria containing MuGam homologue genes

contain the potential to bind NAD<sup>+</sup> which is required for the ligase to function and provide phosphate for the repair of the DNA backbone (Lee *et al.*, 2000). The key amino acid residues found in the adenylation domain of LigA are further illustrated in **Figure 6** where key residues such as Lys112 are shown to be consistently conserved. The importance of key residues for the formation of the secondary structure elements are shown, this suggests that the LigA protein fragments that may be produced in *A. paragallinarum* JF4211 would not form correct secondary and tertiary structures due to the separation of the sequences.



**Figure 6.** Figure 1 from (Singleton *et al.*, 1999) showing the conserved residues of the NAD<sup>+</sup>-dependent DNA ligase (LigA) from different organisms (*Enterococcus faecalis* (Efa), *Mycobacterium tuberculosis* (Mtu), *Bacillus stearothermophilus* (Bst), *Escherichia coli* (Eco) and *Tiedemannia filiformis* (Tfi). It illustrates the significance of the residues that are consistently conserved on the LigA protein, in relation to the position to bind NAD as well as forming the correct secondary structures.

*A. paragallinarum* FARPER-174 was also identified as being LigA deficient. It also has a premature stop codon much further along the sequence so in theory the adenylation domain is completely conserved and produced in the cell. The truncated protein could then bind to NAD<sup>+</sup> and become covalently bound to AMP and perhaps even repair DNA if recruited to the DSB. The DNA binding domain is most likely the only part of the protein that is affected by the frameshift as it is at the C-terminus and does not take part in DNA repair mechanism. *A. paragallinarum* FARPER-174 is described as being NAD-hemin-independent as none of the cellular functions rely on NAD as a cofactor (Tataje-Lavanda *et al.*, 2019). The absence of a functional LigA gene in these two closely related species supports the notion that LigA is not required for the survival of these species, the gene for MuGam is still conserved in *A. paragallinarum* JF4211 despite the lack of LigA. The MuGam protein therefore may not be dependent on the presence of LigA to provide a survival benefit to the cell.

### 3.3. Discussion

A MuGam homologue is consistently conserved in a large number of bacterial genomes. It may contribute to recruiting DNA ligase to double strand breaks to initiate DNA repair through the NHEJ pathway. This recruitment may not be limited to the LigA protein and may also extend to the LigD protein which similarly to the Ligase IV protein in humans is ATP-dependent and only present in some bacterial genomes. The evidence for this comes from *A. paragallinarum* JF4211 as the MuGam homologue gene is conserved however a functional copy of the LigA protein cannot be produced. Therefore, it is possible that the MuGam homologue gene was conserved due to survival benefit that results from recruiting LigD to the sites of DSBs for DNA repair.

# Results

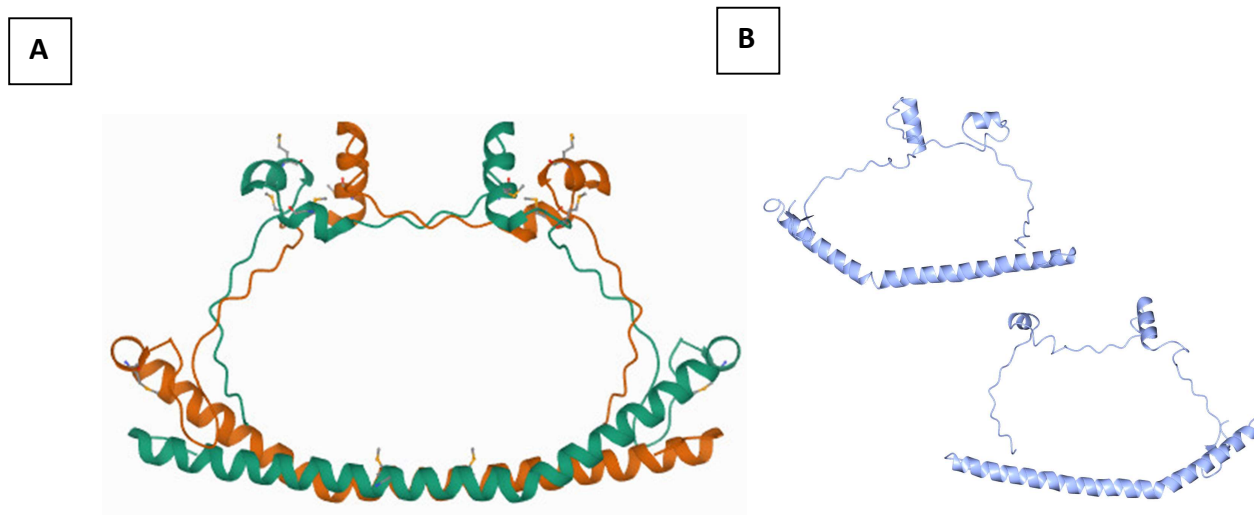
## Chapter 4 – 3D structure analysis of MuGam homologues



#### 4.1. Introduction

##### 4.1.1 3-Dimensional structure for MuGam homologue from *Desulfovibrio vulgaris*

The 3-dimensional crystal structure of one MuGam homologue protein from *Desulfovibrio vulgaris* (DvGam) (**Fig 7**) has been solved using X-ray diffraction (2P2U – Bonnano et. al 2007 - yet to be published). The structure was solved to a resolution of 2.75 Å and showed the structure of monomeric polypeptide chains which were then combined to create a theoretical dimeric protein structure (**Fig 1**).



**Figure 7.** 3-dimensional solved structure for putative host-nuclease inhibitor protein Gam from *Desulfovibrio vulgaris*. **A:** Modelled dimeric 3-D structure. **B:** Crystal structures solved by X-ray crystallography of monomeric MuGam polypeptides.

DvGam is a homologue of MuGam though the sequence alignment shows DvGam to have only 26% sequence identity and 50% sequence similarity with the MuGam protein sequence. This is comparatively low when compared to other MuGam homologues such as HiGam (61% identical residues 75% sequence similarity with MuGam). The dimeric structure of the DvGam protein was never solved, only predicted based on the structure of the monomeric state and the polypeptide sequence is significantly different from the MuGam protein and other MuGam homologues. Low sequence homology and the structural differences in the dimeric form leads us to question the validity of this model for MuGam and the MuGam homologues being investigated in this study.

##### 4.1.2 MuGam as an orthologue to Human Ku70/80

The Ku70/80 heterodimer (Ku heterodimer) is a DNA binding protein involved in the non-homologous end joining (NHEJ) DNA repair pathway in Eukaryotes. It was proposed that the DNA binding domain of the Ku heterodimer showed sequence similarities with the sequence for the MuGam protein (**Figure 8**). Though the sequence alignment suggests the similarity is low as quoted on the paper as “17% identity and 27% similarity with human Ku80, and 13% identity and 9% similarity with human Ku70” (d’Adda di Fagagna *et al.*, 2003). They suggest that the sequence homology could suggest 3D structure similarity between the MuGam dimer and the Ku heterodimer, this led them to develop the model for MuGam based on a truncated Ku70/80 3D structure.



## 3D structure analysis of MuGam homologues

Each relies heavily on multiple sequence alignment, SWISS-MODEL, HHpred, Phyre2 use the alignments to find previously solved structures and fit the sequence into multiple existing 3D structures. The closest solved 3D structure in terms of sequence similarity to MuGam homologues, such as HiGam, is DvGam which results in 100% confidence in structural similarity. This is consistent though Phyre2, SWISS-MODEL and HHpred. As discussed, the sequence alignment for MuGam with DvGam shows a low sequence similarity which suggests the potential limitation to this method of predicting structures.

Jpred 4 (Drozdetskiy *et al.*, 2015) utilises the JNet algorithm which uses multiple sequence alignment to fit discrete sequences from the query to existing solved structures independently of the overall structure of the solved proteins. This means it is more likely to predict discrete secondary structures within the query sequence rather than match the sequence to complete existing structures. Using the advanced settings it is possible to skip the sequence homology search of existing solved 3D structures for the query sequence, this may improve the prediction of discrete secondary structure elements by removing the full protein sequence homology from being the primary alignment method.

## 4.2 Results

### 4.2.1 SWISS-MODEL secondary structure predictions

The bioinformatic tool used the primary sequence for HiGam as the query and found that the existing 3D model with the greatest query coverage and sequence similarity was the structure for DvGam (PDB: 2P2U). The subsequent closest matches in order of query coverage with sequence alignment were all within the first 50% of the sequence where in each case there is either a single long  $\alpha$ -helix or a bundle of  $\alpha$ -helices. The sequence homology between HiGam and DvGam from BLASTp sequence alignment shows 26% identity with 46% similarity, comparatively low particularly when comparing homologs, e.g. HiGam and ApGam have 58% identities with 78% similarity.

### 3D structure analysis of MuGam homologues

Protein Name	3D structure PDB code	Sequence identities (%)	Coverage	Coverage region
Host-nuclease inhibitor protein Gam, putative	2P2U	29.63	92%	
Nucleosome assembly protein	2Z2R	23.08	29.89%	
protein design 2L4HC2_11	5J2L	19.30	32.76%	
Xrcc4-MYH7-(1562-1622) chimera protein	5CJ4	16.95	37.93%	
Protein SET	2E50	16.39	35.06%	
NAP1-related protein 1	5DAY	12.73	31.61%	
Nucleosome assembly protein 1, putative	3FS3	12.73	31.61%	
Nucleosome Assembly Protein	6N2G	12.50	32.18%	
NAP1-related protein 2	6JQV	12.28	32.76%	
Mitofusin domain HR2 V686M/I708M mutant	1T3J	7.58	41.95%	

**Table 8.** SWISS-MODEL output for HiGam query shows a list of the top 10 matches sorted by highest percentage of sequence alignment identities. The solved 3D structure for DvGam has the strongest sequence identity at 29.63. The subsequent structures have only the N-terminal region with any sequence similarity. All shown are either long  $\alpha$ -helix structures or helix bundles with Nucleosome assembly protein (NAP) showing up on a consistent basis.

#### 4.2.2 HHpred secondary structure predictions

The HHpred tool uses a very similar method to the SWISS-MODEL tool to align the query sequence with existing solved 3D structures in the PDB database. We see a similar trend of sequence alignment with the DvGam 3D structure showing the greatest query sequence coverage with subsequent hits with coverage in the first half of the protein sequence at the N-terminal end.


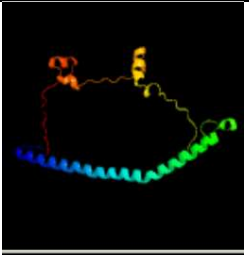

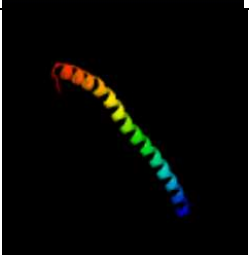

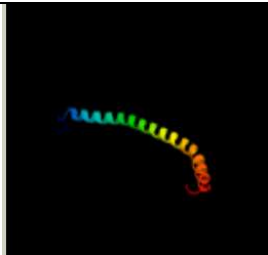

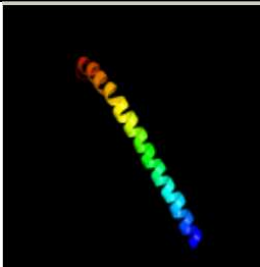

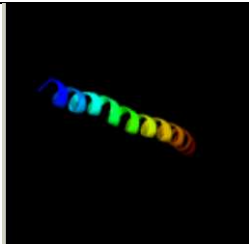


**Figure 9.** Schematic view of the top hits with DvGam as the top hit (red). The bar position and length correspond to the position and length of the region with sequence alignment. Again, the trend shows significant structural similarity to existing structures in the first half of the HiGam amino acid sequence. All these structures that align to this region have a corresponding long  $\alpha$ -helix or helix bundle that we have seen previously.

### 4.2.3 Phyre2 secondary structure predictions


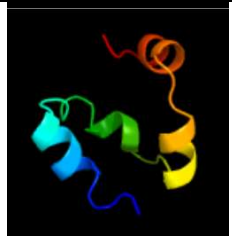





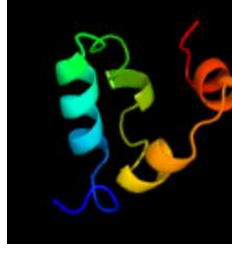

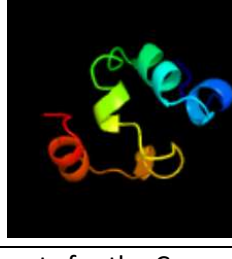
The multiple sequence alignment showed yet another unique set of proteins that align with the HiGam query sequence with a similar trend. The DvGam structure shows up as the model with the highest confidence and subsequent structures show a consistent  $\alpha$ -helix or  $\alpha$ -helical bundle.

### 3D structure analysis of MuGam homologues

Protein Name	PDB structure code	Model confidence	Alignment region	3D structure of aligned sequence
Host-nuclease inhibitor protein Gam (DvGam)	2P2U	100.0		
NAP-1 related protein 1	C5DAY	91.9		
NAP-like protein	D2E50	91.7		
Nucleosome assembly protein (NAP)	C3KYP	88.4		
Serine/Threonine-protein kinase 4	C2J08	87.7		

**Table 9.** The top 5 hits in order of model confidence from the Phyre2 bioinformatic tool with the HiGam amino acid sequence as the query. Results are consistent with what we have seen from the other prediction tools.

### 3D structure analysis of MuGam homologues

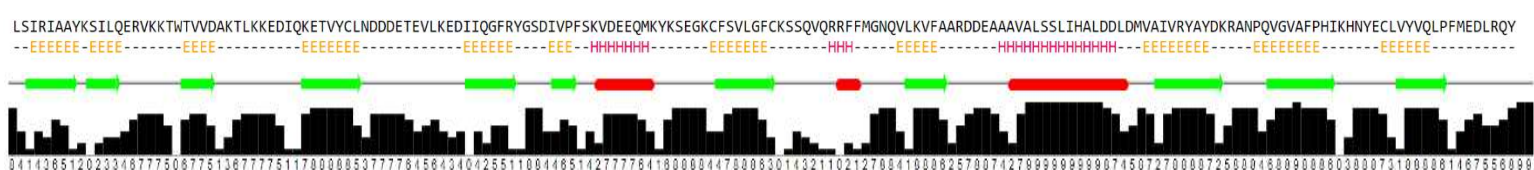
Protein Name	PDB structure code	Model confidence	Alignment region	3D structure of aligned sequence
SAM domain-like	D10XJ	53.6		
RNA-binding protein	C2B6G	50.6		
RNA-binding protein	C2FE9	48.6		
SAM domain of diacylglycerol kinase delta	C3BG7	48.4		
SAM domain of VTS1P in complex with RNA.	C2ESE	46.0		

**Table 10.** Hits 14-18 ordered by confidence (%). Shows the secondary structure alignments for the C-terminal region. The region consistently appears to be structurally similar to a Sterile alpha motif (SAM) domain and an RNA binding region. Hits 7-18 where lower percentage confidence  $\alpha$ -helices in the same region as shown in **Table 9**.

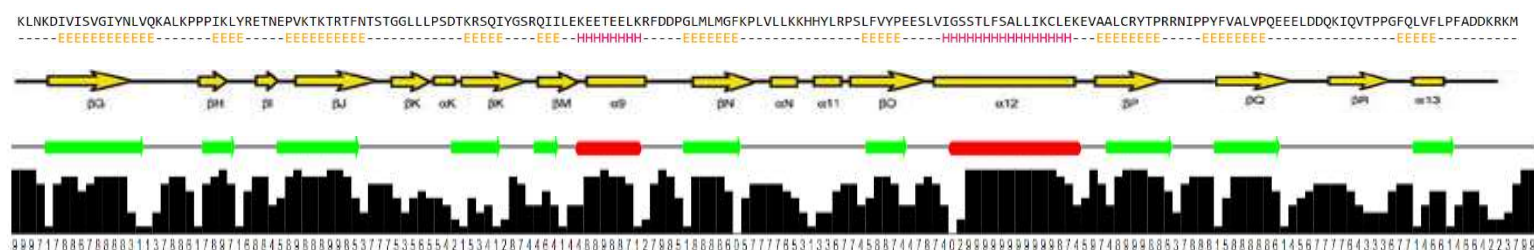
### 3D structure analysis of MuGam homologues

#### 4.2.4 JPred 4 secondary structure prediction

##### Ku80 DNA binding domain

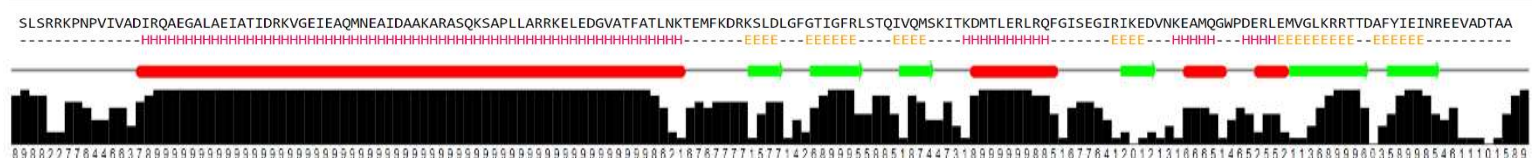


##### Ku70 DNA binding domain

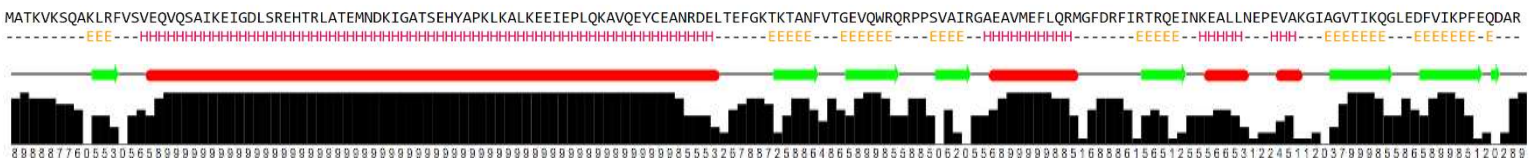


**Figure 10** Shows the JPred 4 secondary structure prediction for Ku70 and Ku80 DNA binding domains (As defined by Ku70: aa262-446, Ku80: aa 275-433 (d'Adda di Fagagna *et al.*, 2003)). Red:  $\alpha$ -helix Green:  $\beta$ -sheet, prediction confidence is shown with the assigned numbers and black bar graphs (higher value corresponds to greater confidence in modelled structure) and peptide primary sequence is shown with corresponding predicted secondary structures. Both Ku70 and Ku80 show extremely similar secondary structure regions with high confidence despite the relatively low sequence similarity (25% identities and 51% positives).

##### DvGam



##### HiGam



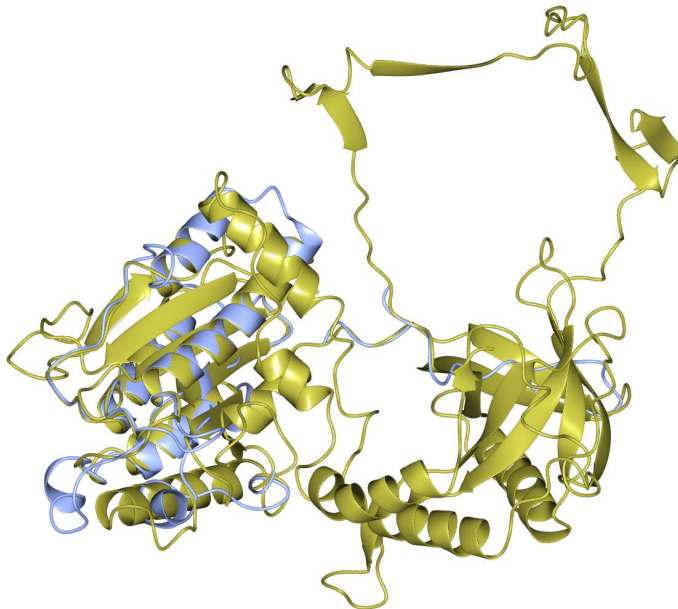
**Figure 11.** Shows the JPred 4 secondary structure prediction for DvGam and HiGam peptide sequences. Red:  $\alpha$ -helix, Green:  $\beta$ -sheet, prediction confidence is shown with the assigned numbers and black bar graphs (higher value corresponds to greater confidence in modelled structure) and peptide primary sequence is shown with corresponding predicted secondary structures. The predicted secondary structure elements look extremely similar and again show the long single  $\alpha$ -helix structure at the N-terminal end. The predicted structures also bear little resemblance to the



predicted structures for Ku70 and Ku80 DNA binding domains. The predicted secondary structures for DvGam are the same as seen on the solved 3D structure for DvGam (2P2U) (**Figure 7**)

### 4.2.5 One-to-one threading HiGam sequence with Ku DNA binding site structure

Overall, there is no evidence that the HiGam amino acid sequence will conform to the binding domain of either Ku70 or Ku80. The sequence homology between HiGam and Ku80 DNA binding region (aa 275-433 as defined by (d'Adda di Fagagna *et al.*, 2003)) according to BLASTp is low. The regions that show sequence homology give 39% identity and 48% positives, but this is only for a 27 amino acid stretch of the HiGam sequence. The sequence similarity with the DNA binding domain of Ku70 is even less impressive. The findings did not match those shown in **Figure 8** comparing MuGam with the Ku70/80 DNA binding domains (d'Adda di Fagagna *et al.*, 2003). **Figure 6** shows the potential for the primary sequence of HiGam to produce alternative secondary structures (and tertiary 3D structures) to the DvGam model. The HiGam aa sequence was successfully aligned with part of the sequence for the Ku80 protein to produce a model for a potential conformation with 75% confidence, though it was outside the DNA binding domain of the Ku80 protein.

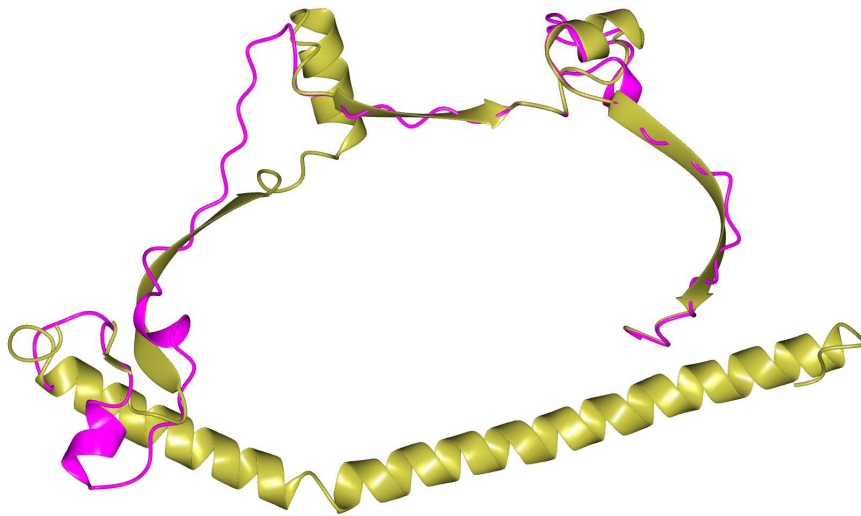


**Figure 12.** The predicted conformation of the HiGam protein sequence when forced into the full Ku80 3D structure (1JEY) when using the One-to-one threading mode on the Phyre2 tool. The HiGam sequence can be superimposed to the Ku80 structure with 75% confidence. This alignment however appears in the C-terminal region of the Ku80 3D structure. With reasonable confidence the HiGam sequence can potentially exist in a conformation outside of the predictions made by the prediction software. Repeating this with the full Ku70 3D structure did not yield any result.

When using the HiGam amino acid sequence as the query in one-to-one threading to the 3D structure of the truncated Ku80 DNA binding domain, the tool only aligned 10 residues with a confidence of 8.53% and 10% sequence identities. This therefore did not produce a predicted secondary structure conformation for HiGam that conforms to the truncated Ku80 DNA binding domain as suggested in **Figure 8**. When repeated using the HiGam sequence with the 3D structure of

truncated Ku70 the one-to-one threading prediction failed to produce any prediction based on sequence alignment and secondary structure prediction similarity.

The Phyre2 one-to-one threading tool was able to align the sequence and produce a 3D structure prediction for the Ku70 DNA binding sequence to the DvGam 3D structure (PDB: 2P2U). There were 70 aligned residues with 9% identities and a model confidence of 1.71%. The structure produced does not suggest that the Ku70 DNA sequence is able to exist in a conformation similar to DvGam (**Figure 7**). When repeated using the Ku80 sequence with the 3D structure of DvGam in Phyre2 one-to-one threading the tool failed to align the sequences and produce a 3D structure prediction.



**Figure 13.** Fitting the Ku70 DNA binding site sequence to the 3D structure of DvGam when using the One-to-one threading mode on the Phyre2 tool. It only selected a short region of the sequence to align to the structure of DvGam with a confidence of 1.71%, 70 aligned residues with 9% identities.

#### 4.3. Discussion

From these findings there is no evidence that HiGam would assume the conformation of anything resembling the DNA binding domain of the Ku70/80 heterodimer. Based on the primary sequences of the HiGam and Ku polypeptides there is very little similarity and no evidence that the sequences could produce similar secondary structures.

The monomeric crystal structure for DvGam is supported by the secondary structure predictors as the sequence for HiGam is predicted to produce the same secondary structure elements. It supports the idea of a single long or bundled  $\alpha$ -helix structure from the N-terminus to the middle of the sequence even when disregarding the DvGam 3D structure as it aligns with unrelated proteins. The Phyre2 results suggest the presence of a SAM-like domain in the C-terminal region of HiGam which takes the conformation of four short  $\alpha$ -helices in a dense bundle (**Table 9-10**).

The evidence shown does not consider the difference in folding and conformation possible when the HiGam homodimer forms. However, the evidence presented does suggest that whatever conformation this would take it probably wouldn't produce a structure that would look like the Ku heterodimer as the predicted secondary structures do not align (**Figure 10-11**). This could make it more difficult to link HiGam to a functional role in non-homologous end joining (NHEJ) pathways for DNA repair.

## Results

# Chapter 5 – Production of recombinant MuGam homologue proteins

## 5.1. Introduction

A previous MSc research student produced a recombinant expression plasmid encoding the gene for Im9-HiGam. This work established that the use of colicin E9 immunity protein (Im9) as an N-terminal solubility tag improved the functional soluble protein yield of the HiGam target protein. This research showed that when Im9-HiGam was overproduced in *Escherichia coli* K-12 MG1655 using the pQE2 expression plasmid the protein could be purified by Ni<sup>2+</sup>-affinity chromatography thanks to the presence of the 6x His-tag. This purified protein product was able to dimerise and function as a linear double stranded DNA binding protein.

To further investigate the conservation of the MuGam homologue genes across the wide range of bacterial species eight MuGam homologue genes were ligated into expression plasmids to produce purified protein samples. These could then be used to investigate protein function and attempt to determine the 3-dimensional structure to provide more information on how the protein may function in the cell and the differences between conserved gene products.

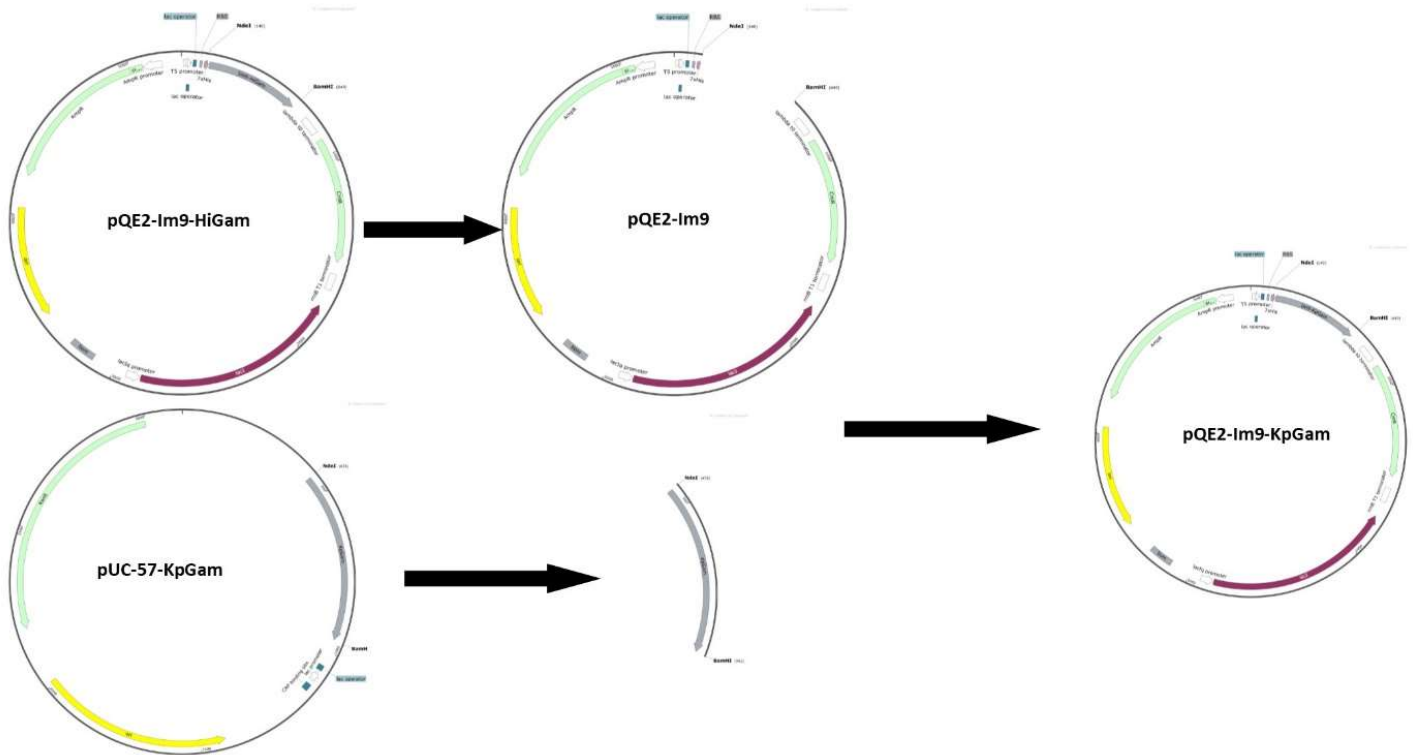
MuGam homologue name	Species origin	Protein accession no.	Gene length (bp)
HiGam	<i>Haemophilus influenzae</i> RdKW 20	WP_005693500.1	510
KpGam	<i>Klebsiella pneumoniae</i> CCIC01000013	WP_040211999.1	522
ApGam	<i>Avibacterium paragallinarum</i> JF4211	WP_017806105.1	507
SsGam	<i>Shigella sonnei</i> CDPH_C88	WP_001107937.1	525
SeGam	<i>Salmonella enterica</i>	WP_044127832.1	525
NmGam	<i>Neisseria meningitidis</i>	WP_049323767.1	522
EcGam	<i>Escherichia coli</i> 0157:H7 str. Sakai	WP_001129553.1	531
RcGam	<i>Rhodobacter capsulatus</i> SB 1003	WP_013066740.1	543
PaGam	<i>Pseudomonas aeruginosa</i> LESB58	WP_034040298.1	378

**Table 11.** Overview of the MuGam homologue proteins that are being investigated.

### 5.1.1 Gene transformation using pQE2 as a vector plasmid

The pQE2 plasmid expression system uses LacI to repress gene expression. In the presence of Isopropyl β- d-1-thiogalactopyranoside (IPTG), the LacI repressor releases from the operator sequence to allow efficient binding of RNA polymerase to the promoter for target gene expression. This system allows for the overproduction of the target protein when IPTG concentration is high as well as extremely limited protein production in the absence of IPTG and particularly when the bacterial growth medium is supplemented with D-Glucose.

## Production of recombinant MuGam homologue proteins



**Figure 14.** Simplified schematic view of the methods described in 2.3 showing the restriction digest of the pUC-57-KpGam and pQE2-Im9-HiGam and subsequent ligation to form the final recombinant pQE2-Im9-KpGam.

### 5.2.3 Transforming *E. coli* DH5- $\alpha$ with pUC-57-Gam plasmids by heat shock

Transformation by heat shock of the product of ligation into *E. coli* DH5- $\mu$  for amplification and purification of the plasmids. As shown in **Fig 14** the ligation product (pQE2-Im9-KpGam for example) is formed by using the appropriate restriction endonuclease enzymes (detailed in **Materials and Methods 2.6.3**) to excise the desired gene sequence from the carrier plasmid to be ligated with the expression plasmid which was digested using the corresponding restriction enzymes.

Another protein produced by *Haemophilus influenzae* Rd KW20 is HI1450, a small double stranded DNA (dsDNA) mimic protein (Parsons, Liu and Orban, 2009). This protein may have the potential to bind to HiGam and act as a competitive inhibitor by binding to the DNA binding region of HiGam and blocking linear DNA binding. In order to investigate the interaction between the two proteins the recombinant HI1450 protein would need to be cloned and purified.

## 5.2. Results

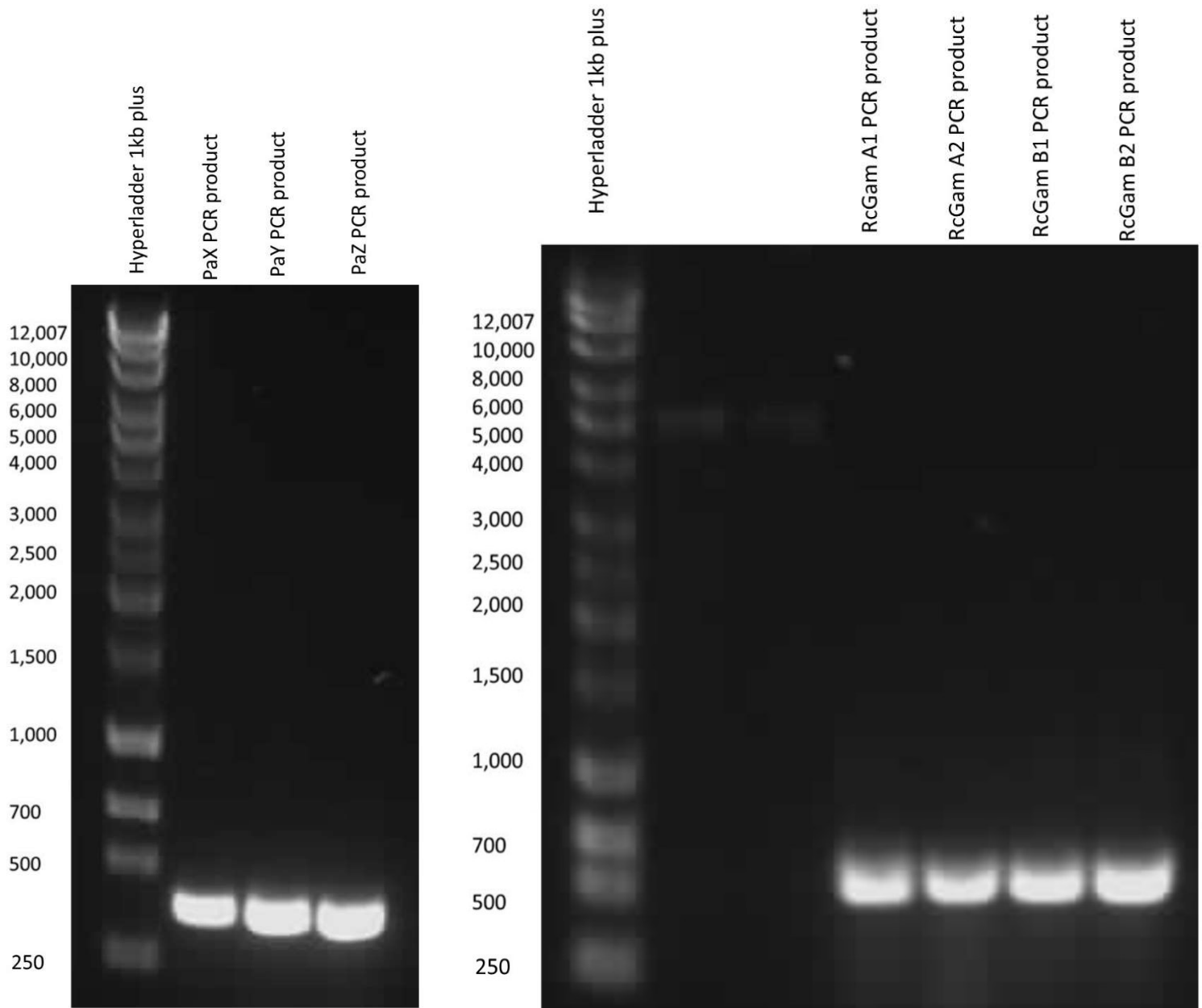
### 5.2.1 Diagnostic PCR and double digest of recombinant pQE2-Im9-Gam homologue

The diagnostic PCR or double digest provide evidence that the MuGam homologue gene was successfully ligated with the expression plasmid and that transformation with the Im9-Gam-homologue gene was successful. Diagnostic PCR was used where primers were available (RcGam and

## Production of recombinant MuGam homologue proteins

PaGam samples) while diagnostic double digest was used for the others as primers were not available.

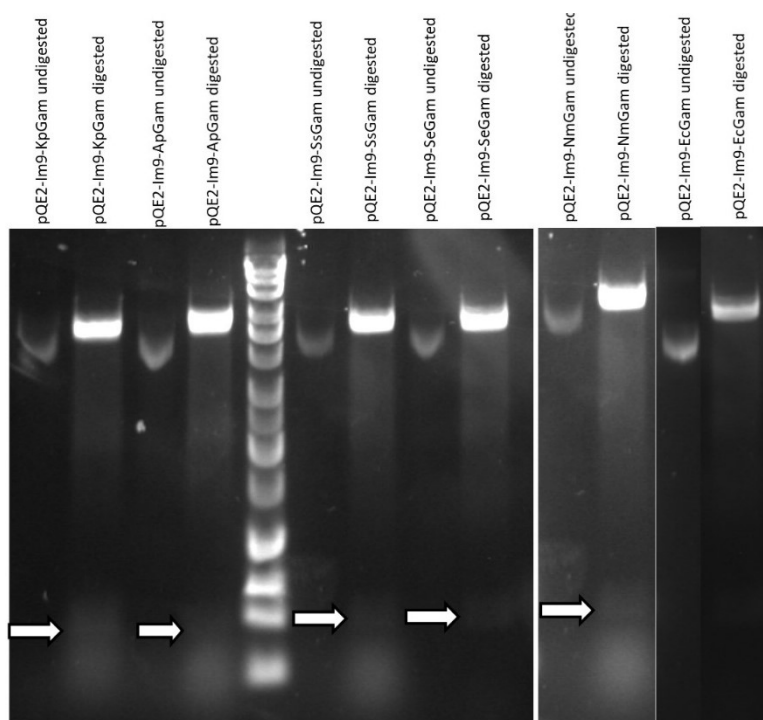
Each recombinant plasmid (selection with 100 µg/ml ampicillin) was formed by ligating the appropriate synthesised gene into the pQE2-Im9 expression plasmid and used to successfully transform *E. coli* DH5- $\alpha$  cells. The purified plasmid was used to transform *E. coli* K-12 MG1655 cells for the production of recombinant protein.



**Figure 15.** 1% (w/v) agarose-TAE electrophoresis gel images presenting evidence of successful amplification of target genes from transformed *E. coli* DH5-  $\alpha$  cells. Left: Diagnostic PCR of purified pQE2-Im9-PaGam from three colonies X, Y and Z grown on ampicillin selection LB plates after transformation by electroporation using primers to bind either side of the PaGam gene. Bright bands are present at the expected position to indicate the PaGam gene between 250 and 500 bp Right: Diagnostic PCR of purified pQE2-Im9-RcGam from two colonies A and B grown on ampicillin selection LB plates. Bands at approx. 500 bp is expected for the RcGam gene.

## Production of recombinant MuGam homologue proteins

The diagnostic double restriction digest should cleave the plasmid DNA at the sites for NdeI and BamHI-HF and result in two DNA bands, the vector pQE2-Im9 and the insert Gam gene (approx. 500bp). The gene inserts were identified in five of the six digested plasmids. The pQE2-Im9-EcGam did not appear to form correctly and resulted in a false positive (Figure 15). The undigested plasmid appeared shorter than those of the other recombinant plasmids and the 531bp gene insert band is not visible.



**Figure 16.** 1% (w/v) agarose-TAE gel electrophoresis showing the diagnostic double digest for each Gam homolog ligated with pQE2-Im9 and transformed to *E. coli* DH5- $\alpha$ . The gene insert can be identified in each homolog with exception to EcGam which appears to be a false positive.

The digested plasmids consistently showed a degree of degradation that is not consistent with the expected cleavages of the DNA at the NdeI and BamHI-HF sites. This can be identified by the smearing below the top/vector plasmid band as well as the undefined band below the 250bp molecular weight marker. This extra degradation is most likely a result of the long (overnight) incubation period and could represent star activity.

The purified plasmids for each pQE2-Im9-Gam were used to transform the expression strain *E. coli* MG1655 and can subsequently be tested for protein production after induction of recombinant gene expression with IPTG.

### 5.2.2 Plasmid sequencing

A commercial DNA sequencing service (Eurofins) was used to confirm the presence of the correct gene sequence in the pQE2-Im9 expression plasmid. This provides a much more accurate confirmation of the gene sequence and gives confidence that overexpression of the gene will produce the correct target proteins. Details of the sequencing results are provided in the **Appendix**.

The forward pQE2-Im9-RcGam sequence gives the expected sequence to produce the correct RcGam polypeptide sequence though a single point mutation has occurred to differentiate it from

## Production of recombinant MuGam homologue proteins

the natural gene sequence at the wobble base in the valine codon so it does not give rise to amino acid sequence change.

However, the reverse sequence for RcGam showed a deviation from the expected genetic sequence at the 3' end resulting in the loss of the stop codon, most likely due to the loss of sequencing accuracy near the end of the sequence. The sequencing was repeated, but instead of using the pQE2-Im9 plasmid sequencing primers, we premixed the DNA sample with another forward primer (RKG9) used to amplify the RcGam gene. The resulting sequence gave us much greater coverage of the 3'-terminus region in the gene and confirmed that the correct sequence had been obtained. The point mutation mentioned above was seen again in this sequence data.

The sequencing results for Im9-NmGam revealed the presence of a premature stop codon directly downstream from the Im9 encoding sequence. It is likely that the wrong sequence was ligated into the pQE2-Im9 vector plasmid which retained the BamHI-HF recognition sequence and formed a MluI restriction site where the NdeI restriction site should have been. This restriction site may have a small amount activity with NdeI resulting in the false positive diagnostic digest result.

Each of the other MuGam homologue plasmid sequences were confirmed by the sequencing as detailed in **Appendix (11.3)**.

### 5.2.3 Determining optimal growing and induction conditions for MuGam homologue protein overproduction

In order to maximise the yield of soluble recombinant protein, the conditions for growth during induction were optimised to find the best conditions for protein production prior to a large-scale grow up. The Gam homologue proteins will be purified from the soluble fractions of the cell lysates so the method must be optimised for soluble protein yield. Each individual protein may potentially be produced at greater efficiency under different conditions, so each protein was tested under each condition in case there was a difference in soluble protein yield across each Gam homologue.

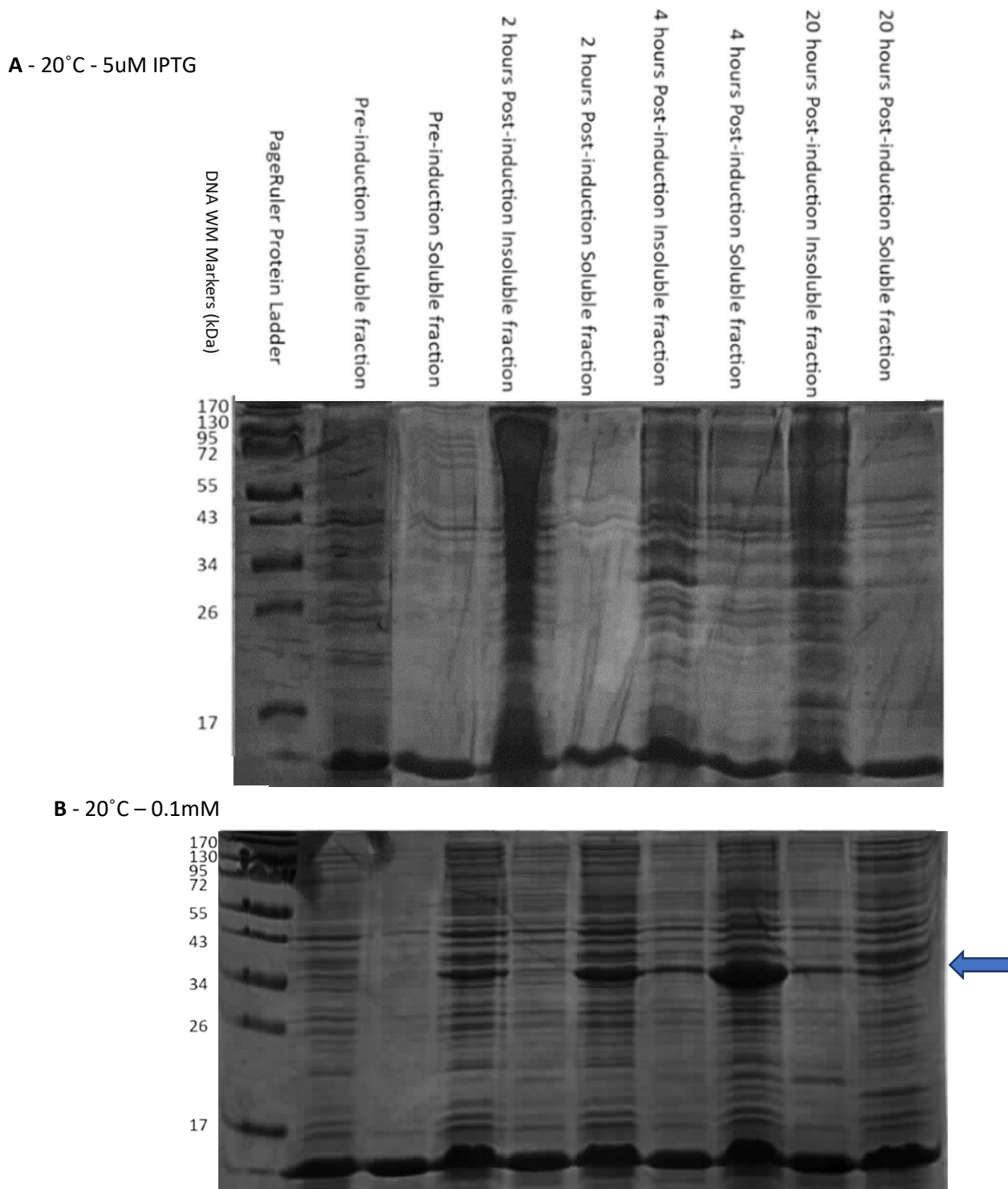
The following conditions were investigated to find the optimal conditions for soluble protein production: post-induction bacterial growth temperature, IPTG concentration, and length of time for post-induction bacterial growth. These conditions were chosen to test the widest range of potential variables to see their effects on soluble protein yield.

The tests were carried out initially using *E. coli* K-12 MG1655 cells transformed with pQE2-Im9-PaGam as well as *E. coli* K-12 MG1655 cells transformed with pQE2-Im9-RcGam. This preliminary screen was used to inform on the growth conditions for the other MuGam homologue proteins.

The results from **Figures 17** and **18** suggested that for the greatest yield of soluble Im9-RcGam protein the most ideal conditions were to induce protein overproduction using 0.1mM IPTG and growing at 37°C for 4 hours. This was determined due to the band associated with the target protein with the correct molecular weight (MW) appearing to be the most intense when collected and analysed under these conditions. The same conditions also appear to be optimal for Im9-PaGam as shown in **Figures 19** and **20**.

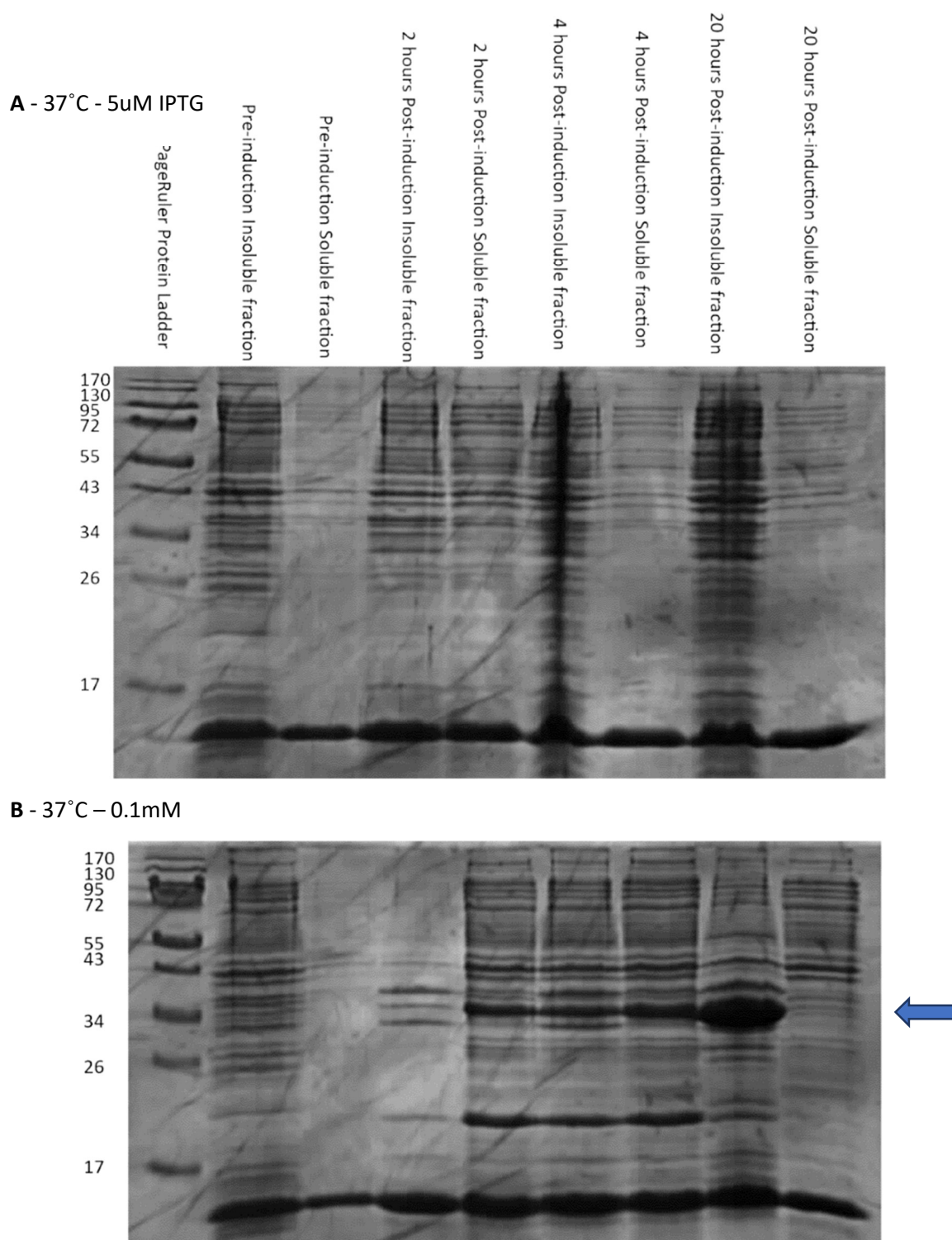


## Production of recombinant MuGam homologue proteins



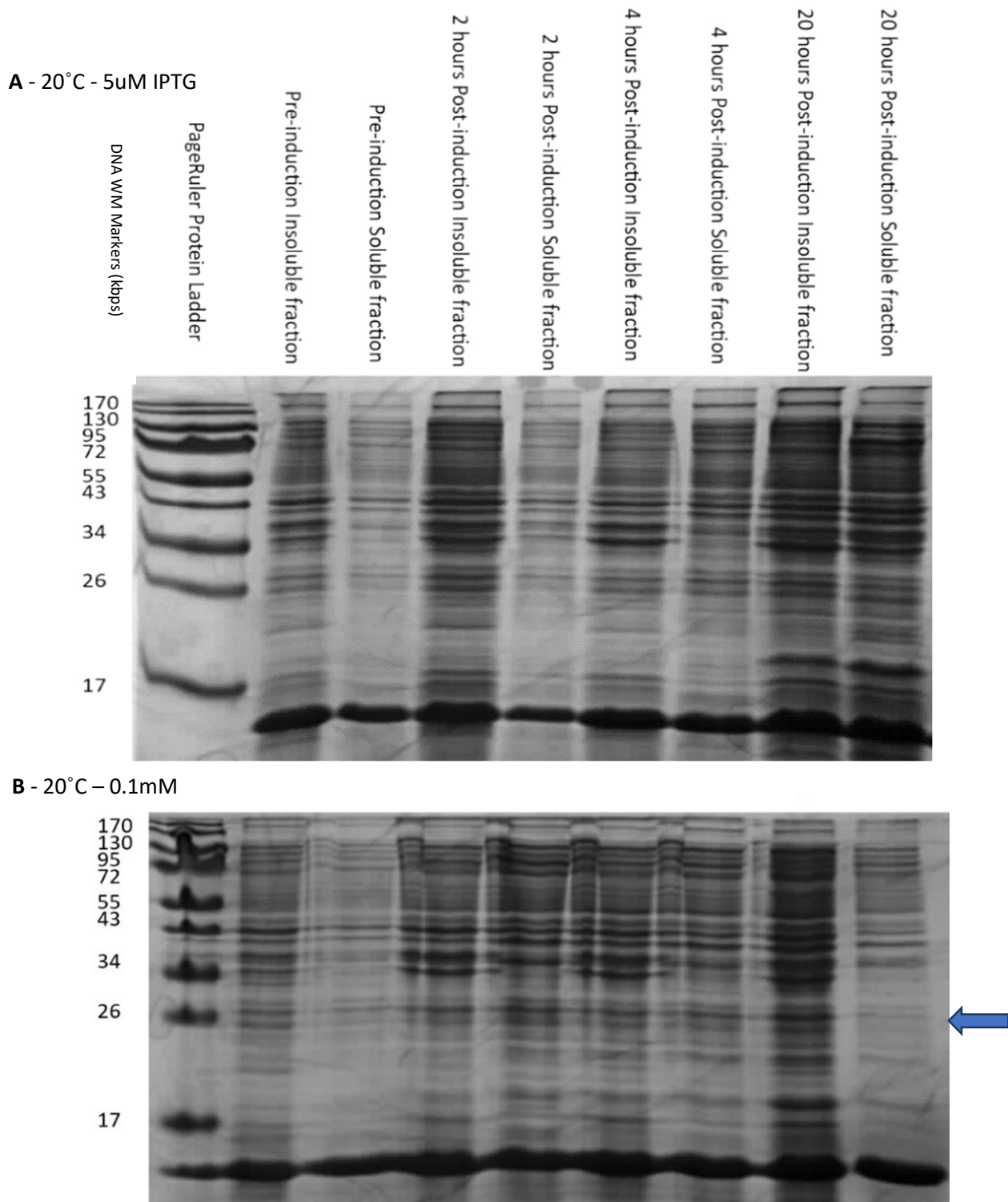
**Figure 17.** SDS-PAGE analysis of *E. coli* K-12 MG1655 soluble and insoluble protein fractions after IPTG induction for overproduction of Im9-RcGam protein from the pQE2-Im9-RcGam recombinant plasmid incubated at 20°C post-induction with either **A**: 5 $\mu$ M or **B**: 0.1mM IPTG. Each lane of **A** is labelled with the amount of time of incubation after induction with IPTG and whether it contains the soluble or insoluble fraction of the cell lysate. The lanes for Image **B** are shown in the same position as image **A**. The expected molecular weight of the Im9-RcGam protein is 31.87kDa, expected band associated with overproduced protein marked with blue arrow.

## Production of recombinant MuGam homologue proteins



**Figure 18.** SDS-PAGE analysis of *E. coli* K-12 MG1655 soluble and insoluble protein fractions after IPTG induction for overproduction of Im9-RcGam protein from the pQE2-Im9-RcGam recombinant plasmid incubated at 37°C post-induction with either **A**: 5μM or **B**: 0.1mM IPTG. Each lane of **A** is labelled with the amount of time of incubation after induction with IPTG and whether it contains the soluble or insoluble fraction of the cell lysate. The lanes for Image **B** are shown in the same position as image **A**. The expected molecular weight of the Im9-RcGam protein is 31.87kDa, expected band associated with overproduced protein marked with blue arrow.

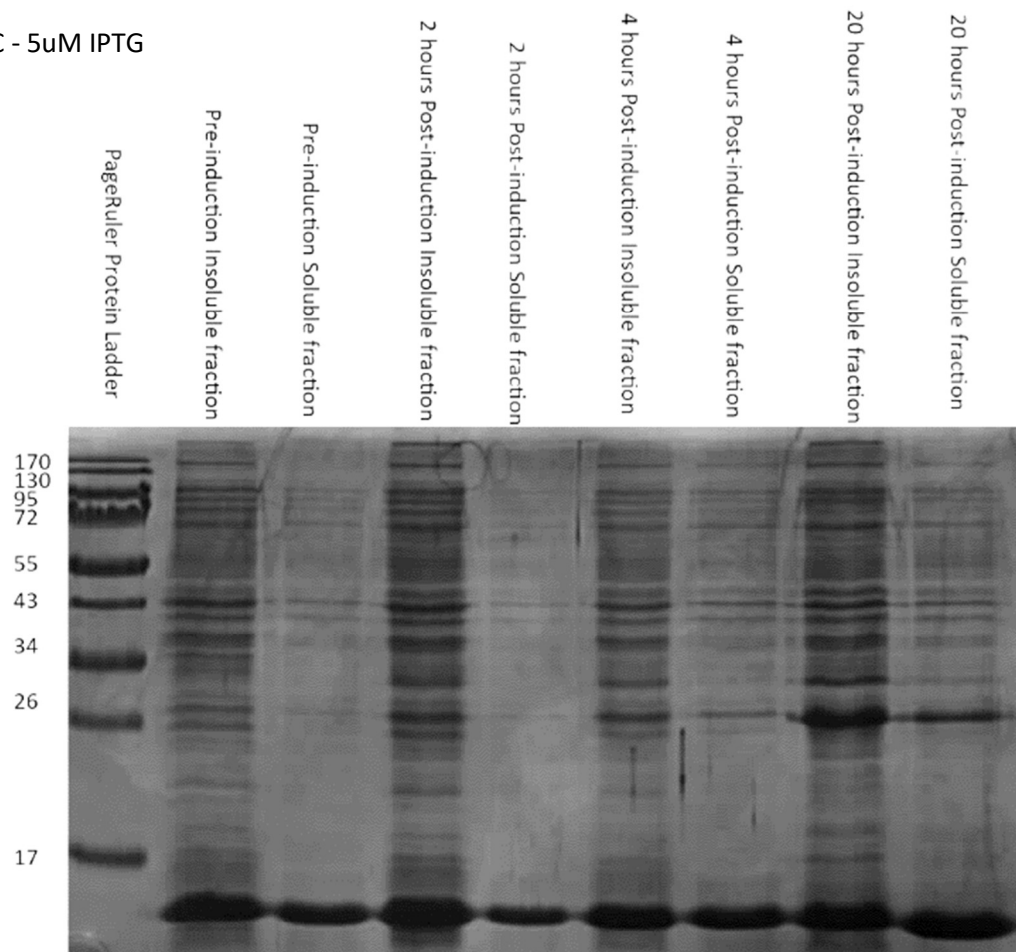
## Production of recombinant MuGam homologue proteins



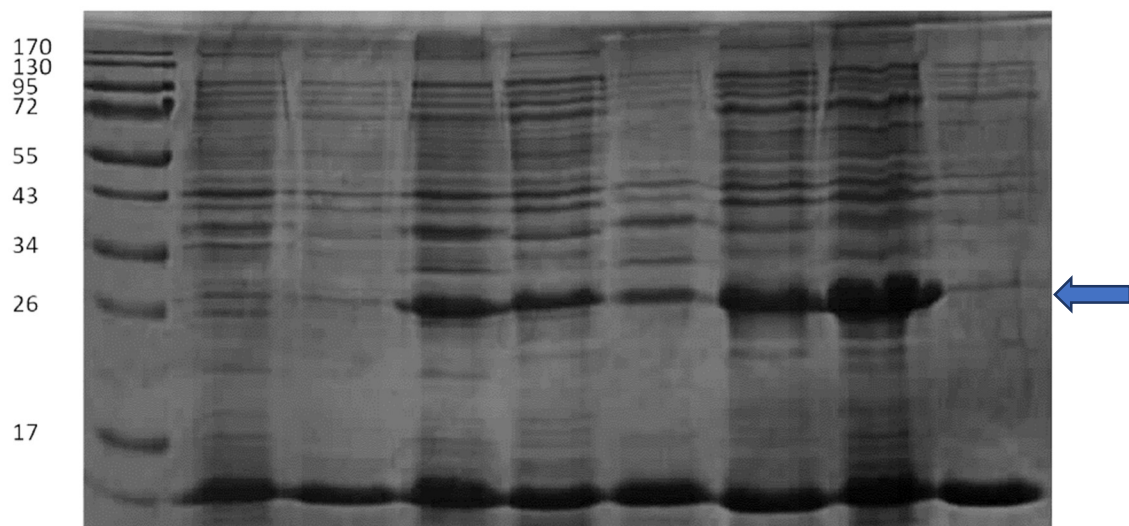
**Figure 19.** SDS-PAGE analysis of *E. coli* K-12 MG1655 soluble and insoluble protein fractions after IPTG induction for overproduction of Im9-PaGam protein from the pQE2-Im9-PaGam recombinant plasmid incubated at 20°C post-induction with either **A**: 5 $\mu$ M or **B**: 0.1mM IPTG. Each lane of **A** is labelled with the amount of time of incubation after induction with IPTG and whether it contains the soluble or insoluble fraction of the cell lysate. The lanes for Image **B** are shown in the same position as image **A**. The expected molecular weight of the Im9-PaGam protein is 25.07kDa, expected band associated with overproduced protein marked with blue arrow.

## Production of recombinant MuGam homologue proteins

**A - 37°C - 5 $\mu$ M IPTG**



**B - 37°C - 0.1mM**

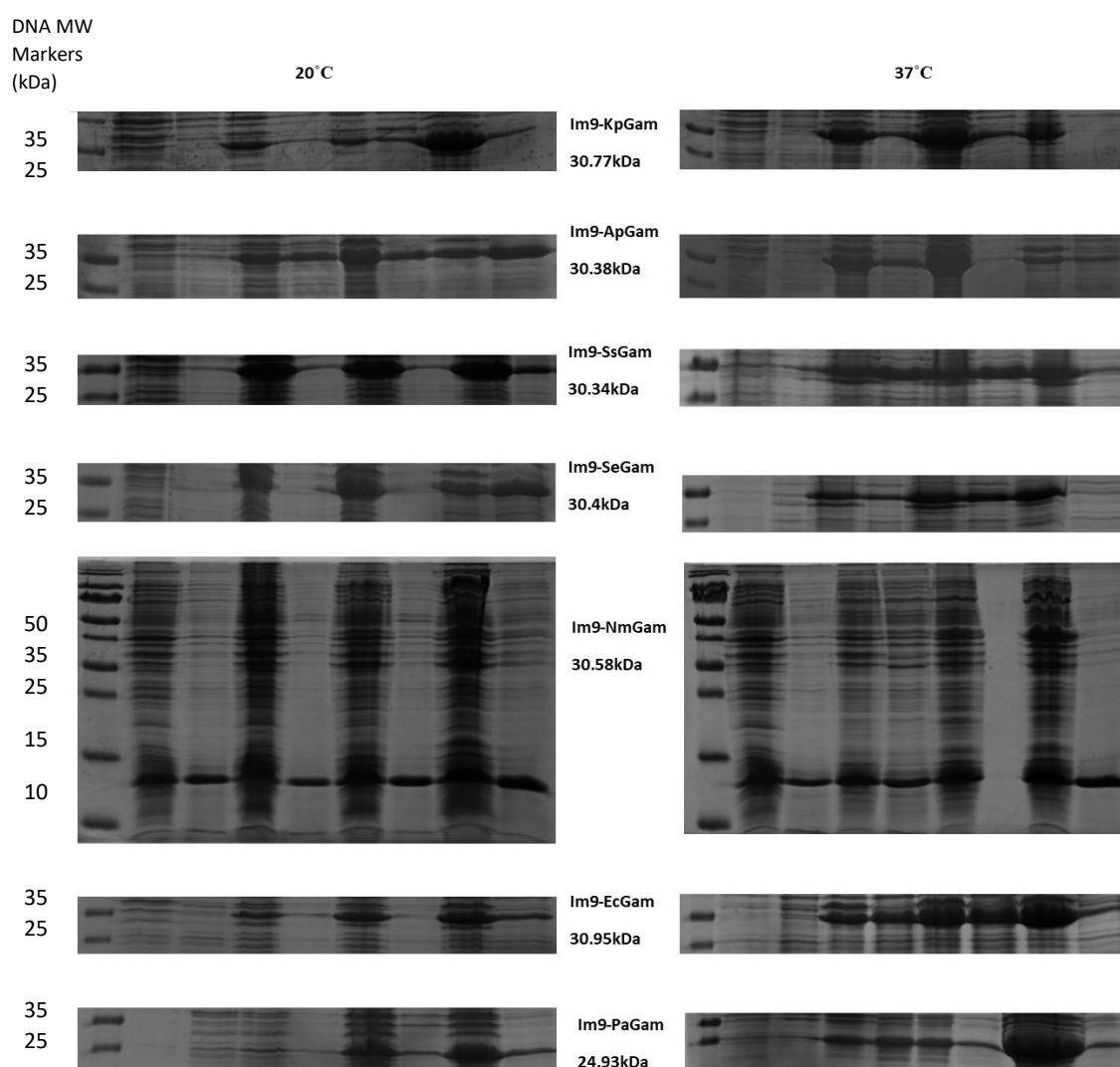


**Figure 20.** SDS-PAGE analysis of *E. coli* K-12 MG1655 soluble and insoluble protein fractions after IPTG induction for overproduction of Im9-PaGam protein from the pQE2-Im9-PaGam recombinant plasmid incubated at 20°C post-induction with either **A**: 5 $\mu$ M or **B**: 0.1mM IPTG. Each lane of **A** is labelled with the amount of time of incubation after induction with IPTG and whether it contains the soluble or insoluble fraction of the cell lysate. The lanes for Image **B** are shown in the same position as image **A**. The expected molecular weight of the Im9-PaGam protein is 25.07kDa, expected band associated with overproduced protein marked with blue arrow.

## Production of recombinant MuGam homologue proteins

When scaling up these conditions (0.1mM IPTG induction, incubation at 37°C for 4 hours), the total amount of soluble protein yield was very low. After the Im9-RcGam and Im9-PaGam proteins were purified by Ni<sup>2+</sup>-affinity chromatography, the monomeric protein concentrations were only 1.9µM and 6.3µM, respectively. This is significantly lower than that obtained when purifying Im9-HiGam where protein yield produced monomeric protein concentration of 67.2µM. The optimal conditions used for the production of Im9-HiGam were 0.1mM IPTG induction and incubation overnight at 20°C.

Using this information, the subsequent Im9-MuGam homologue proteins were overproduced by 0.1mM IPTG induction and 20 hours incubation at 20°C then cells were harvested, lysed and the proteins were purified by Ni<sup>2+</sup>-affinity chromatography. In each case a much less intense band corresponding to the target protein was observed compared with the previously defined optimal conditions (0.1mM IPTG induction and incubation for 4 hours at 37°C) as seen in **Figures 18-20**.



**Figure 21.** 15% (w/v) polyacrylamide-SDS gel images showing lysates for each transformed *E. coli* MG1655 culture containing each Gam homolog gene. Each gel lane from left to right contains 1. Page Ruler pre-stained protein ladder with the molecular weights listed (LHS) 2. Pre-induction insoluble fraction 3. Pre-induction Soluble fraction 4. 2 hours (2h) post-induction insoluble 5. 2h post-induction soluble 6. 4h post-induction insoluble 7. 4h post-induction soluble 8. Overnight post-induction insoluble 9. Overnight post-induction soluble.

## Production of recombinant MuGam homologue proteins

The test over-production of the other Im9-Gam homologues showed that there is consistently a concentrated band of overproduced protein in the soluble fraction when incubated overnight at 20°C, and these conditions were used when producing Im9-HiGam. At 37°C after 4 hours there is often an intense band in the soluble fraction, though due to the short incubation time the total protein produced is relatively low when compared to incubating overnight. Therefore, the ideal conditions based on these experiments were to induce with 0.1mM final concentration IPTG and incubate at 20°C overnight to maximise yield of recombinant Im9-Gam protein.

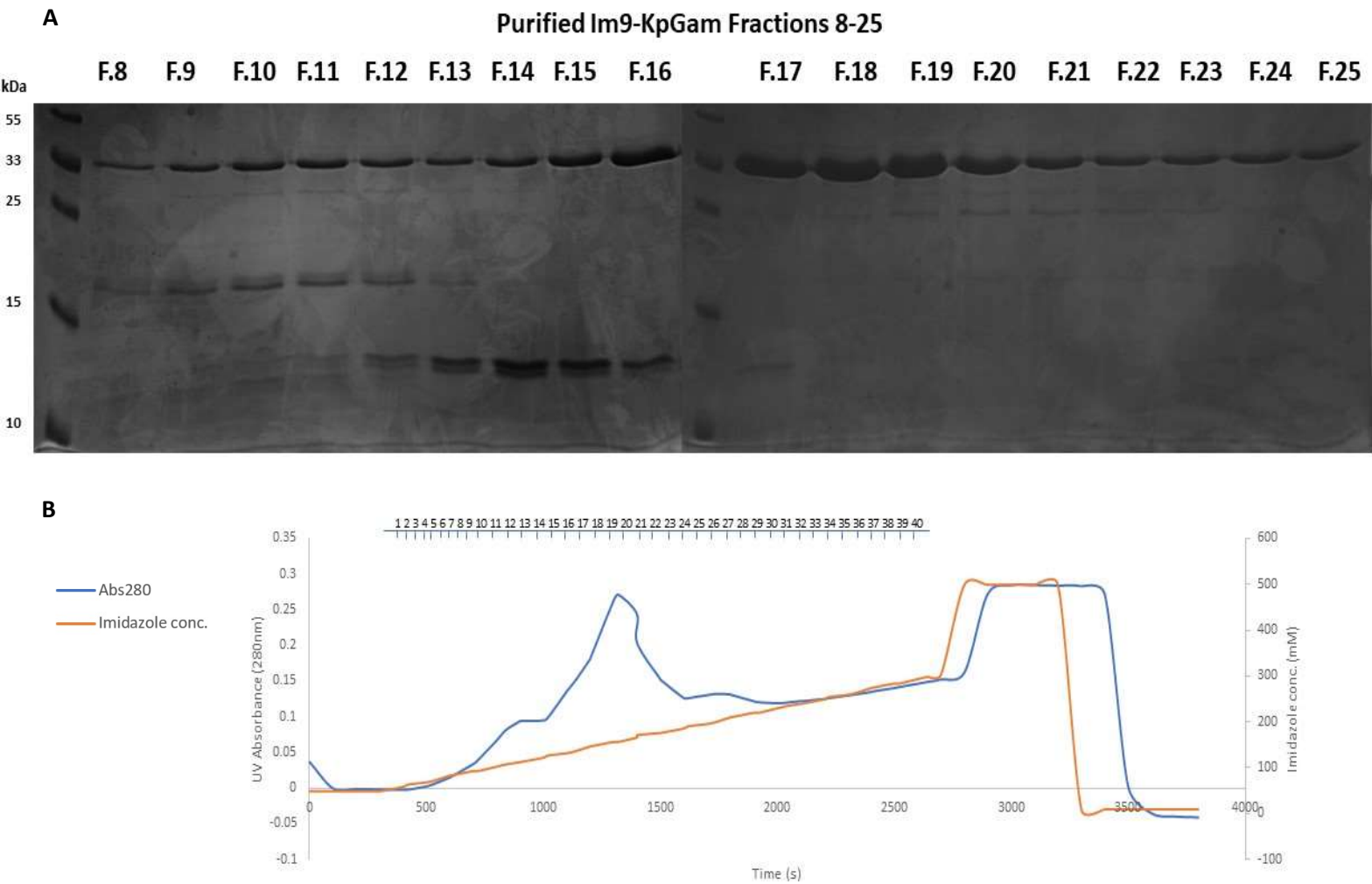
Im9-NmGam is the outlier, despite evidence of a successful transformation using the diagnostic double digest (**Figure 15**), the induction using IPTG did not result in overproduction of Im9-NmGam, and there is no evidence of a significantly intense band at the expected molecular weight in the post-induction samples.

### 5.2.4 MuGam homologue protein purification by Ni<sup>2+</sup>-affinity chromatography

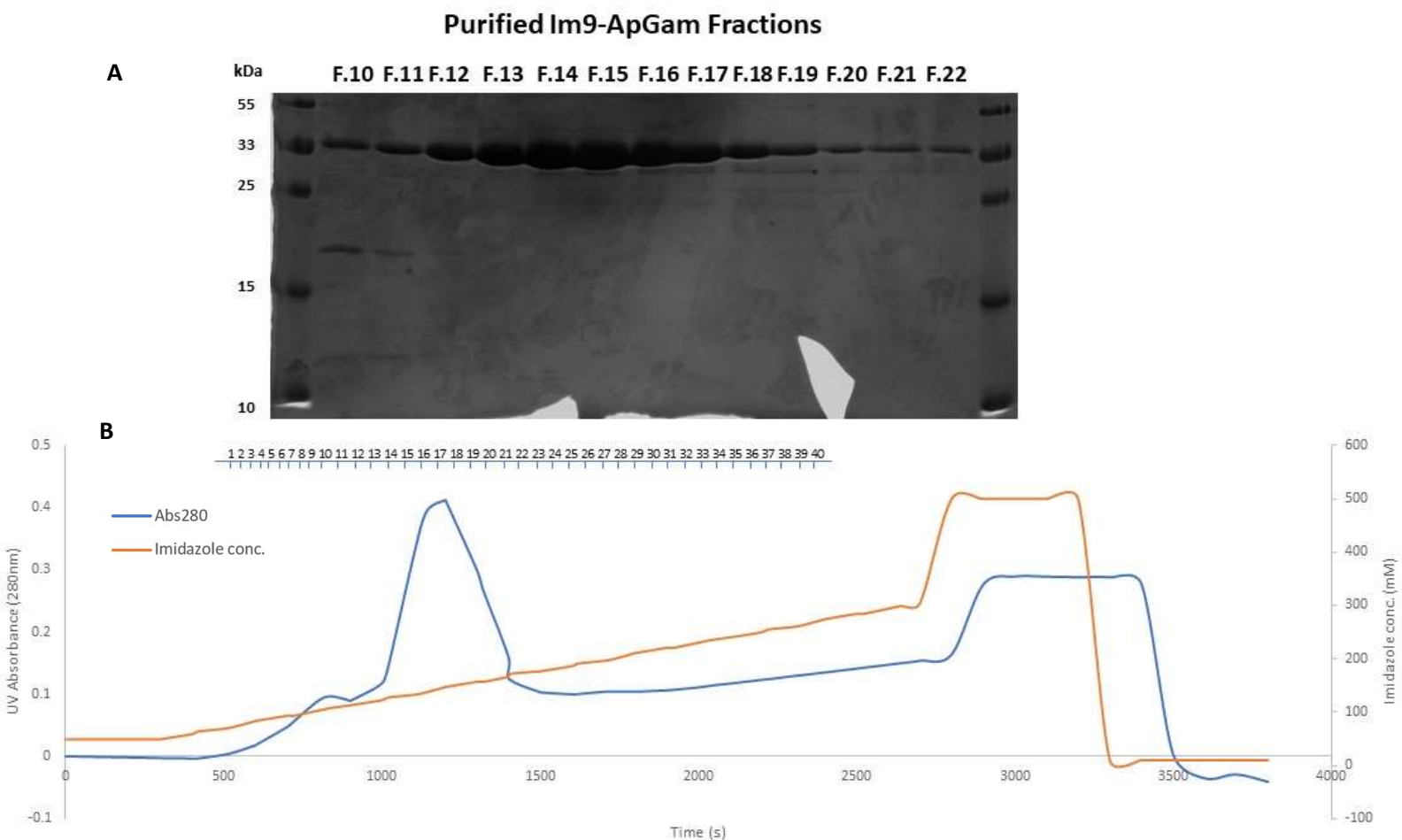
The 6x-His tag present at the N-terminus of each of the Im9-MuGam homologue proteins could be leveraged to separate the target proteins from the rest of the soluble proteins in the total bacterial cell lysate. This allows the overproduced target protein to anneal to the stationary phase in the Ni<sup>2+</sup>-chelating column until the imidazole concentration increased to the threshold required to elute the target protein from the column. The soluble fraction of the total cell lysate was run through the Ni<sup>2+</sup>-chelating column using an imidazole concentration of 10mM, and the majority of remaining proteins that may have bound weakly to the column were washed off using a 49.2mM imidazole concentration. The imidazole concentration is gradually increased at a constant rate and absorbance at 280nm is measured to identify the fractions which contain the target protein.

Im9-NmGam was not successfully overproduced in the *E. coli* K-12 MG1655 cells so the transformed cells were not collected post-induction with IPTG for Im9-NmGam protein purification.

Each Im9-MuGam homologue protein was eluted from the column at slightly different imidazole concentrations but each of the target proteins were completely eluted at an imidazole concentration of 157.0 – 196.7mM. The target proteins were eluted in forty 1 mL fractions, and the fractions containing the recombinant protein were pooled together and dialysed.



**Figure 22. A:** 10% (w/v) polyacrylamide-SDS gel image showing the protein fractions analysed to identify the fractions which contain the most target Im9-KpGam protein and the least amount of unwanted protein. The chosen fractions were then pooled to investigate protein function. **B:** The UV absorbance at 280nm (blue trace) was used to identify when protein eluted from the column as the imidazole concentration (orange) increased. The Im9-KpGam eluted into fractions 8-25. Only fractions 17-25 were pooled and dialysed because fractions 8-16 contained a significant amount of non-target protein.



**Figure 23. A:** 10% (w/v) polyacrylamide-SDS gel image showing the protein fractions analysed to identify the fractions which contain the most target Im9-ApGam protein and the least amount of unwanted protein. The chosen fractions were then pooled to investigate protein function. **B:** The UV absorbance at 280nm (blue trace) was used to identify when protein eluted from the column as the imidazole concentration (orange) increased. The Im9-ApGam eluted into fractions 10-22. Only fractions 11-22 were pooled and dialysed because fraction 10 contained a significant amount of non-target protein.



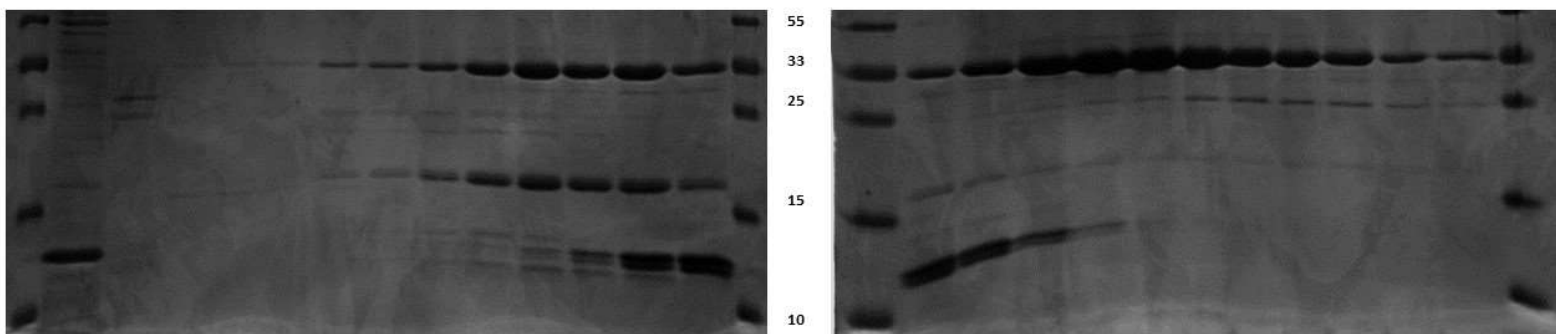
A

Purified Im9-SsGam Fractions

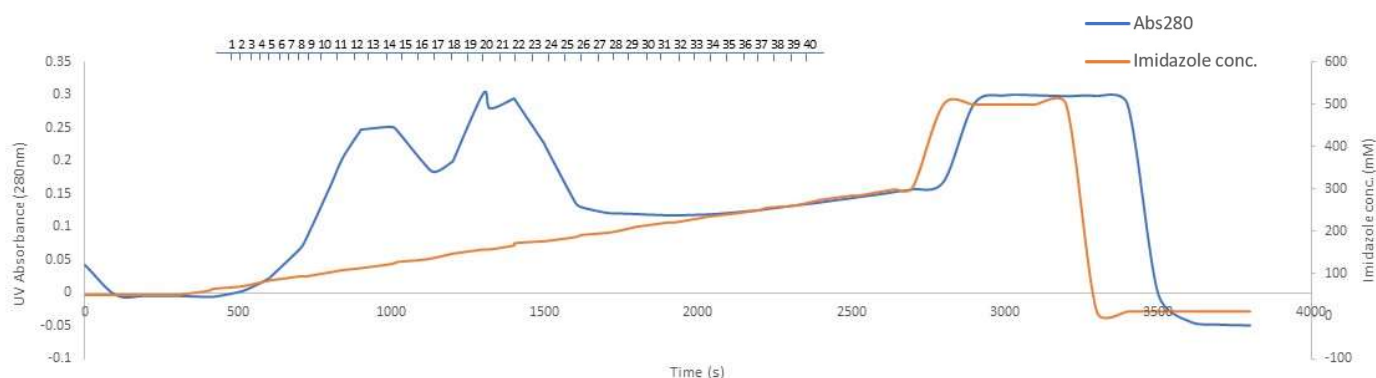
F.3 F.4 F.5 F.6 F.7 F.8 F.9 F.10 F.11 F.12 F.13

kDa

F.14 F.15 F.16 F.17 F.18 F.19 F.20 F.21 F.22 F.23 F.24



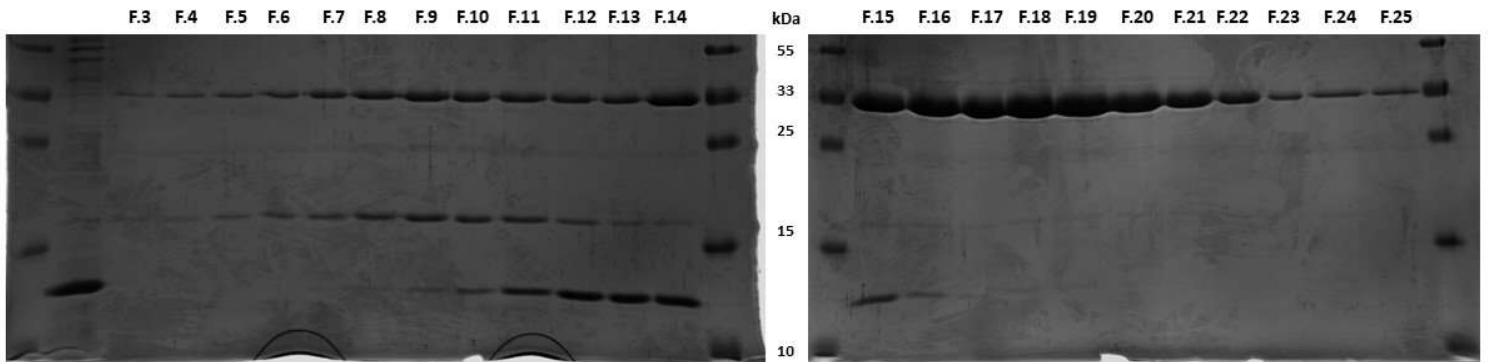
B



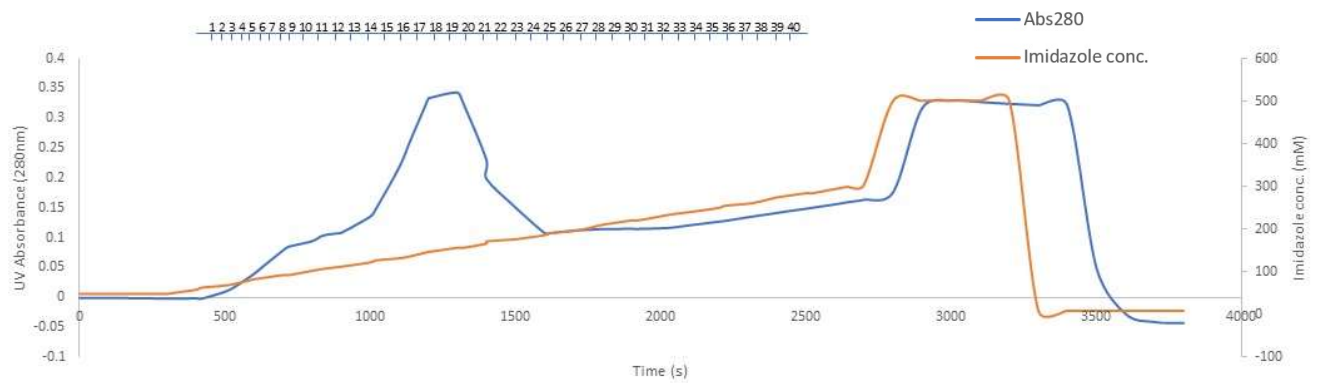
**Figure 24. A:** 10% (w/v) polyacrylamide-SDS gel image showing the protein fractions analysed to identify the fractions which contain the most target Im9-SsGam protein and the least amount of unwanted protein. The chosen fractions were then pooled to investigate protein function. **B:** The UV absorbance at 280nm (blue trace) was used to identify when protein eluted from the column as the imidazole concentration (orange) increased. The Im9-SsGam eluted into fractions 3-24. Only fractions 18-24 were pooled and dialysed because fractions 3-17 contained a significant amount of non-target protein.

A

Purified Im9-SeGam fractions



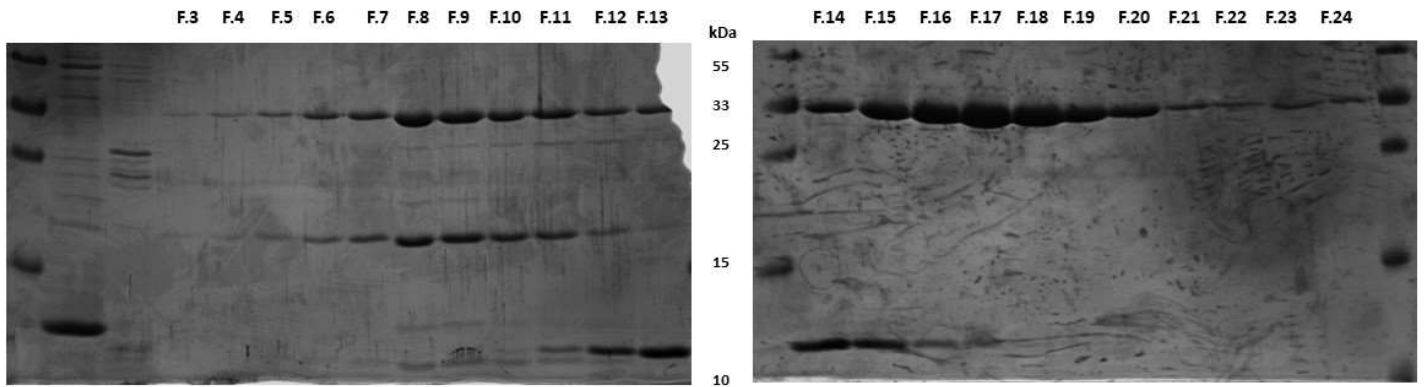
B



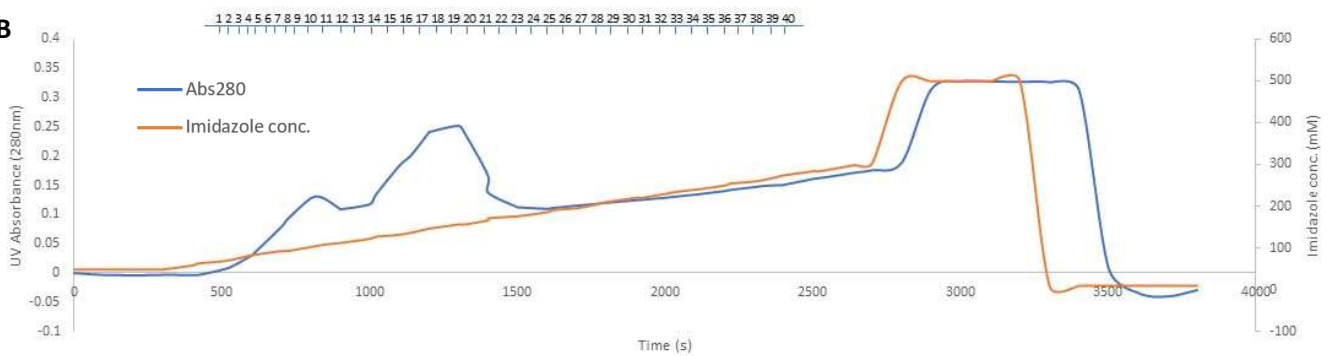
**Figure 25. A:** 10% (w/v) polyacrylamide-SDS gel image showing the protein fractions analysed to identify the fractions which contain the most target Im9-SeGam protein and the least amount of unwanted protein. The chosen fractions were then pooled to investigate protein function. **B:** The UV absorbance at 280nm (blue trace) was used to identify when protein eluted from the column as the imidazole concentration (orange) increased. The Im9-SeGam eluted into fractions 3-25. Only fractions 16-25 were pooled and dialysed because fractions 3-15 contained a significant amount of non-target protein.

Purified Im9-EcGam fractions

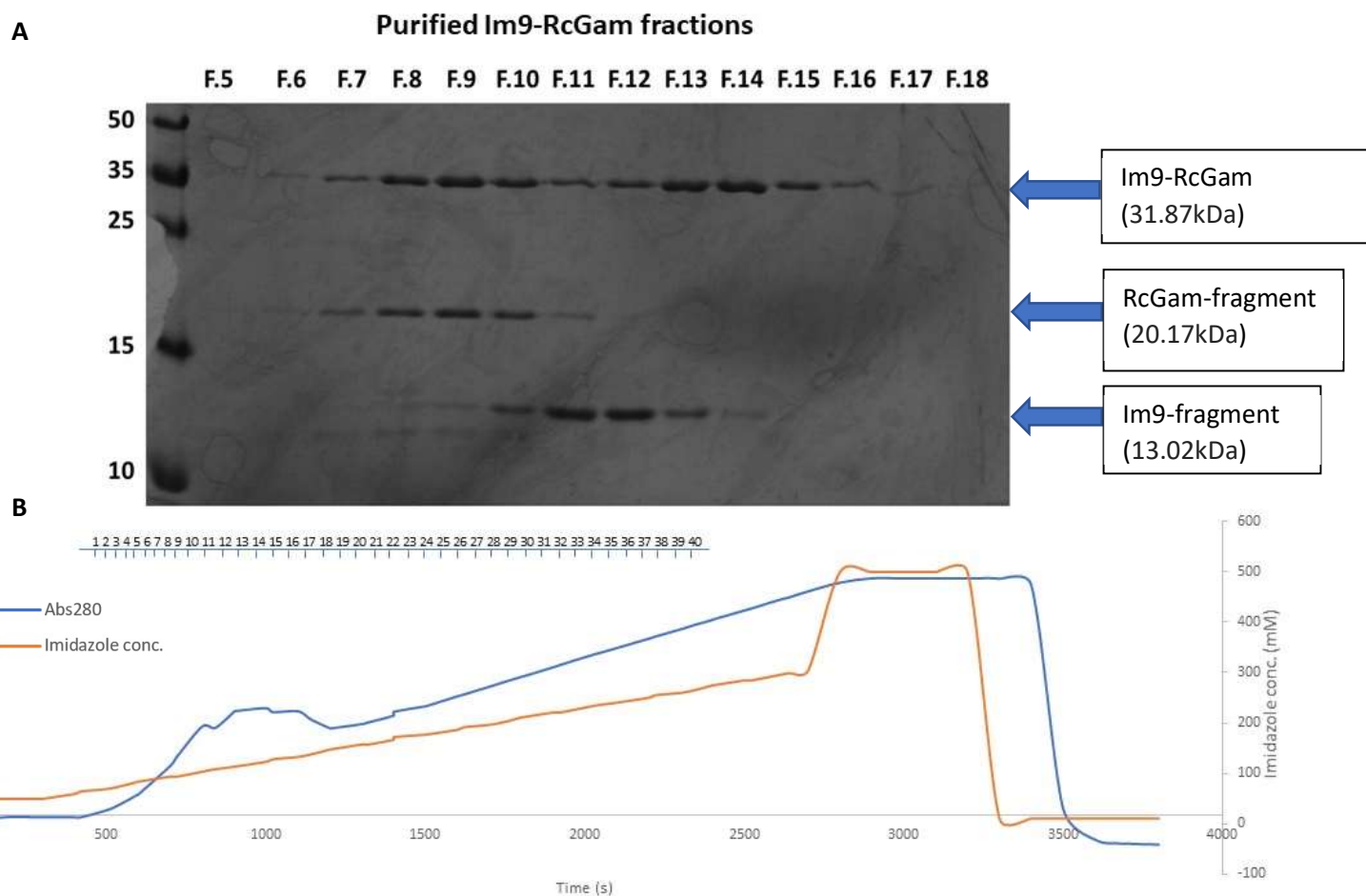
A



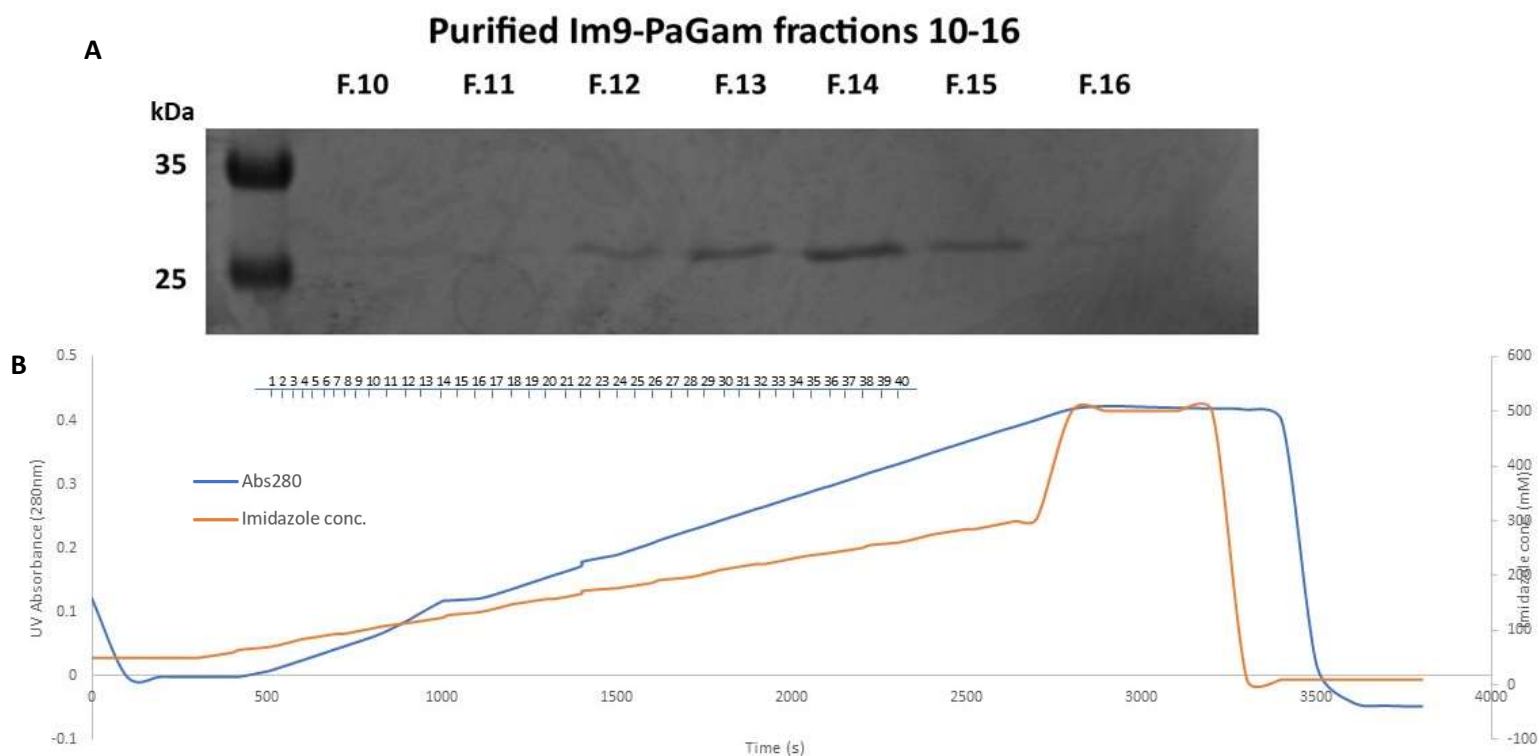
B



**Figure 26. A:** 10% (w/v) polyacrylamide-SDS gel image showing the protein fractions analysed to identify the fractions which contain the most target Im9-EcGam protein and the least amount of unwanted protein. The chosen fractions were then pooled to investigate protein function. **B:** The UV absorbance at 280nm (blue trace) was used to identify when protein eluted from the column as the imidazole concentration (orange) increased. The Im9-EcGam eluted into fractions 3-24. Only fractions 17-24 were pooled and dialysed because fractions 3-16 contained a significant amount of non-target protein.



**Figure 27. A:** 10% (w/v) polyacrylamide-SDS gel image showing the protein fractions analysed to identify the fractions which contain the most target Im9-RcGam protein and the least amount of unwanted protein. The chosen fractions were then pooled to investigate protein function. **B:** The UV absorbance at 280nm (blue trace) was used to identify when protein eluted from the column as the imidazole concentration (orange) increased. The Im9-RcGam eluted into fractions 5-18. Only fractions 12-16 were pooled and dialysed because fractions 5-13 contained a significant amount of non-target protein. Fractions 6-9 were also pooled and dialysed to analyse the unexpected non-target proteins. The three distinct protein bands are labelled with their predicted protein names as discussed in 5.2.5. as well as their estimated molecular weights (**Figure 29**).



**Figure 28. A:** 10% (w/v) polyacrylamide-SDS gel image showing the protein fractions analysed to identify the fractions which contain the most target Im9-EcGam protein and the least amount of unwanted protein. The chosen fractions were then pooled to investigate protein function. **B:** The UV absorbance at 280nm (blue trace) was used to identify when protein eluted from the column as the imidazole concentration (orange) increased with the number of each fraction shown above the graph. The Im9-EcGam eluted into fractions 10-16. The yield of purified protein was very low and fractions 10-16 were all pooled for dialysis and subsequent analysis.

Each of the target MuGam homologue proteins with the exception of Im9-NmGam was successfully purified using  $\text{Ni}^{2+}$ -affinity chromatography. **Figures 26** and **27** show that the protein yield was significantly lower than the other Im9-MuGam homologues likely as a result of the change in cell collection time and temperature as discussed in section 2.3. In many cases there are multiple peaks in the absorbance at 280nm as the fractions eluted from the column. This can be seen on the SDS-PAGE gel images where multiple bands are seen with different MWs compared to the expected target protein.

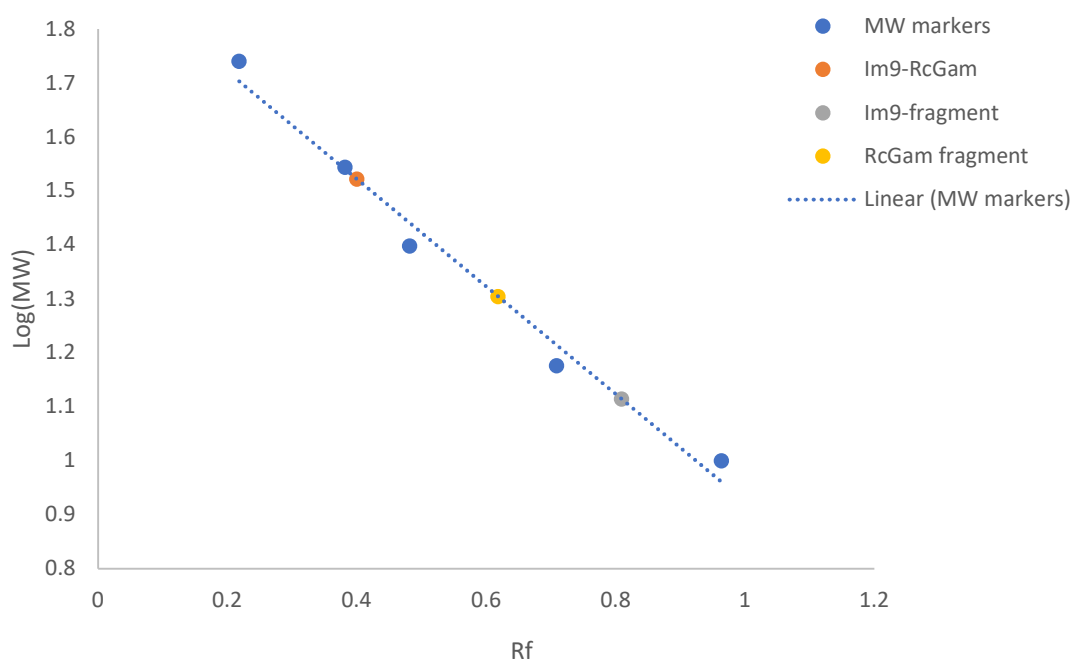
### 5.2.5 Analysing non-target proteins eluted from FPLC purification of target Im9-MuGam homologue proteins

For Im9-KpGam, Im9-SsGam, Im9-SeGam, Im9-EcGam and Im9-RcGam there was a significant amount of unexpected non-target protein. In each case two major bands appeared that do not match the molecular weight of the target protein. When the MW of these non-target protein bands are estimated and added together the total MW was similar to the MW of the target Im9-MuGam homologue protein. The expected MW of the target protein Im9-RcGam is 31.87kDa, the estimated

## Production of recombinant MuGam homologue proteins

MW based on **Figure 28** is 33.26kDa with the two non-target protein bands having an estimated MW of 20.17kDa and 13.02kDa respectively (**Figure 28**). When the estimated MWs of the non-target protein bands are added the total estimated MW is 33.19kDa which is extremely similar to the estimated MW of Im9-RcGam. This suggests that the non-target protein bands have resulted from proteolytic breakdown likely at a specific sequence.

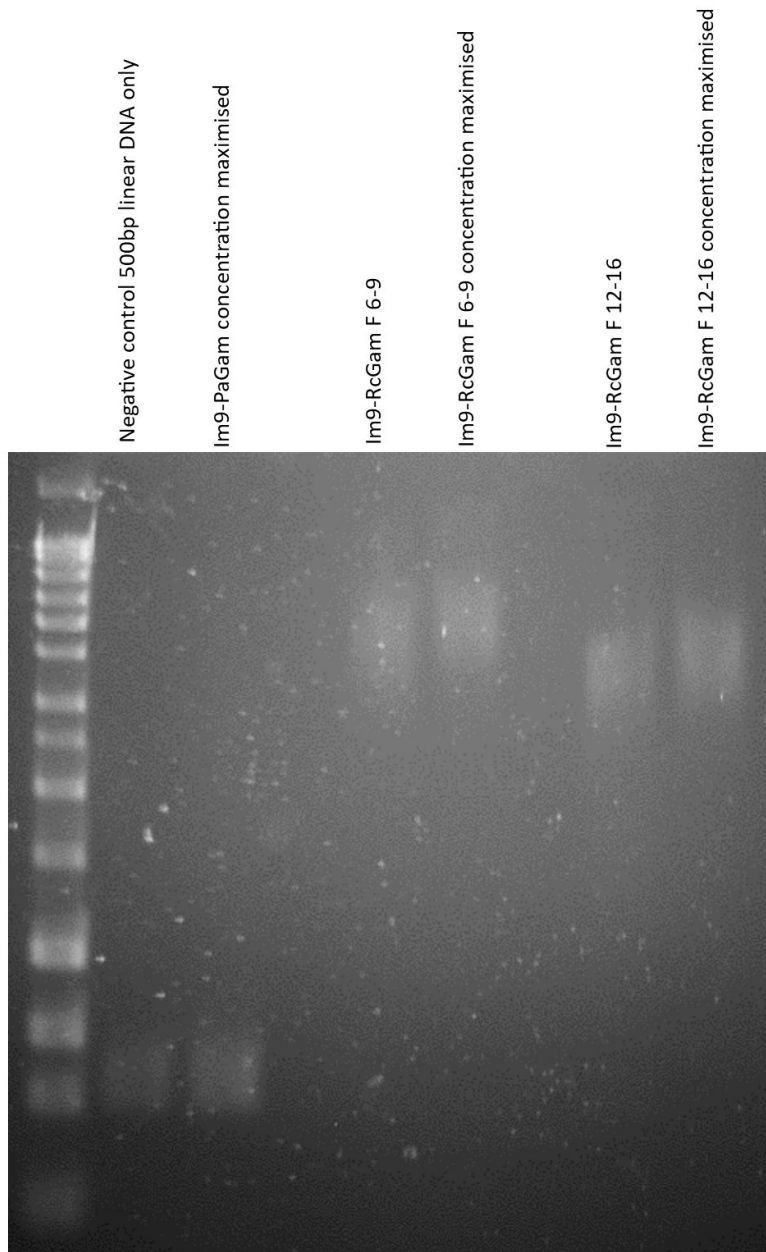
Based on what we know about Im9-MuGam homologues such as Im9-HiGam we expect the protein to readily form a dimer in solution. The fragments likely correspond to degradation of the protein at a sequence between the Im9- solubility tag and the RcGam protein itself. The smaller fragment (13.02kDa estimated MW) will most likely be the Im9 solubility tag with the N-terminal 6x His tag, allowing it to anneal to the column and not be eluted during the wash step. The larger fragments (20.17kDa estimated MW) is therefore at least the majority of the RcGam protein that has successfully folded and formed a heterodimer with full-length Im9-RcGam monomers. This means that in the case of RcGam, fractions 14-17 will predominantly contain Im9-RcGam homodimers where fractions 5-13 will contain a mixture of Im9-RcGam homodimers, Im9-RcGam – RcGam heterodimers and free Im9 protein. The protein bands are labelled on **Figure 27** to illustrate the positions of the Im9-RcGam protein, RcGam protein fragment and Im9 protein fragment discussed here.



**Figure 29.** The MW calibration graph used to estimate the MW of the Im9-RcGam protein and the two protein fragments as seen in **Figure 26**. The  $\text{Log}_{10}$  of the molecular weight for each molecular weight marker was calculated and plotted against the  $R_f$  for each corresponding band on the gel from **Figure 26**, using **equation 1 (2.14)** the distance that the protein band travelled through the gel is converted to  $R_f$ . Using this graph and the equation for the linear regression ( $y = -0.9957x + 1.9202$ ) where  $R^2 = 0.9819$  the largest MW band (Im9-RcGam) (orange) was estimated to be 33.26kDa with the second (RcGam fragment) (yellow) estimated as 20.17kDa and the final band (Im9-fragment) (grey) was estimated at 13.02kDa.

## Production of recombinant MuGam homologue proteins

To analyse the function of the purified protein fractions for RcGam, the concentration was measured after the pooled protein fractions were concentrated using a 5,000 MWCO centrifugal concentrator. The two pooled fractions were analysed by EMSA to investigate whether the two protein samples could bind to linear DNA. Im9-RcGam fractions 6-9 should contain a mixture of Im9-RcGam and RcGam fragment heterodimers while pooled fractions 12-16 should only contain Im9-RcGam homodimers.



**Figure 30.** 1.5% (w/v) agarose-TAE gel image showing Im9-RcGam fractions 6-9 and fractions 12-16 analysed by EMSA using 500bp linear DNA. Both samples were able to cause a shift in the mobility of the DNA suggesting the Gam homolog protein can bind to the linear DNA. The shift in the mobility of the DNA was different between the two fraction samples, with fractions 6-9 causing the DNA to shift to a greater MW than the sample containing fractions 12-16. Im9-PaGam did not appear to bind to the linear DNA and induce a shift in the mobility of linear DNA.

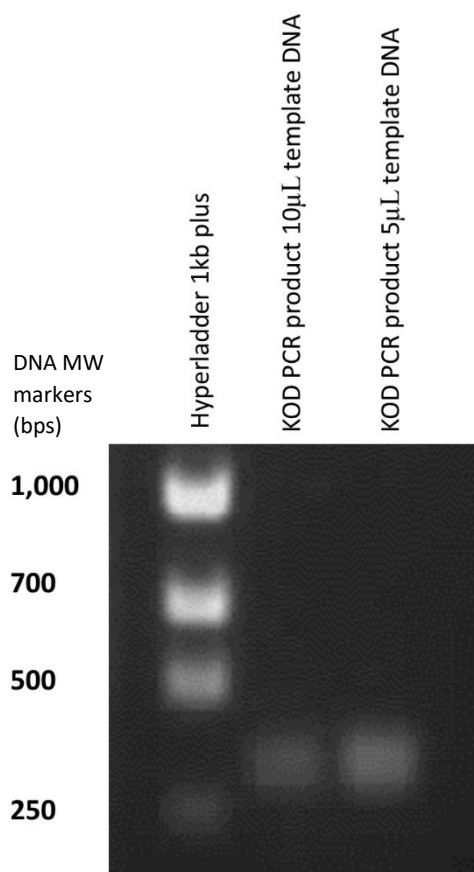
## Production of recombinant MuGam homologue proteins

The difference in the mobility shift for the linear DNA suggests that the samples of Im9-RcGam are different and supports the idea that the larger fragment is the RcGam fragment and it is forming an Im9-RcGam-RcGam fragment heterodimer that is able to bind to linear DNA. The greater DNA shift in mobility may therefore be a result of the Im9-RcGam-RcGam fragment heterodimer having a smaller DNA binding footprint compared to the Im9-RcGam homodimer resulting in more dimer units binding and the greater MW.

### 5.2.6 Producing pET15b-HI1450 recombinant expression plasmid

The HI1450 gene is present in the genome of *Haemophilus influenzae* Rd KW20, and the gene was amplified from the genome using PCR with primers designed to bind to either side of the hi1450 gene. The PCR amplicon was purified from an agarose gel using the Wizard® SV PCR clean up system (Promega Corporation).

The KOD PCR resulted in some slightly smeared bands that appeared in the expected area for a 324 bp gene between the 250 bp and 500 bp molecular weight markers in the DNA ladder. The gel shows that the KOD PCR was able to produce linear DNA fragments that are in the expected region for the 324 bp HI1450 gene.



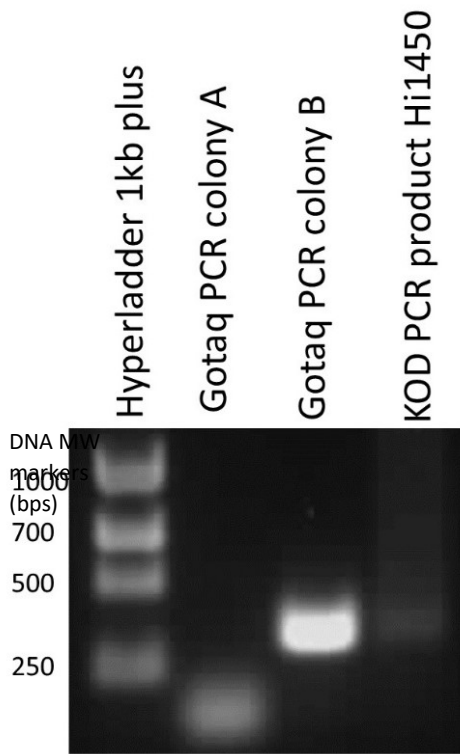
**Figure 31.** Products of the KOD PCR for two samples using the conditions listed above with lane 2 using 10µL of template DNA (therefore only 28µL PCR water). The band in lane 3 appears more intense than the band in lane 2. Other conditions were also tested but none showed the expected band in the 250-500bp region. The bands in lanes 2 and 3 both appear to be in the correct region for a 324bp fragment of DNA.

The resulting bands were purified as above, ligated into the Zero® Blunt® TOPO® vector plasmid (Topo) and transformed into *Escherichia coli* DH5-α cells using heat shock. Only two colonies grew on the plate after overnight incubation at 37°C. Both colonies were used to investigate the success of the ligation and transformation. Using the same primers for the initial amplification of the hi1450



## Production of recombinant MuGam homologue proteins

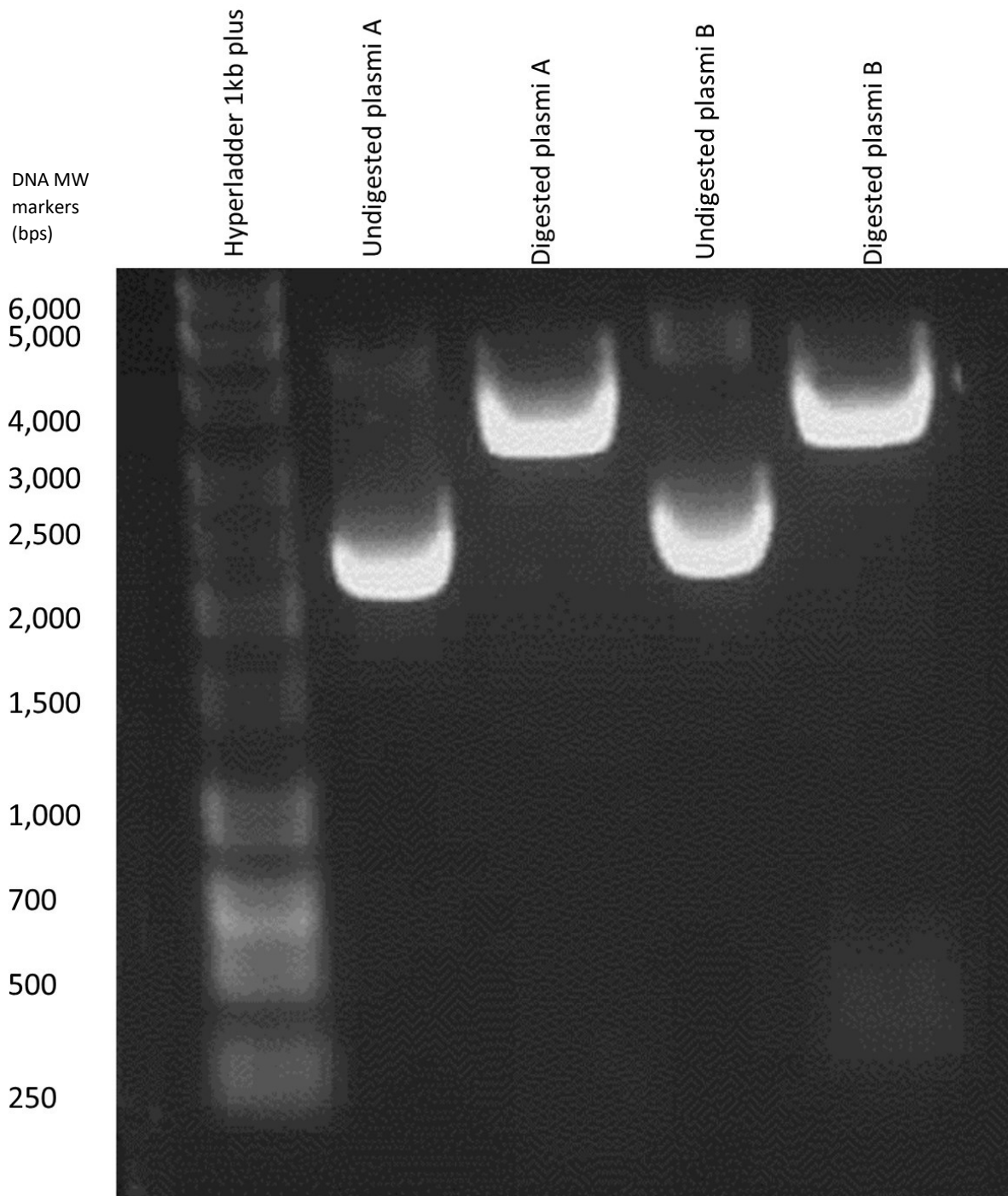
gene from the *H. influenzae* Rd KW20 genome, a diagnostic PCR confirmed the presence of the hi1450 gene in the plasmid recovered from the transformed colonies.



**Figure 32.** Resulting bands from diagnostic Gotaq PCR from two colonies A and B confirming the presence of the HI1450 gene in colony B. The band produced from colony A is well below the 250 bp molecular weight marker suggesting that the transformation of the correct HI1450-Topo plasmid did not occur. The band produced from colony B lines up with the band produced from the original KOD PCR product for HI1450 and is in the region we would expect for the length of the gene (324bp).

The result indicates that the *E. coli* DH5- $\alpha$  cells from colony B are transformed with the HI1450-Topo recombinant plasmid as the PCR has produced a bright band at the expected length and therefore successfully bound to the forward and reverse primers designed to bind to the HI1450 gene. This result is reinforced by the use of the diagnostic double restriction digest, where the plasmid from colony B produced a linear plasmid band (approx. 3,500bps) and a band corresponding to the linear hi1450 gene (approx. 350bps) (**Figure 31**).

## Production of recombinant MuGam homologue proteins



**Figure 33.** Diagnostic double restriction digestion of plasmids from colonies A and B. The expected HI1450 -Topo plasmid should have a 3.5kb linear Topo plasmid band and the 324bp HI1450 gene as the two products. This is faintly seen from the plasmid retrieved from colony B.

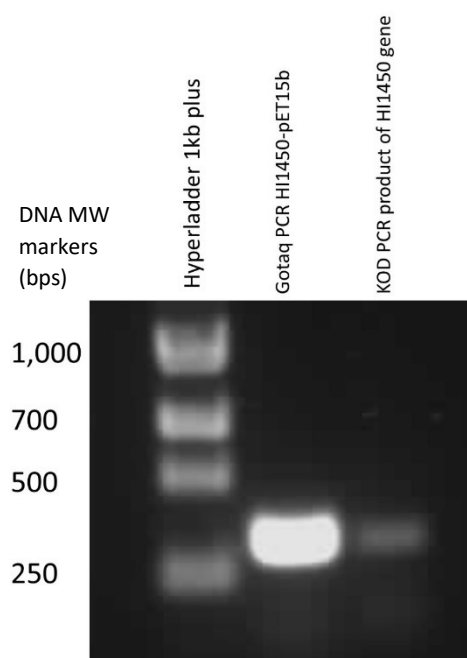
The diagnostic double restriction digest of purified plasmid from colony B produced a linear Topo plasmid at approx. 3.5 kbp and a fragment between 250 and 500 bp consistent with the HI1450 gene insert (324 bp). Colony A shows no signs of an inserted fragment from a successful transformation of the expected HI1450-Topo recombinant plasmid.

## Production of recombinant MuGam homologue proteins

The purified HI1450-Topo plasmid from colony B was digested using the appropriate restriction endonuclease enzymes and ligated into a pET15b vector plasmid digested with the same enzymes. The ligation product was transformed into *E. coli* DH5- $\alpha$ .

PCR was used to confirm the presence of the expected HI1450 gene in the plasmid transformed into *E. coli* DH5-alpha cells. The intense DNA band at the expected molecular weight (324bp) indicated the correct insert had been cloned. The product of the PCR suggests that the cells were transformed with the correct recombinant HI1450-pET15b plasmid as indicated by the band in the same position as that of the KOD PCR product used initially to form the recombinant plasmid.

This information tells us that the HI1450 gene is present in the cells of the transformed colony though it gives little information whether the recombinant plasmid has ligated as expected. The diagnostic double restriction digest and sequencing data give a better idea of the success of the ligation.

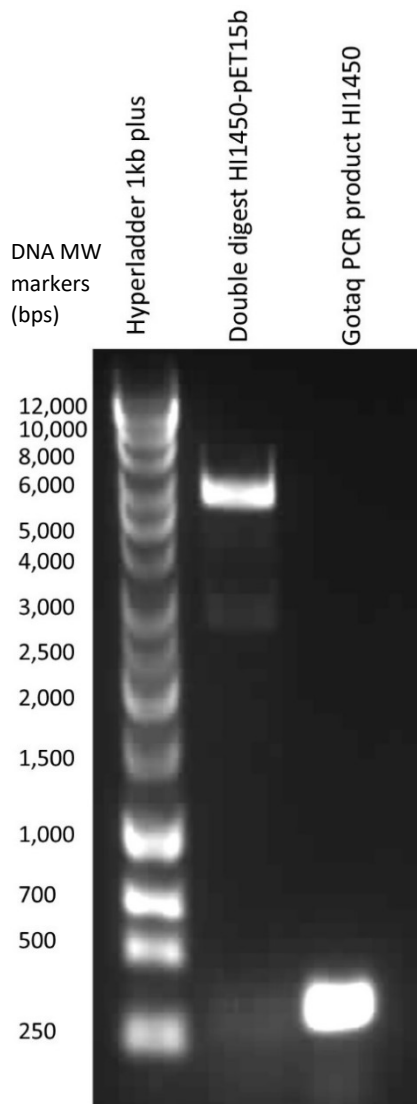


**Figure 34.** 1% (w/v) agarose + TAE gel image of the diagnostic Gotaq PCR product of the transformed *E. coli* DH5-alpha cell colony with the HI1450-pET15b ligation product (lane 2). The presence of the strong band in the same position as the original product of the KOD PCR from *H. influenzae* Rd KW 20 (which had been diluted 10-fold). The band migrates in the region between 250 and 500bp which is indicative of the 324bp HI1450 gene.

The resulting gel image (**Figure 33**) for the double digest of HI1450-pET15b purified plasmid shows bands at approx. 6,000bp, 3,000bp and one at the same position as the Gotaq PCR product which was indicative of the HI1450 gene. The intensity of the band at approximately 6,000bp is very strong and it matches the expected length for the linear pET15b plasmid (5,708bp).

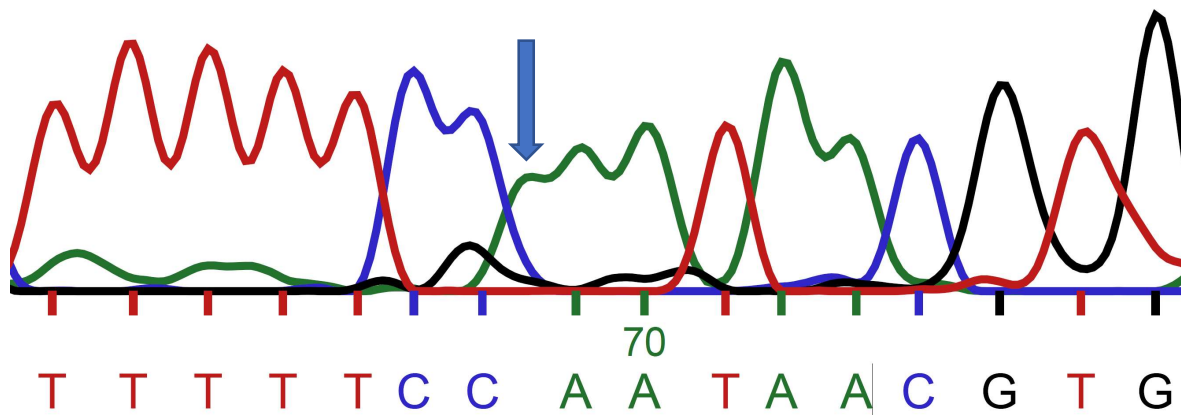
The weak band at around 3,000bp is not an expected product of the reaction and most likely represents the product of a single cut somewhere in the centre of the linear pET15b plasmid producing two bands of similar length, possibly due to star activity of one of the restriction enzymes. This can be caused by the presence of a nucleotide sequence that is similar but not identical to the definite restriction site for the enzyme resulting in low affinity binding and enzymatic activity at the site. This can occur as a result of an overnight (17.5 hour) incubation time (Wei *et al.*, 2008).

## Production of recombinant MuGam homologue proteins



**Figure 35.** 1% (w/v) agarose + TAE gel image of the diagnostic double restriction digest of the miniprep purified HI1450-pET15b plasmid. This result indicated a successful digestion of the circular plasmid either side of the HI1450 gene to yield a band for linear pET15b plasmid and evidence of a band at the same molecular weight as the Gotaq PCR product representing the HI1450 gene.

The sequencing results confirmed the structure of the expected recombinant plasmid as shown in **Figure 36**. The results received for the sequence were not exactly as expected. The sequencing data produced from the top strand appeared to match the expected gene insert, while the reverse strand did not yield the expected sequence. In the reverse sequence data at the terminal end of the gene there appeared to be a frame shift that caused the loss of the functional stop codon at the end of the gene and a change to the amino acid sequence. The problem originated from a missing base call in the sequencing data where a short run of adenine nucleotides was not completely identified. When the missing base is added back into the sequence at this location the expected sequence is restored.

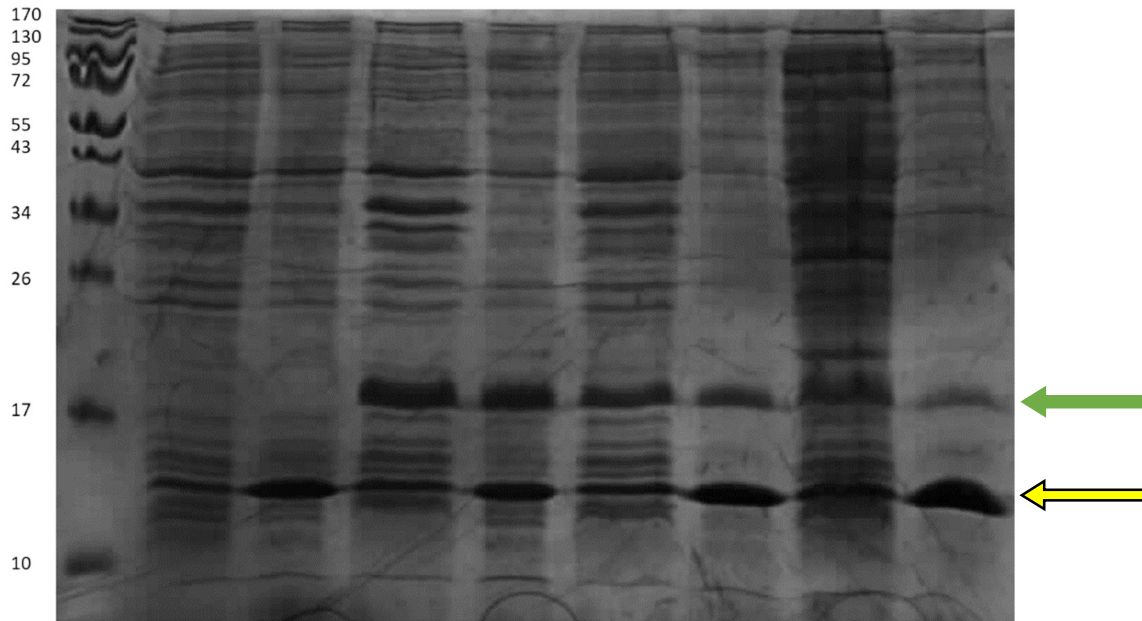


**Figure 36.** The sequencing data output received in the region with the unexpected anomalous sequence. The blue arrow indicates the location of a peak that is not identified, which resulted in an incorrect sequence. If the A nucleotide is added back into the sequence at this location the correct protein is obtained from the sequence.

### 5.2.7 Overexpression of HI1450

The expected molecular weight of the HI1450 protein with the 6x His tag and thrombin site was 14.69kDa which is very similar to the molecular weight of lysozyme (14.307 kDa). This makes it very difficult to distinguish the IPTG induced protein of interest from lysozyme in **Figure 37**. The most noticeable induced protein was one at around 17-18kDa (indicated with the green arrows on **Figure 37**) which was not expected for the HI1450 protein. The identity of this band is unknown. There is no evidence of these bands in the pre-induced samples and intense bands in this position for the post-induction samples in both soluble and insoluble fractions, highly indicative of the expected IPTG induction.

The intensity of the bands between the 10 and 17kDa (indicated with yellow arrows in **Figure 37**) molecular weight marker bands is consistently more intense in the soluble fractions than the insoluble fractions and the mobilities do not match up exactly. It is possible that the band for the lysozyme and HI1450 proteins are combining and/or overlapping. Though it is important to note that it appears as though this trend extends to the pre-induced samples, and this may be due to difference in the lysozyme concentration between the soluble and insoluble fractions. This trend was not visible when analysing Im9-MuGam homologues.

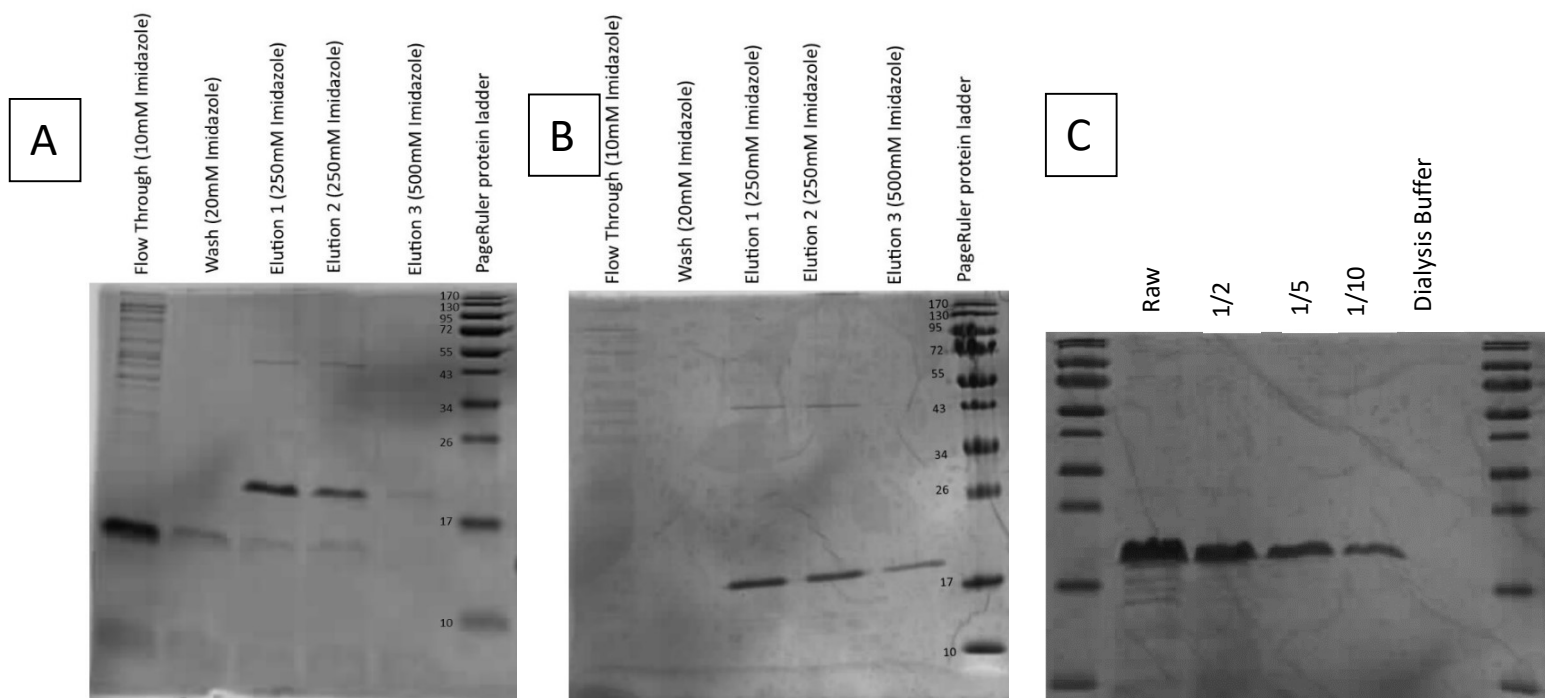


**Figure 37.** 15% (w/v) polyacrylamide-SDS gel electrophoresis analysis of HI1450 over-production using IPTG induction and *E. coli* BL21(DE3) cells transformed with HI1450-pET15b plasmid. The expected molecular weight of the 6xHis tagged HI1450 protein is 14.69kDa. The bands labelled with the yellow arrow are in the expected position and indicate the presence of the HI1450 protein, while the green arrow labels protein bands that appear to have resulted from induced overproduction (approx. 18kDa). The protein band indicated with the yellow arrow is characteristic of the lysozyme used to lyse the *E. coli* BL21(DE3) cells. This suggests that the band labelled with the green arrow is the overproduced HI1450 protein which showed abnormal mobility for its predicted MW.

In order to confirm that the target HI1450 protein was being overproduced the cells were induced and the cell lysates were purified using the HisPur purification system (**Materials and Methods – 2.9.4**). This confirmed that the target protein with the 6x His tag was being overproduced and that it showed a consistent abnormal mobility in the SDS-PAGE gel.

If the expected HI1450 protein was produced in the transformed *E. coli* BL21(DE3) using IPTG induction, then it should bind with high affinity to a Ni<sup>2+</sup> spin column containing a Nickel charged affinity resin designed to bind and trap proteins with the 6xHis tag. This would separate it from the other proteins including lysozyme and therefore be visible in the elution fraction after SDS-PAGE. To ensure that we were able to distinguish between the target protein and the lysozyme, the lysis protocol prior to purification was done with (**Figure 38A**) and without lysozyme (**Figure 38B**) at a final concentration of 1 mg/mL.

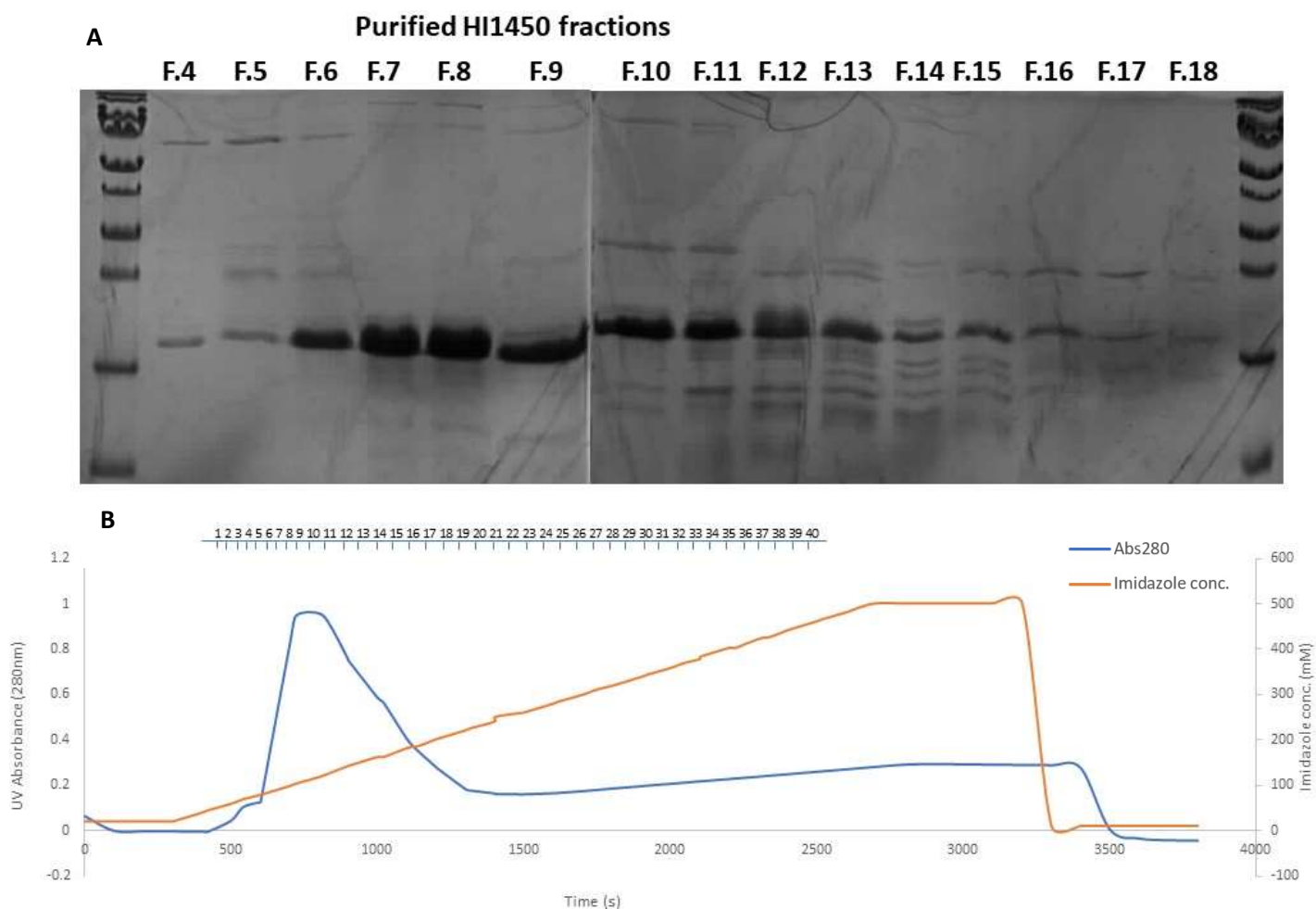
## Production of recombinant MuGam homologue proteins



**Figure 38.** 15% (w/v) polyacrylamide-SDS gel electrophoresis was used to analyse the purified target 6x His tagged HI1450 protein **A:** shows the protein bands purified by HisPur™ spin column with the use of 1 mg/ml lysozyme in the protocol. **B:** protein bands purified in the same manner without the addition of lysozyme. **C:** Purified and pooled fractions serially diluted to get a qualitative assessment of the protein purity. The images prove that in the presence and absence of Lysozyme the target 6x His tagged HI1450 protein was purified and migrated with an apparent MW of approx. 18kDa.

The dominant band of the resulting purification was approx. 18kDa, thus overproduction of this protein was induced by IPTG and it bound to the immobilised Ni<sup>2+</sup> on both the spin column and chromatography column. The evidence suggested that this was the HI1450 protein and that the mobility through the gel is anomalous for the 14.69kDa protein.

The protein was therefore overproduced and purified in order to be used to assess the potential for a protein-protein interaction between Im9-HiGam and HI1450. The purification was successful though the analysed protein fractions showed a number of non-target proteins that likely bound to the HI1450 protein during Ni<sup>2+</sup> affinity chromatography and eluted with the target protein. This could be evidence that the HI1450 protein was folding correctly and binding to numerous DNA binding proteins.



**Figure 39. A:** 10% (w/v) polyacrylamide-SDS gel image showing the protein fractions analysed to identify the fractions which contain the most target HI1450 protein and the least contaminant protein. The chosen fractions were then pooled to investigate protein function. **B:** The UV absorbance at 280nm (blue trace) was used to identify when protein eluted from the column as the imidazole concentration (orange) increased with the number of each fraction shown above the graph. The Hi1450 eluted in fractions 4-18. Only fractions 6-10 were pooled and dialysed because fractions 4-5 and 11-18 contain a significant amount of non-target protein.

### 5.3 Discussion

The Im9-MuGam homologues that were successfully overproduced in *E. coli* K-12 MG1655 cells were all successfully purified using FPLC. It was common to see the target protein breakdown into two protein fragments which most likely correspond to a small fragment containing the 6x His tagged Im9 protein and a larger MuGam homologue containing fragment. This would suggest that there is a site in the junctional sequence that is resulting in instability or targeted proteolysis. The sequence that may be causing this has not been identified and there does not appear to be a sequence present that corresponds to a common target site for bacterial proteolysis. Further probing of the primary and secondary structures of the expected region of degradation could provide more information and explain the consistent fragmentation of the Im9-MuGam homologue proteins. The protein degradation is not seen in all of the Im9-MuGam homologue proteins, Im9-ApGam and Im9-HiGam show very little evidence that the protein is degrading at the same rate as



## Production of recombinant MuGam homologue proteins

the other Im9-MuGam homologue proteins. The site of degradation is most likely somewhere in the MuGam homologue amino acid sequence as opposed to the Im9 sequence or the linker sequence as the degradation is only observed in some of the Im9-MuGam homologues including Im9-HiGam, Im9-ApGam and Im9-PaGam.

The transformation of the *E. coli* BL21(DE3) with recombinant HI1450-pET15b was successful and the plasmid contains the expected sequence. The protein was successfully overproduced in the cells and purified by Ni<sup>2+</sup>-affinity chromatography. The anomalous mobility through the SDS-PAGE gel was most likely the result of the high density of negatively charged residues present on the HI1450 protein. This results in a net-negative charge and inconsistent SDS binding (Malhotra and Sahal, 1996). The inconsistency of SDS binding results in inconsistent charge across the denatured protein during SDS-PAGE analysis and therefore reducing the mobility of the protein through the polyacrylamide gel.

## Results

# Chapter 6 – Characterisation of DNA Binding Properties for Gam Homologues

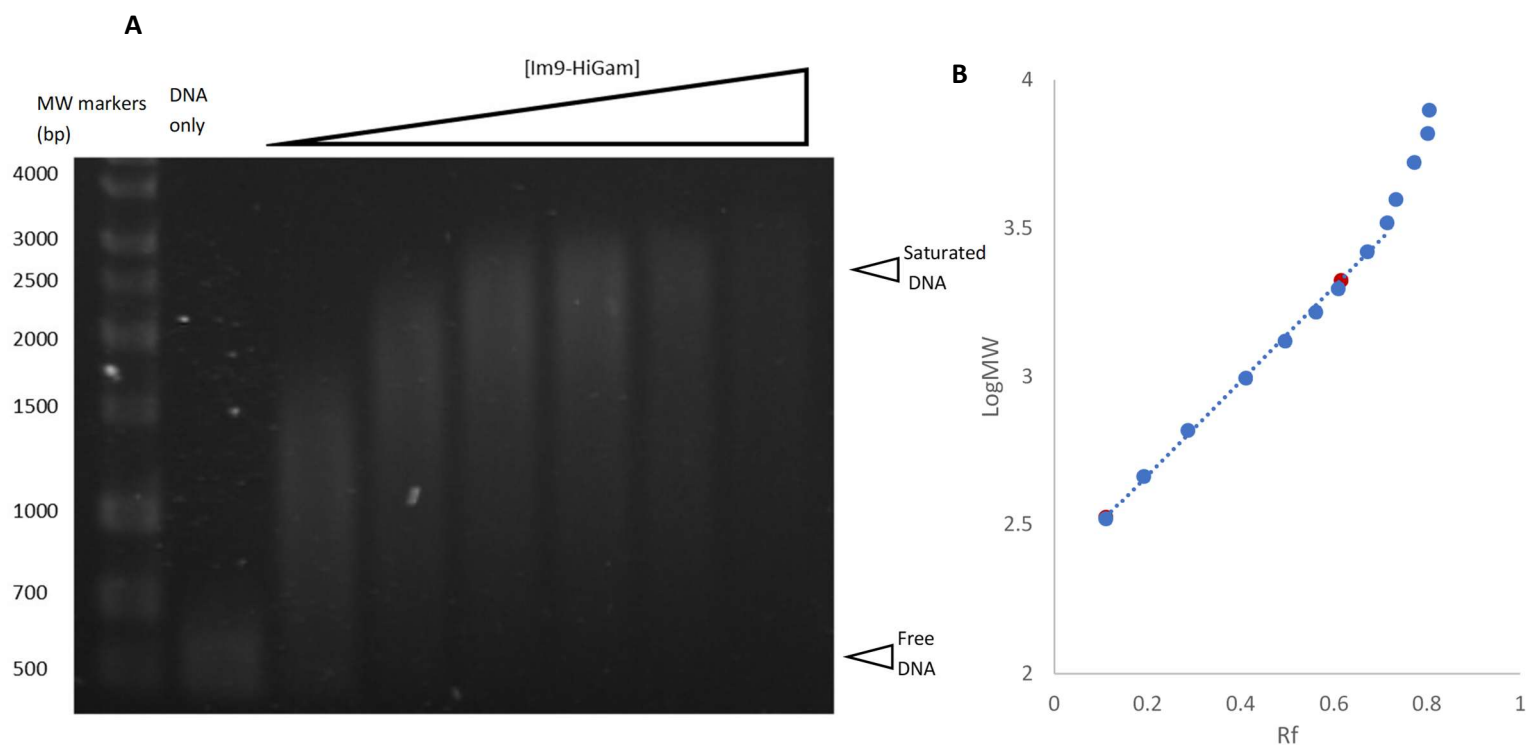
## **6.1. Introduction**

The MuGam homologues identified in different bacterial genomes are consistently conserved in either complete or incomplete prophage regions. If these genes are transcribed and the gene products produced within the host cells (relative to other co-expressed prophage genes) then it suggests that the genes have been conserved because the gene product is beneficial to the host. The DNA binding characteristics of the cloned MuGam homologues have been analysed to determine if the gene products can interact with linear DNA.

Electrophoretic mobility shift assays (EMSAs) are used to investigate the function of the MuGam homologue proteins as it can be used to show whether the proteins are able to bind to the DNA based on a shift in the mobility of the DNA through an agarose-TAE gel during electrophoresis. The shift in the DNA mobility results from the increase in molecular weight (MW). The experiment also produced semi-quantitative results which can be used to determine binding characteristics of the protein such as the DNA binding footprint.

### **6.1.1 Action of HiGam in the presence of linear DNA**

Unpublished data has already shown that in the presence of HiGam linear DNA will move more slowly through a agarose-TAE gel during electrophoresis. This shift in mobility was not observed when introducing HiGam to circular plasmid DNA. This confirms that HiGam will bind exclusively to linear DNA due to the presence of exposed ends. It also showed that when HiGam is in excess that multiple HiGam dimers will bind to the linear DNA. This implies that the protein will bind to the end and slide along the DNA to allow other dimers to bind until the strand is completely covered by protein. The shift in the MW of the DNA can be used to calculate the number of HiGam dimers that bound to the linear DNA at saturation and therefore also the size of the DNA binding footprint for the HiGam dimer.



**Figure 40.** A) Electrophoretic mobility shift assay of 0.3 $\mu$ M (6.67 $\mu$ g/mL) 500 base pair (bp) linear DNA on a 1.5% (w/v) agarose-TAE gel. Shows the shift in mobility of DNA in the presence of increasing Im9-HiGam protein concentration ([Im9-HiGam]) within the range of 0.45 $\mu$ M - 3.6 $\mu$ M. The increase in MW of the DNA fragments as a result of the DNA-protein interaction leads to progressively lower mobility through the gel as protein concentration increases. B) The relative mobility of the known MW marker bands were subject to linear regression to estimate the MW of free and fully protein-saturated DNA bands. This used the  $\text{Log}_{10}(\text{MW})$  in kDa plotted against the  $R_f$  as calculated by **equation 1 (2.14)**. Linear regression of the  $R_f$  data for the MW ladder bands between 500 – 5000bps ( $y = 1.587x + 2.35$ ,  $R^2 = 0.9968$ ) was used to estimate the  $\text{log}(\text{MW})$  of the free and saturated DNA plotted as red points on the graph. When the DNA was fully saturated with Im9-HiGam the MW increased by 1778.81kDa as a result of bound Im9-HiGam (dimeric MW estimated from doubling the monomeric MW = 30.64kDa), therefore an estimated 29 Im9-HiGam dimers bound to the 500bp linear DNA. The dimeric Im9-HiGam has a binding footprint of approximately 17 bp (approx. 58.62 $\text{\AA}$ ).

## 6.2. Results

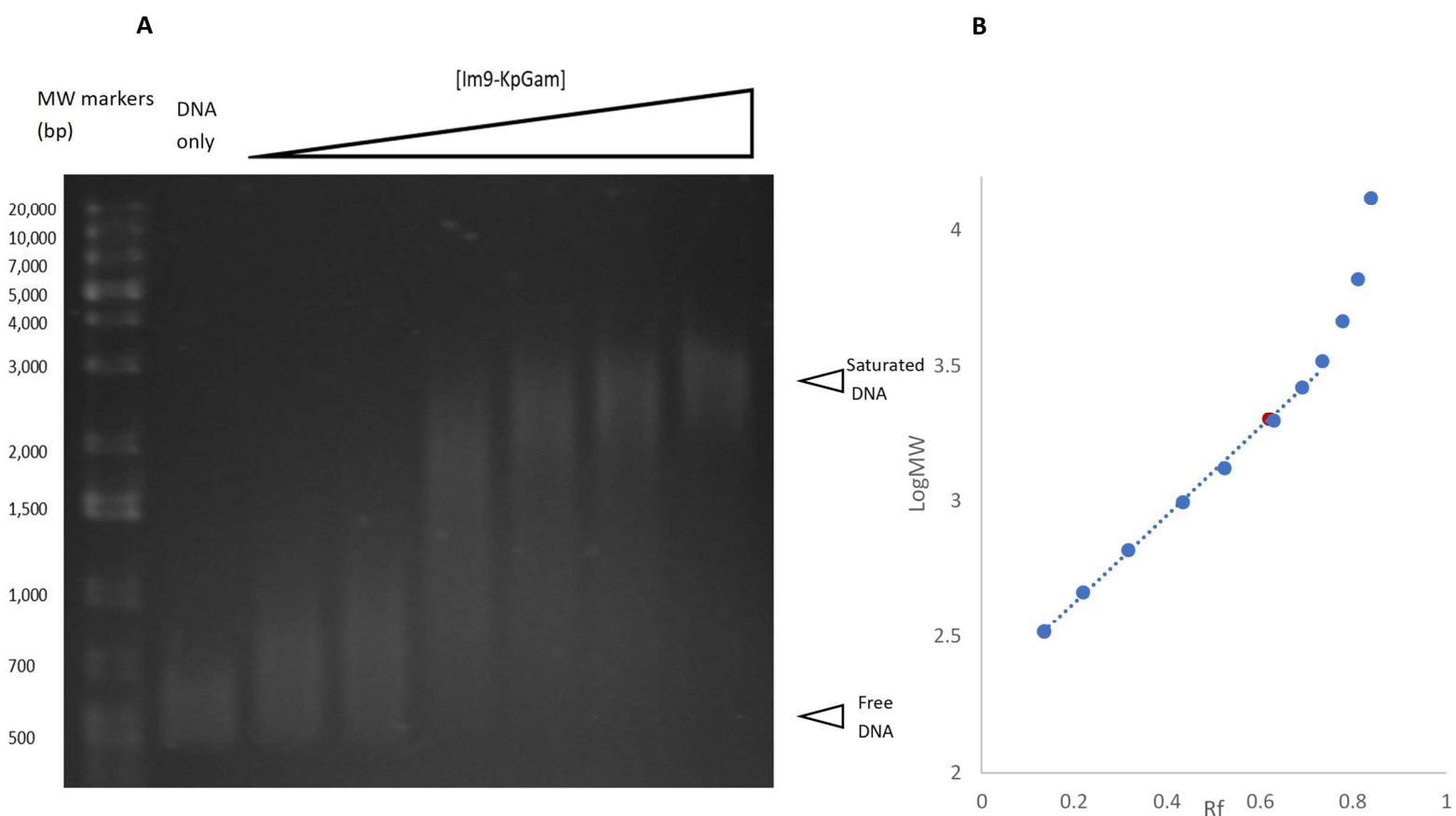
The function of the HiGam protein and other MuGam homologues *in vitro* can be used to infer potential function *in vivo*. Here the MuGam homologues are shown to bind to linear DNA *in vitro* as evidence by the shift in mobility of linear DNA through 1.5% agarose + TAE gel. However, it is necessary to determine if the protein is present in the cells under natural (non-induced) circumstances. If the protein provides benefit to the survival of the cell it is reasonable to assume that the protein is present in reasonable abundance in the soluble cell matrix.

### 6.2.1 Electrophoretic mobility shift assay (EMSA) to determine Gam homologue function

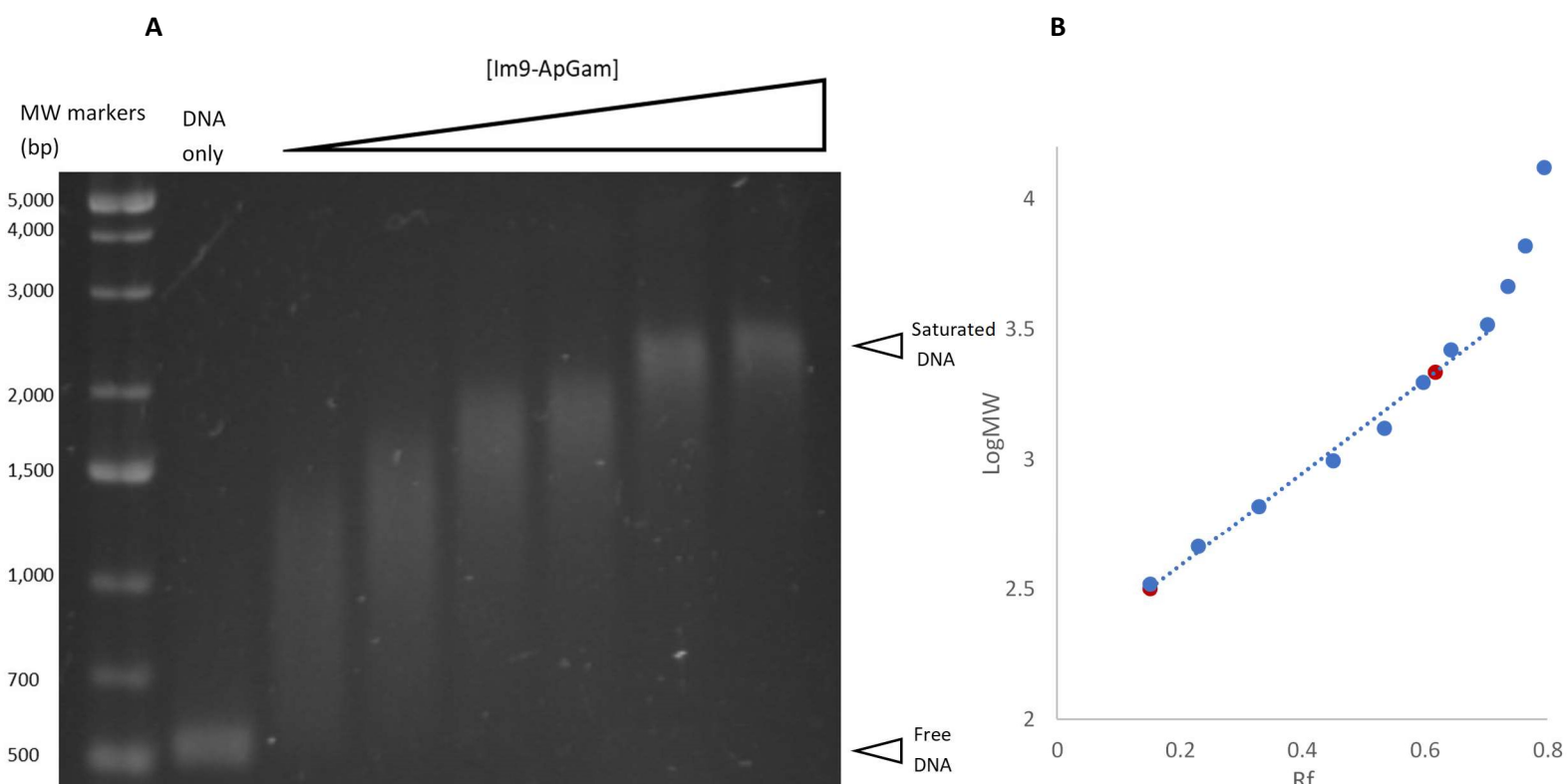
Showing that other MuGam homologue proteins with different amino acid sequences are functional when overproduced and purified from *Escherichia coli* K-12 MG1655 would support the hypothesis that there is some evolution pressure to retain the functionality of these gene products.

## Characterisation of DNA Binding Properties for Gam Homologues

It also provides an opportunity to find a homologue that can produce a dimeric crystal structure, which so far has not been found.

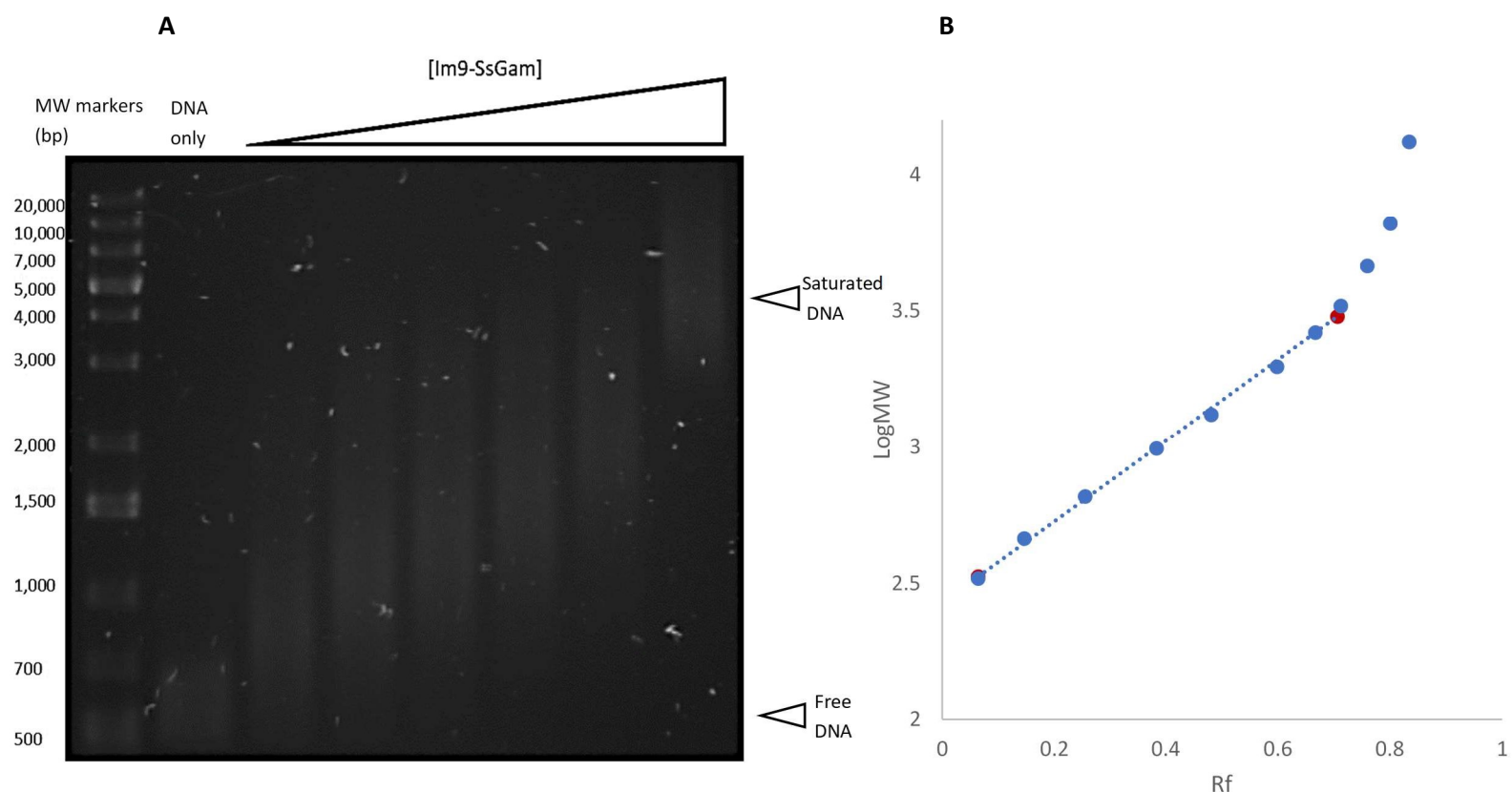


**Figure 41.** A) Electrophoretic mobility shift assay of  $0.3\mu\text{M}$  ( $6.67\mu\text{g}/\text{mL}$ ) 500bp linear DNA on a 1.5% (w/v) agarose-TAE gel. Shows the shift in mobility of DNA in the presence of increasing Im9-KpGam protein concentration ( $[\text{Im9-KpGam}]$ ) within the range of  $0.45\mu\text{M}$  -  $3.6\mu\text{M}$ . The increase in MW of the DNA fragments as a result of the DNA-protein interaction leads to progressively lower mobility through the gel as protein concentration increases. B) The relative mobility of the known MW marker bands were subject to linear regression to estimate the MW of free and fully protein-saturated DNA bands. This used the  $\log_{10}(\text{MW})$  in kDa plotted against the  $R_f$  as calculated by **equation 1 (2.14)**. Linear regression of the  $R_f$  data for the MW ladder bands between 500 – 5000bps ( $y = 1.6174x + 2.3012$ ,  $R^2 = 0.9973$ ) was used to estimate the  $\log(\text{MW})$  of the free and saturated DNA plotted as red points on the graph. When the DNA was fully saturated with Im9-KpGam the MW increased by 1678.18kDa as a result of bound Im9-KpGam (dimeric MW estimated from doubling the monomeric MW = 30.93kDa), therefore an estimated 27 Im9-KpGam dimers bound to the 500bp linear DNA. The dimeric Im9-KpGam has a binding footprint of approximately 19 bp (approx.  $62.97\text{\AA}$ ).



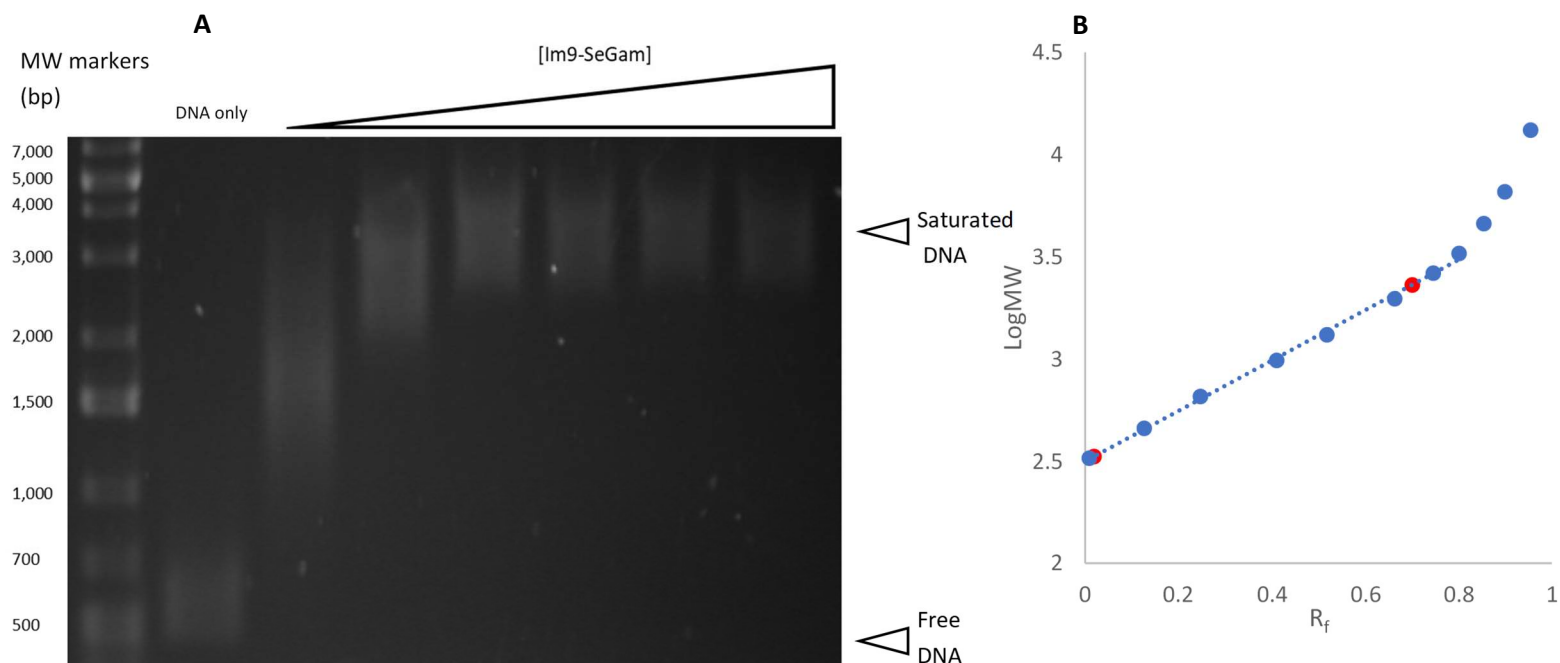
**Figure 42.** A) Electrophoretic mobility shift assay of  $0.3\mu\text{M}$  ( $6.67\mu\text{g}/\text{mL}$ ) 500bp linear DNA on a 1.5% (w/v) agarose-TAE gel. Shows the shift in mobility of DNA in the presence of increasing Im9-ApGam protein concentration ( $[\text{Im9-ApGam}]$ ) within the range of  $0.45\mu\text{M}$  -  $3.6\mu\text{M}$ . The increase in MW of the DNA fragments as a result of the DNA-protein interaction leads to progressively lower mobility through the gel as protein concentration increases. B) The relative mobility of the known MW marker bands were subject to linear regression to estimate the MW of free and fully protein-saturated DNA bands. This used the  $\log_{10}(\text{MW})$  in kDa plotted against the  $R_f$  as calculated by **equation 1 (2.14)**. Linear regression of the  $R_f$  data for the MW ladder bands between 500 – 5000bps ( $y = 1.7914x + 2.23$ ,  $R^2 = 0.9893$ ) was used to estimate the  $\log(\text{MW})$  of the free and saturated DNA plotted as red points on the graph. When the DNA was fully saturated with Im9-ApGam the MW increased by 1384.608kDa as a result of bound Im9-ApGam (dimeric MW estimated from doubling the monomeric MW = 30.53kDa), therefore an estimated 22 Im9-ApGam dimers bound to the 500bp linear DNA. The dimeric Im9-ApGam has a binding footprint of approximately 23 bp (approx.  $77.28\text{\AA}$ ).

## Characterisation of DNA Binding Properties for Gam Homologues



**Figure 43.** A) Electrophoretic mobility shift assay of  $0.3\mu\text{M}$  ( $6.67\mu\text{g}/\text{mL}$ ) 500bp linear DNA on a 1.5% (w/v) agarose-TAE gel. Shows the shift in mobility of DNA in the presence of increasing Im9-SsGam protein concentration ([Im9-SsGam]) within the range of  $0.45\mu\text{M}$  -  $3.6\mu\text{M}$ . The increase in MW of the DNA fragments as a result of the DNA-protein interaction leads to progressively lower mobility through the gel as protein concentration increases. B) The relative mobility of the known MW marker bands was subject to linear regression to estimate the MW of free and fully protein-saturated DNA bands. This used the  $\log_{10}(\text{MW})$  in kDa plotted against the  $R_f$  as calculated by **equation 1 (2.14)**. Linear regression of the  $R_f$  data for the MW ladder bands between 500 – 5000bps ( $y = 1.4871x + 2.4303$ ,  $R^2 = 0.9973$ ) was used to estimate the  $\log(\text{MW})$  of the free and saturated DNA plotted as red points on the graph. When the DNA was fully saturated with Im9-SsGam the MW increased by 2685.06kDa as a result of bound Im9-SsGam (dimeric MW estimated from doubling the monomeric MW = 30.50kDa), therefore an estimated 44 Im9-SsGam dimers bound to the 500bp linear DNA. The dimeric Im9-SsGam has a binding footprint of approximately 11 bp (approx.  $38.62\text{\AA}$ ).

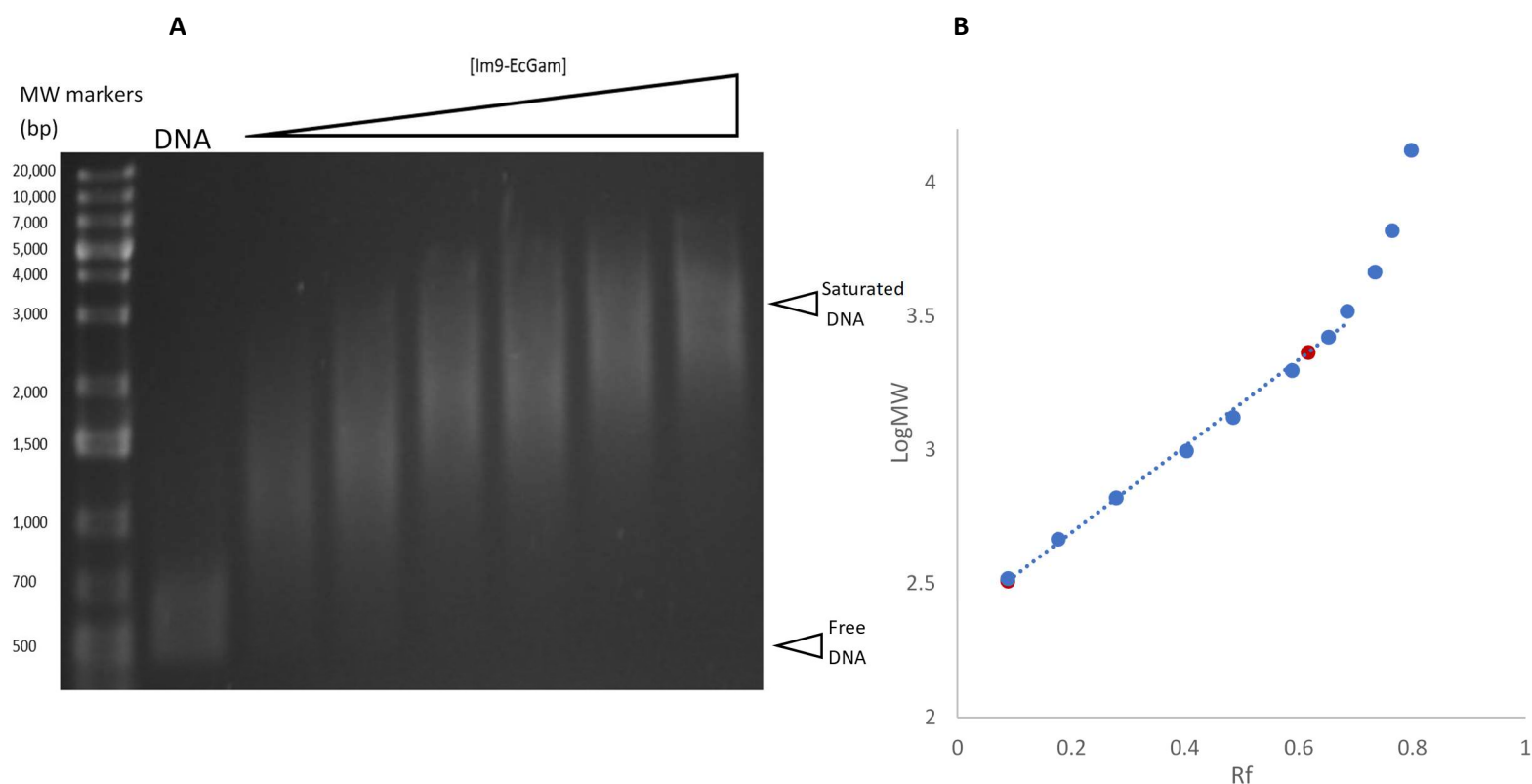
## Characterisation of DNA Binding Properties for Gam Homologues



**Figure 44.** A) Electrophoretic mobility shift assay of  $0.3\mu\text{M}$  ( $6.67\mu\text{g/mL}$ ) 500bp linear DNA on a 1.5% (w/v) agarose-TAE gel. Shows the shift in mobility of DNA in the presence of increasing Im9-SeGam protein concentration ( $[\text{Im9-SeGam}]$ ) within the range of  $0.45\mu\text{M}$  -  $3.6\mu\text{M}$ . The increase in MW of the DNA fragments as a result of the DNA-protein interaction leads to progressively lower mobility through the gel as protein concentration increases. B) The relative mobility of the known MW marker bands was subject to linear regression to estimate the MW of free and fully protein-saturated DNA bands. This used the  $\text{Log}_{10}(\text{MW})$  in kDa plotted against the  $R_f$  as calculated by **equation 1 (2.14)**. Linear regression of the  $R_f$  data for the MW ladder bands between 500 – 5000bps ( $y = 1.6168x + 2.3659$ ,  $R^2 = 0.9958$ ) was used to estimate the  $\text{log}(\text{MW})$  of the free and saturated DNA plotted as red points on the graph. When the DNA was fully saturated with Im9-SeGam the MW is estimated to increase by 1977.37 kDa as a result of bound Im9-SeGam (dimeric MW estimated from doubling the monomeric MW = 30.53kDa), therefore an estimated 32 Im9-SeGam dimers had bound to the 500bp linear DNA. The dimeric Im9-SeGam has a binding footprint of approximately 16 bp (approx.  $53.14\text{\AA}$ ).

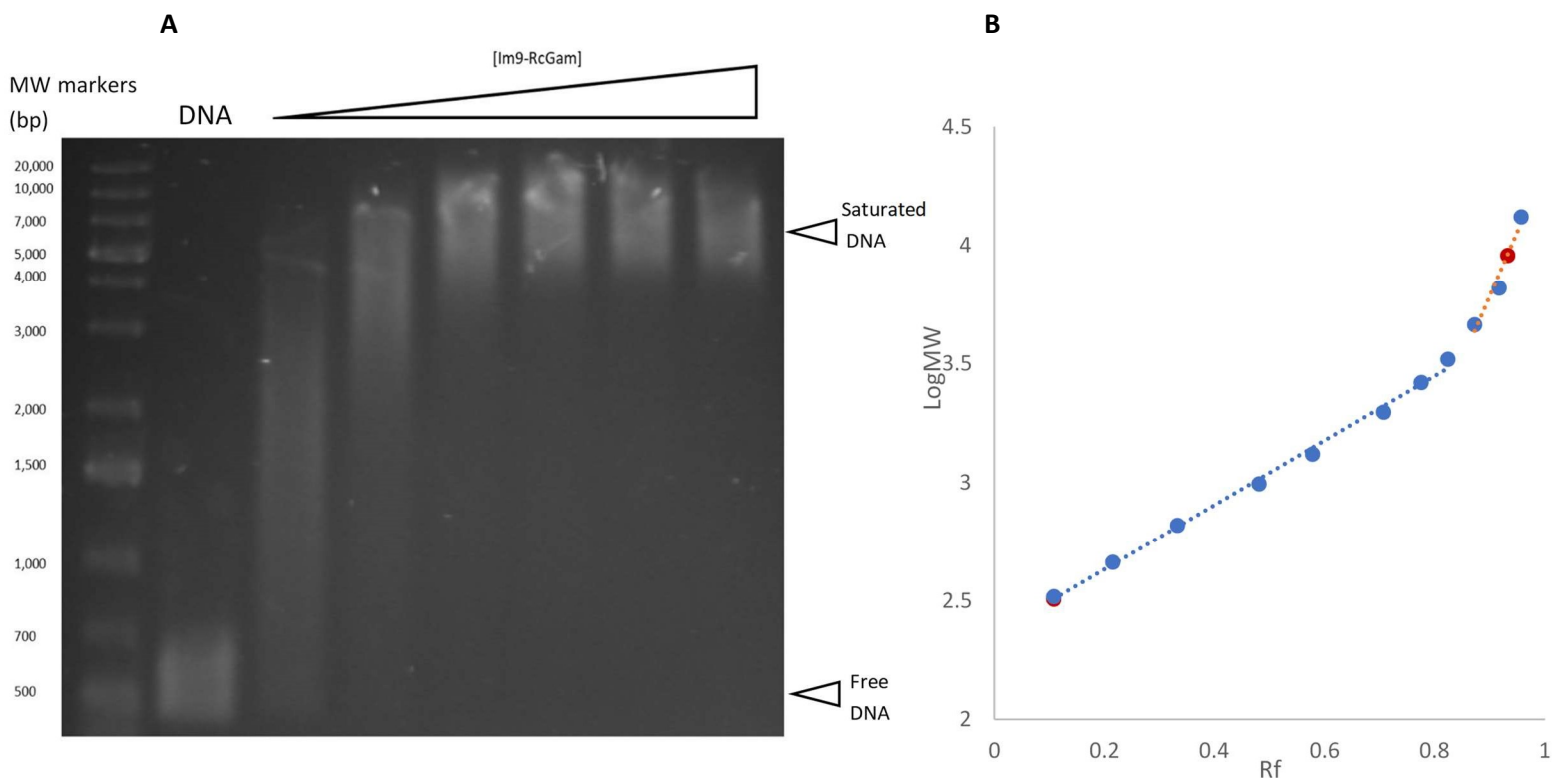


## Characterisation of DNA Binding Properties for Gam Homologues



**Figure 45.** A) Electrophoretic mobility shift assay of  $0.3\mu\text{M}$  ( $6.67\mu\text{g}/\text{mL}$ ) 500bp linear DNA on a 1.5% (w/v) agarose-TAE gel. Shows the shift in mobility of DNA in the presence of increasing Im9-EcGam protein concentration ( $[\text{Im9-EcGam}]$ ) within the range of  $0.45\mu\text{M}$  -  $3.6\mu\text{M}$ . The increase in MW of the DNA fragments as a result of the DNA-protein interaction leads to progressively lower mobility through the gel as protein concentration increases. B) The relative mobility of the known MW marker bands was subject to linear regression to estimate the MW of free and fully protein-saturated DNA bands. This used the  $\log_{10}(\text{MW})$  in kDa plotted against the  $R_f$  as calculated by **equation 1 (2.14)**. Linear regression of the  $R_f$  data for the MW ladder bands between 500 – 5000bps ( $y = 1.2288x + 2.5041$ ,  $R^2 = 0.9975$ ) was used to estimate the  $\log(\text{MW})$  of the free and saturated DNA plotted as red points on the graph. When the DNA was fully saturated with Im9-EcGam the MW increased by 1992.269kDa as a result of bound Im9-EcGam (dimeric MW estimated from doubling the monomeric MW = 31.10kDa), therefore an estimated 32 Im9-EcGam dimers had bound to the 500bp linear DNA. The dimeric Im9-EcGam has a binding footprint of approximately 16 bp (approx.  $53.14\text{\AA}$ ).

## Characterisation of DNA Binding Properties for Gam Homologues

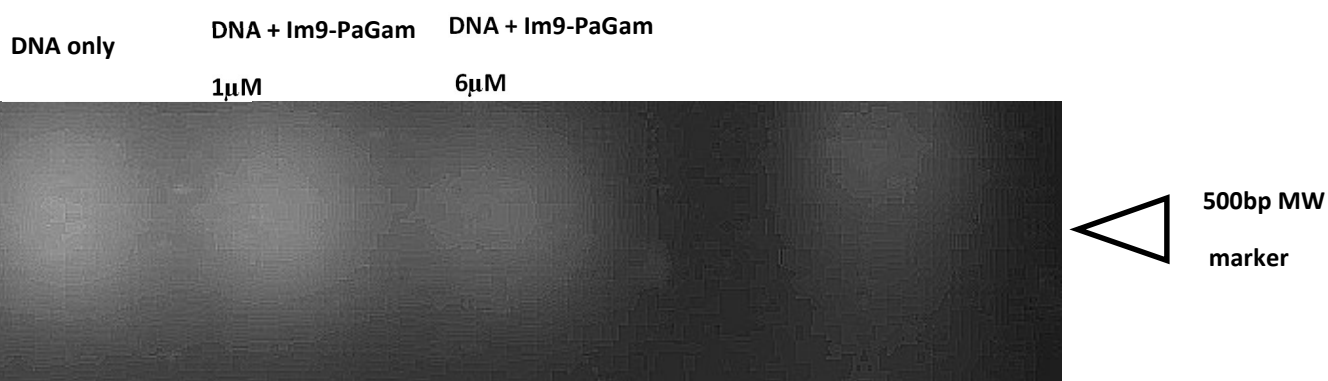


**Figure 46.** A) Electrophoretic mobility shift assay of  $0.3\mu\text{M}$  ( $6.67\mu\text{g}/\text{mL}$ ) 500bp linear DNA on a 1.5% (w/v) agarose-TAE gel. Shows the shift in mobility of DNA in the presence of increasing Im9-RcGam protein concentration ( $[\text{Im9-RcGam}]$ ) within the range of  $0.45\mu\text{M}$  -  $3.6\mu\text{M}$ . The increase in MW of the DNA fragments as a result of the DNA-protein interaction leads to progressively lower mobility through the gel as protein concentration increases. B) The relative mobility of the known MW marker bands was subject to linear regression to estimate the MW of free and fully protein-saturated DNA bands. This used the  $\log_{10}(\text{MW})$  in kDa plotted against the  $R_f$  as calculated by **equation 1 (2.14)**. Linear regression of the  $R_f$  data for the MW ladder bands between 500 – 5000bps was used to estimate the  $\log(\text{MW})$  of the free ( $y = 1.3586x + 2.3618$ ,  $R^2 = 0.9963$ ) and the  $\log(\text{MW})$  for the saturated DNA ( $y = 5.4305x - 1.1009$ ,  $R^2 = 0.9537$ ) plotted as red points on the graph. When the DNA was fully saturated with Im9-RcGam the MW increased by 3438.312kDa as a result of bound Im9-RcGam (dimeric MW estimated from doubling the monomeric MW = 31.50kDa), therefore an estimated 54 Im9-RcGam dimers had bound to the 500bp linear DNA. The dimeric Im9-RcGam has a binding footprint of approximately 9 bp (approx.  $32.71\text{\AA}$ ).

All of the recombinant MuGam homologues tested were able to bind to linear DNA *in vitro*. The shift in the mobility of the 500bp linear DNA was evident in each case as a result of increasing protein concentration. This suggests each protein produced a correctly folded and functional protein despite having different amino acid sequences.

There are differences in the results between the homologues investigated, with the total increase in MW varying from 1384.60kDa (ApGam) to 3438.312kDa (RcGam), and the concentration required to reach saturation. The MWs of each protein are slightly different which may explain these differences in the estimated DBNA binding footprints.

Im9-PaGam proved to be the exception, as although it was successfully overproduced and purified from the *E. coli* K-12 MG1655 with the expected protein MW, it was not able to produce a mobility shift with the 500bp linear DNA as seen in **Figures 40-46** for the other MuGam homologue proteins. This is likely due to the significantly different MW (25.07kDa monomeric MW compared to Im9-HiGam at 30.64kDa) and difference in sequence alignment similarity when compared to the other MuGam homologues (PaGam protein is unable to produce a sequence alignment with MuGam) though it is identified as a nuclease inhibitor protein from phage. Im9-PaGam therefore is unable to bind to linear DNA whether due to improper folding, difference in 3-dimensional (3D) structure or due to the protein amino acid sequence being truncated.

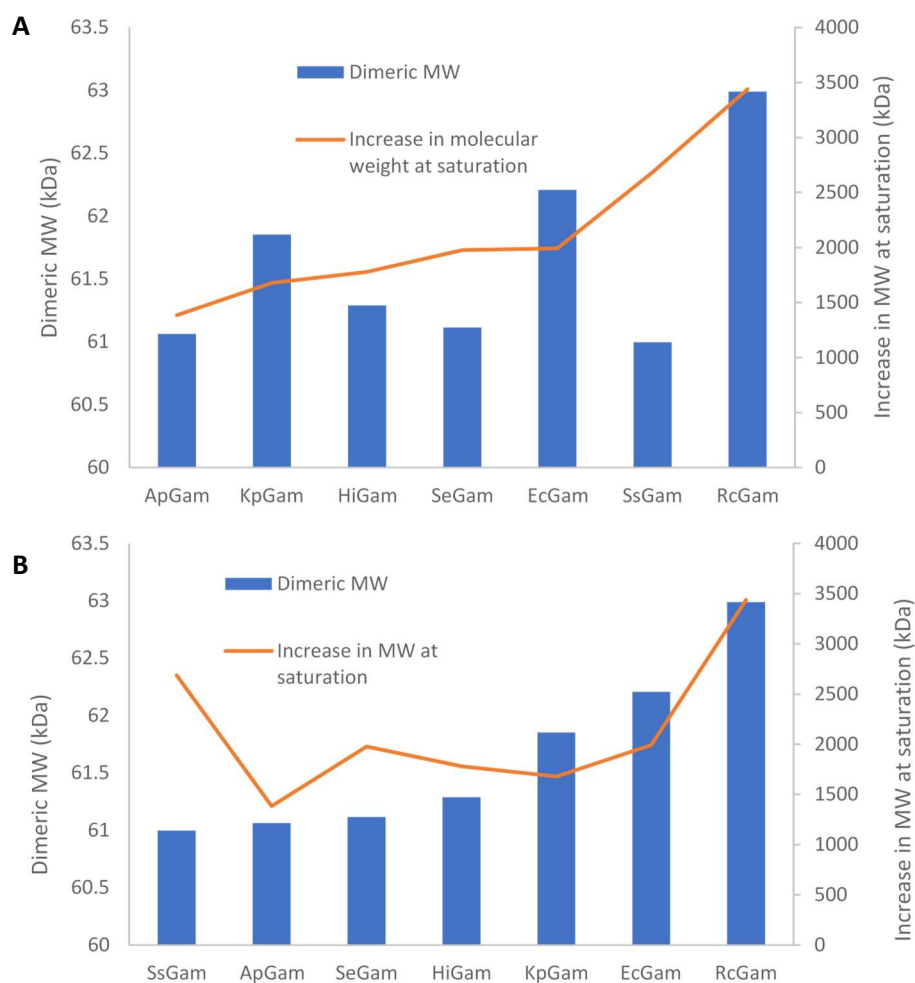


**Figure 47.** The result of electrophoresis of 0.3 $\mu$ M (6.67 $\mu$ g/mL) 500bp linear DNA in the presence and absence of Im9-PaGam. The addition of Im9-PaGam even at extreme excess to the DNA was unable to produce an electrophoretic mobility shift in the DNA.

### 6.2.2 Relationship between MW and point of saturation

The relationship between dimeric protein MW and mobility shift at saturation (increase in MW measured on the gel) are not necessarily related. The 3D structure of the protein should hold a greater correlation with the number of bound units to the 500bp linear DNA. If these proteins share similar 3D structures, then with similar binding footprints heavier dimers will increase the total MW at saturation. However, it is reasonable to assume that a protein with fewer residues would be slightly smaller and therefore have a smaller binding footprint, which would result in more dimeric units bound at saturation.

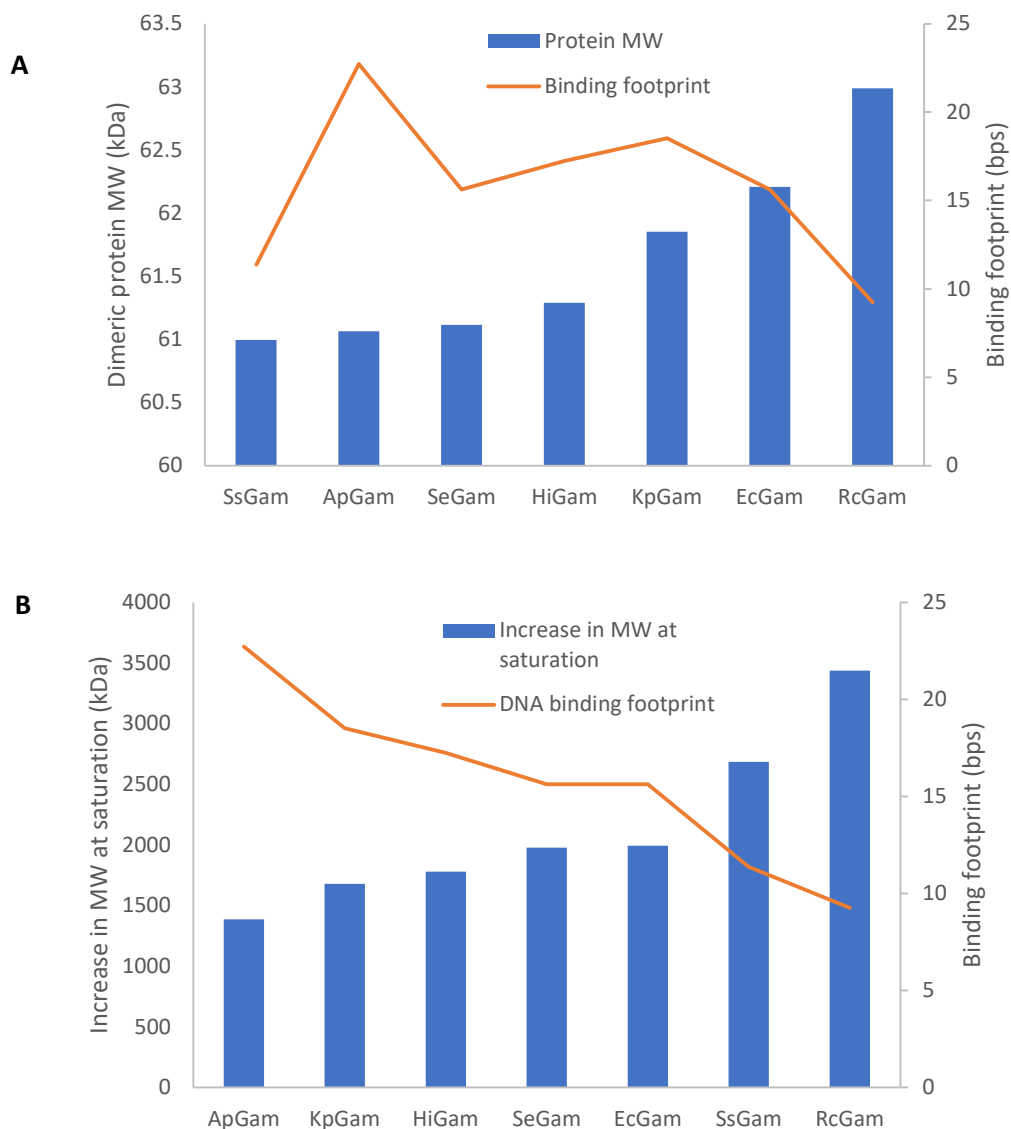
## Characterisation of DNA Binding Properties for Gam Homologues



**Figure 48. A:** The relationship between dimeric MW and the increase in MW of the 500bp linear DNA at saturation when Gam homologue proteins are ordered in ascending order of increase in MW on the gel at saturation. **B:** The same graph as (A) in ascending order of dimeric MW of the Gam homologue protein.

The relationship between dimeric MW of the Gam homologue protein and the total increase of MW of the 500bp linear DNA at saturation is unclear when ordered by the increase in total MW at saturation (**Figure 48A**), though a pattern appears to arise when ordered in ascending order of protein dimeric MW (**Figure 48B**). As discussed there is a potential in conflicting relationships where a heavier dimeric MW will produce a greater increase in MW for the 500bp linear DNA assuming that the binding footprints are the same, but that a lighter dimeric MW results in a shorter binding footprint due to decreased size of 3D structure, therefore allowing more units to bind to increase shift in the mobility. As dimeric Gam homologue MW increased the total increase in MW of 500bp linear DNA creates a U-shaped curve relationship where the lightest protein will produce a similar shift in mobility to the heaviest and proteins with dimeric MW between the two extremes show a decreased shift in mobility of the linear DNA (**Figure 48B**). The relationship between MW and estimated binding footprint is not supported by the data (**Figure 49A**).

## Characterisation of DNA Binding Properties for Gam Homologues



**Figure 49. A:** Compares the MW of each Im9-MuGam homologue (blue) to the estimated binding footprint for each protein (orange). **B:** Shows the total increase in MW at saturation (blue) compared to the estimated DNA binding footprint for each of the MuGam homologue proteins. The length of the DNA binding footprint is directly inversely correlated to the total increase in MW at saturation.

**Figure 49** identifies the direct relationship between the estimated binding footprint and the total increase in MW of the 500bp linear DNA, this relationship potentially contradicts the idea that the MW of the protein is related to the binding footprint, the evidence presented in **Figure 49A** shows the same relationship that was seen in **Figure 48B**. MW and DNA binding footprint do not appear to be directly related. The 3-dimensional structures of the MuGam homologue proteins may be slightly different or there may be ionic or steric differences in amino acids in key areas which result in the difference in the binding footprint due to the ability for the proteins to pack closely together.

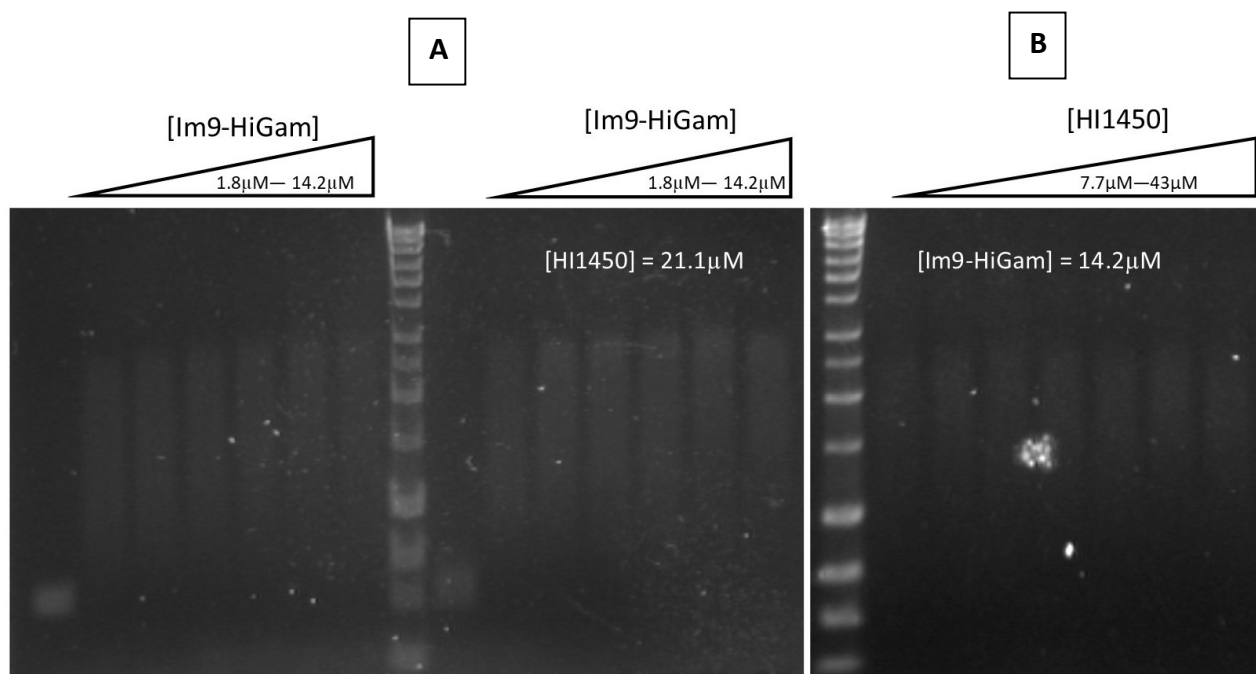
### 6.2.3 Competitive inhibition analysis using HI1450 DNA mimic

HI1450 is a protein known to act as a double stranded DNA (dsDNA) mimic which originates from the *Haemophilus influenzae* Rd KW20 genome (Parsons, Liu and Orban, 2009), the same bacteria which has been shown to produce the HiGam linear dsDNA binding protein (**Chapter 3- HiGam**

**protein production in *Haemophilus influenzae* Rd KW20**). To investigate whether these two proteins may interact the two proteins were mixed together and incubated prior to the addition of linear 500bp DNA for EMSA analysis to see if the addition of the DNA mimic protein acts as a competitive inhibitor to Im9-HiGam binding to the linear DNA.

The Im9-HiGam was able to bind to the linear 500bp DNA and result in a shift in the mobility when loaded onto the 1.5% (w/v) agarose-TAE gel using a loading buffer with no SDS (**Figure 50**). However, there does not appear to be any evidence that the presence of HI1450 reduced the shift in mobility and therefore that HI1450 and Im9-HiGam interact. This is when the two proteins are incubated together at 20°C in the appropriate volume of sodium phosphate buffer for 30 minutes prior to the addition of the linear DNA.

If Im9-HiGam and HI1450 were able to bind together then we would expect to see reduced shift in the DNA in the presence of the HI1450 when in excess of both the DNA and Im9-HiGam. The results do not show this however (**Figure 50**) as the extent of shift in mobility of DNA does not appear to change, either with constant or increasing DNA mimic protein concentration.



**Figure 50.** 1.5% (w/v) agarose + TAE electrophoresis gel images showing the effects of competitive inhibition by HI1450 on Im9-HiGam binding to linear DNA. **A:** EMSA analysed by 1.5% agarose gel electrophoresis showing the shift in mobility of 500bp linear DNA with increasing Im9-HiGam concentration in the presence and absence of a constant concentration of HI1450. **B:** EMSA with constant concentration of Im9-HiGam and increasing concentration of HI1450. There is no visual or measurable reduction in the shift in the mobility of the linear DNA when HI1450 is present with Im9-HiGam, even when HI1450 is greatly in excess.

Where HI1450 is in significant excess to both the Im9-HiGam and the 500bp linear DNA there is no evidence of any interaction between HI1450 and Im9-HiGam. There was no reduction in the shift in mobility in the presence of the HI1450 DNA mimic when compared to its absence. Either the HI1450 and Im9-HiGam protein do not interact or the interaction is too weak to endure the addition of linear DNA. This would then suggest that the binding affinity for Im9-HiGam to linear dsDNA is significantly greater than the affinity for HI1450 DNA mimic protein. In either instance this result suggests that if both proteins are produced in the *H. influenzae* Rd KW20 cell then the HI1450 protein will not inhibit HiGam from binding linear DNA.

### 6.3 Discussion

Seven separate MuGam homologues have now been identified as being capable of binding to linear dsDNA *in vitro* when overproduced and purified from *E. coli* K-12 MG1655. Each showed significant shift in the mobility of the DNA corresponding to the increase in MW as a result of the protein-DNA interaction. If the presence of MuGam in the host cell is increasing the rate of precise repair of DSB in host DNA as the evidence provided in Bhattacharyya *et. al* (2019) suggests, this supports the idea that the gene is conserved because it provides a benefit to the host organism as in each case here the protein is shown to bind to linear DNA as expected which can lead to DNA DSB repair by NHEJ.

Im9-PaGam was not shown to produce any shift in the mobility of the linear DNA when tested under the same conditions as the others. This is likely due to the significant difference in the primary amino acid (aa) sequence of the protein when compared with MuGam and the other MuGam homologues. The aa sequence of Im9-PaGam is considerably shorter (225 aa) than that of Im9-HiGam (269aa) and does not produce a sequence alignment when using the protein multiple sequence alignment tool BLASTp. The protein is however identified as a host-nuclease inhibitor protein in the same family as MuGam. The protein is clearly significantly different from MuGam and originates from Pseudomonas phage JBD88a which is not well researched.

It is difficult to speculate on reasons for the differences in the appearance of the shift between different MuGam homologues, including total difference in MW and the pattern of shift at increasing MuGam homologue at standardised concentrations, for example in **Figure 43** the increase in MW of the DNA appears to increase in a linear progression upon increasing concentration of Im9-SsGam as opposed to the more curved relationship between MuGam homologue and DNA shift seen in **Figure 44** by Im9-SeGam. As discussed in **6.2.3** the total shift of mobility is of DNA at saturation with the MuGam homologues may be influenced by protein MW both as a factor of mass and size independently. **Figure 48** demonstrates that total mobility shift at protein binding saturation varies based on the MW of the protein where the lowest MW (SsGam) has very similar total shift to the highest MW (RcGam) with a discernible U-shaped curve relationship forming with the intermediate MW proteins. This supports the idea that a lower MW protein may be smaller in size than a larger MW protein (assuming a similar 3-dimensional structure) and that that allows an increase in total units of bound protein and a smaller binding footprint.

On the 500bp linear DNA 29 Im9-HiGam dimers were estimated to be bound which produced the observed shift in mobility at saturation point (**Figure 40**). In unpublished research it was found that dimeric Im9-HiGam produced a binding footprint of 17-18bp, this is consistent with what was observed in **Figure 40** (17 bp) and is supported by the theoretical 3-D structure for Im9-HiGam based on the Ku70-80 analogue (d'Adda di Fagagna *et al.*, 2003).

Each MuGam homologue protein is shown here to bind at multiple points along the linear DNA, the protein is unable to bind circular DNA (plasmid or supercoiled) which leads to the idea that the protein must bind to the end of the linear DNA and slide along to allow a new protein unit to bind to the end eventually completely saturating the DNA fragment. Whether this occurs naturally when the Mu Phage injects its genomic DNA into bacterial host has not been determined, the model mechanism indicates that a single MuGam dimer will bind to each end to inhibit host nuclease breakdown. These dimers may be held on the ends by recruitment of other proteins, though these results may suggest that MuGam instead saturates the linear genomic Mu phage transposable element to potentially protect the DNA entirely by wrapping it fully with protein. The saturation effect may however be caused by the enormous excess of protein used in the experiments used here, the high affinity of the protein may force already bound dimers to be pushed into the linear

## Characterisation of DNA Binding Properties for Gam Homologues

DNA fragment where under lower concentration conditions the dimer would naturally remain near the DNA end. Determining this would be difficult using EMSAs as they are limited by the observed mobility shift of sufficient DNA concentration to be visible under UV on agarose gel. Other methods such as atomic force microscopy (AFM) could be used to visualise the protein-DNA interaction at a single molecule level (Ruggeri *et al.*, 2019). It may therefore be possible to isolate a single linear DNA strand and single MuGam homologue protein unit to visualise the location of protein binding. If the protein preferentially stays at the end or if it moves freely along the DNA once bound. This can also be done in conjunction with other Mu proteins to determine if MuGam remains at the end in the presence of other Mu proteins known to bind to the linear DNA for transposition (such as MuA and MuB).

There is no observable difference in the mobility of the DNA when the Im9-HiGam is first incubated with HI1450 (**Figure 50**). This would suggest that the DNA mimic does not interact with the Im9-HiGam and therefore cannot competitively inhibit the binding to linear DNA. This could be due to several reasons including the short length of the HI1450 protein (corresponding to approximately 10-15bps) or that the HI1450 DNA mimic protein has not folded correctly and therefore cannot mimic DNA.



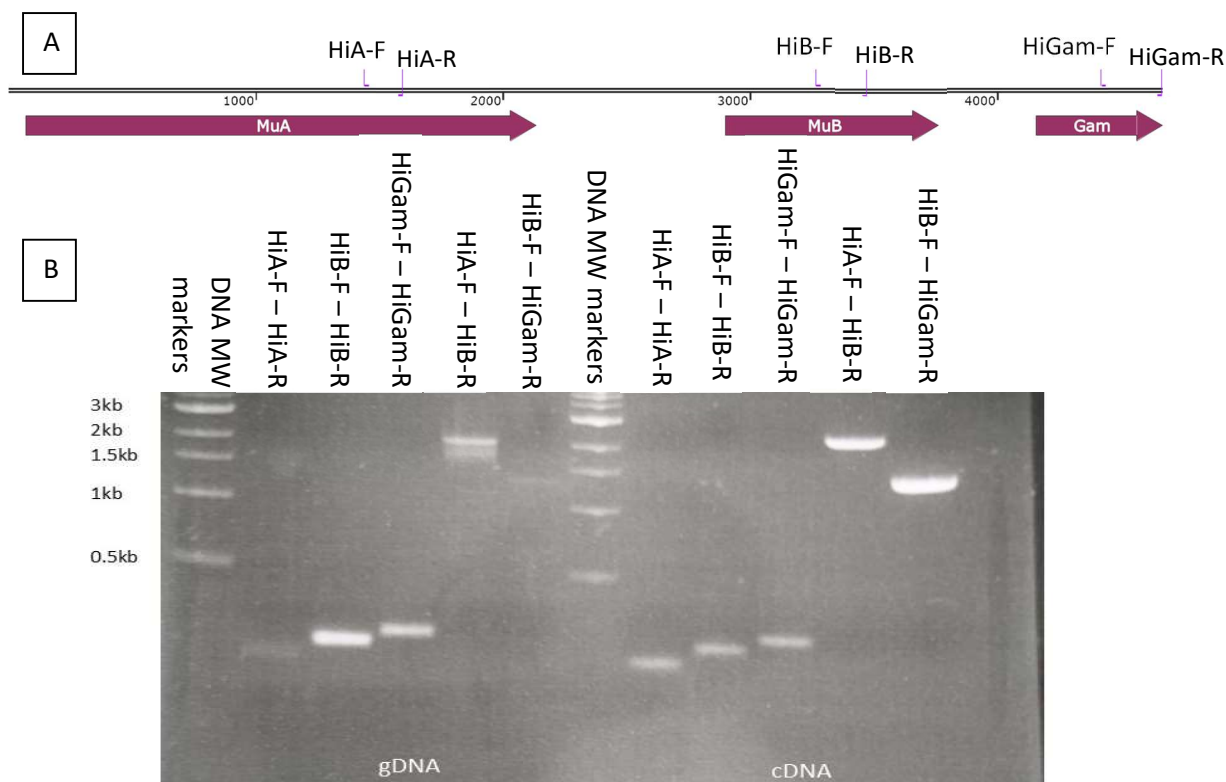
# Results

## Chapter 7 – HiGam protein production in *Haemophilus influenzae* Rd KW20

## 7.1. Introduction

### 7.1.1 Mu phage prophage on *Haemophilus influenzae* Rd KW20 (*H. influenzae* Rd KW20) genome

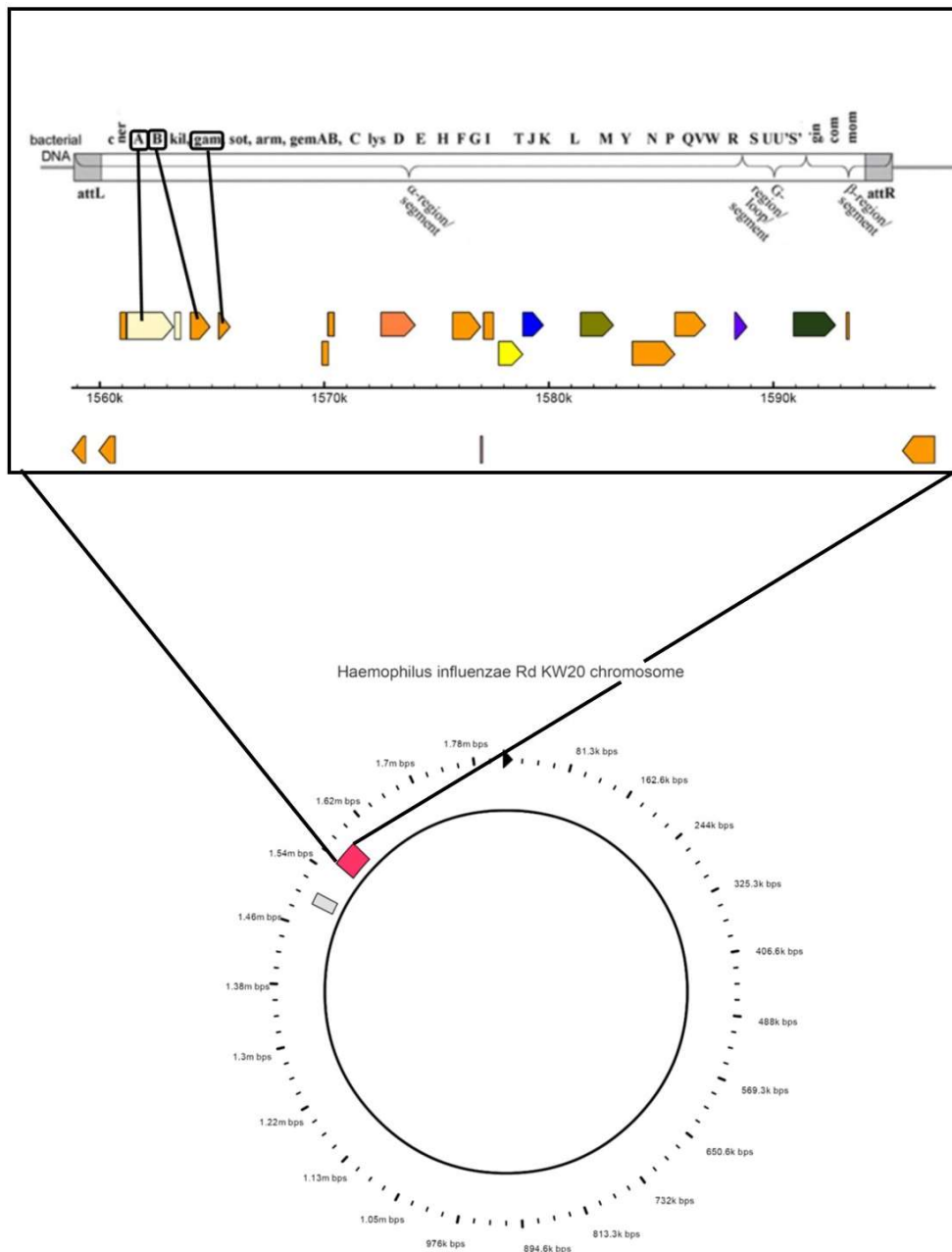
MuGam inhibits host-nuclease activity by blocking the exposed ends of the linear DNA (Williams and Radding, 1981) MuA and MuB will then bind to these ends to perform the transposition (Mizuuchi and Mizuuchi, 1989). Previous research has shown that the MuGam protein is translated from a polycistronic mRNA that contains two other genes to produce proteins MuA and MuB (demonstrated in **Fig 51**) which are involved in transposition of the prophage DNA to the host genome.



**Figure 51.** Evidence that the MuA, MuB and MuGam genes are transcribed onto a single polycistronic mRNA from purified samples of DNA amplified by PCR using selected primers to produce different lengths of DNA corresponding to different regions found on the Mu prophage after transcription (cDNA). Repeat of previous experiment performed in unpublished research. A: demonstrates the PCR primers designed to anneal to specific regions at each gene of interest found on the prophage region of *H. influenzae* Rd KW20. B: 1% (w/v) agarose-TAE gel electrophoresis of PCR products of genomic DNA (gDNA) and complementary DNA (cDNA) produced from reverse transcription of mRNA produced by the transcription in *H. influenzae* Rd KW20 cells when using various combinations of PCR primers (e.g HiA-F). The results of this gel prove that the mRNA is polycistronic containing genes for HiA, HiB and HiGam. If there were separate mRNA sequences produced for each gene then we would expect the PCR to fail to amplify HiA-F – HiB-R and HiB-F – HiGam-R samples as primers would fail to anneal cooperatively.

The genes for HiA, HiB and HiGam (Mu phage genes found in *H. influenzae* Rd KW20 prophage are named with the Hi- prefix) are found in sequences from the Mu prophage region on the

*Haemophilus influenzae* Rd KW20 (*H. influenzae* Rd KW20) genome. The prophage region here is complete as seen in **Figure 52**. The HiA, HiB and HiGam genes are transcribed in the same polycistronic mRNA so it is predicted that each transcript should be translated to produce protein. HiGam is the protein of interest, predicted to provide a survival benefit to the cell, while HiA and HiB are not predicted to provide survival benefit. We have seen evidence that MuGam homologues such as HiGam are consistently conserved even where the prophage is not fully intact. We theorise that the HiGam protein is produced while HiA and HiB are not, which could hold implications about the use of MuGam homologues in the cell.



**Figure 52.** (Bottom) Schematic view of the *H. influenzae* Rd KW20 chromosome from Phage search tool (PHAST) search for *H. influenzae* Rd KW20 (Accession no.: NC\_000907) with highlighted (pink) and expanded view of the prophage region originating from Mu phage. (Top) Expanded view of the complete Mu prophage within *H. influenzae* Rd KW20, HiA, HiB and HiGam genes are identified on the gene map in sequence.

### 7.1.2 Liquid chromatography-Mass spectrometry (LC-MS) to determine presence of HiGam protein in *H. influenzae* Rd KW20 cells

It is important to determine whether the HiGam protein is produced from the prophage gene under normal conditions in *H. influenzae* Rd KW20 cells to infer potential function and a survival benefit for the host bacterium.

LC-MS is a technique used to identify compounds on a massive scale with the sensitivity to search for specific compounds, utilising the separation of liquid chromatography and the compound analysis by mass spectrometry allows the identification of thousands of unique compounds in a sample (Colgrave *et al.*, 2019). In this case we analyse the soluble fraction of the total cell lysis of a cell culture to identify fragments of polypeptide compounds which can be sequence aligned to existing protein sequences to form a map of all the proteins that are produced in the cell (to the detection limit).

In unpublished work, the HiGam protein was identified in *H. influenzae* Rd KW20 cytoplasmic extract by mass spectrometry (MS). The MS data was not able to produce any evidence of HiA or HiB protein being present in the cell extract. This could suggest either that the concentration of the proteins within the cell were too low to be detected or that the proteins are not present in the cell. Their absence could suggest incomplete folding resulting in targeted enzymatic proteolysis, instability resulting in the breakdown of the protein product produced from the polycistronic mRNA. Advancements in equipment and technique have made it possible to identify protein fragments with significantly lower abundance. The soluble fraction of the cell lysate was analysed again using the same method with an updated methodology to try to identify the HiA and HiB proteins and determine the relative abundances.

## 7.2. Results

### 7.2.1 Identification of key polypeptide fragments using LC-MS

LC-MS output data were peak picked, combined and searched against the *H. influenzae* subset of the SwissProt database aligned with the expected sequence for target proteins HiA, HiB and HiGam. The data presented are from database searching using PEAKSX Studio. The LC-MS analysis was able to identify and match unique peptide fragment sequences for 1643 unique proteins with a  $p < 0.05$  confidence and a minimum of two peptides identifications per protein, and the empirical false discovery rate determined against a reversed database search was 2.1%. This list included peptide fragment sequences found exclusively in HiGam and HiB listed in **Table 12**. No corresponding fragments were found for HiA, the relative quantities of protein fragment are either too low to be identified or they are not detected due to similarities to other sequences and limitations in the analysis method.

HiGam-derived peptides	HiB-derived peptides	HiA-derived peptides
EIGDLSR	QHLSDSQITQAQLAR	<b>YESLALIGTSHK</b>
IGATSEHYAPK	EAGVNAGALSAYLNDNYK	<b>AFSHGGLGDYVDK</b>
GAEAVMEFLQR	GNIADVEAK	<b>TTPEQTAVALQK</b>
EALLNEPEVAK	EFVEAPAFIETATSR	
	VMQSITETGGGLR	

**Table 12.** Peptide sequences identified using LC-MS. The amino acid sequences identified for HiGam and HiB from the total protein content of *H. influenzae* Rd KW20 using LC-MS. HiGam identified fragments were matched with  $-10\log(P) = 65.31$  (equivalent to  $p < 0.01$  confidence) and HiB with  $-10\log(P) = 54.32$  (equivalent to  $p < 0.01$  confidence). The most theoretically MS amenable fragments to potentially identify HiA in the total cell lysate are also provided in the table (in bold). None of these fragments were identified in the sample or any others that matched corresponding sequences found in the HiA protein sequence.

### 7.3. Discussion

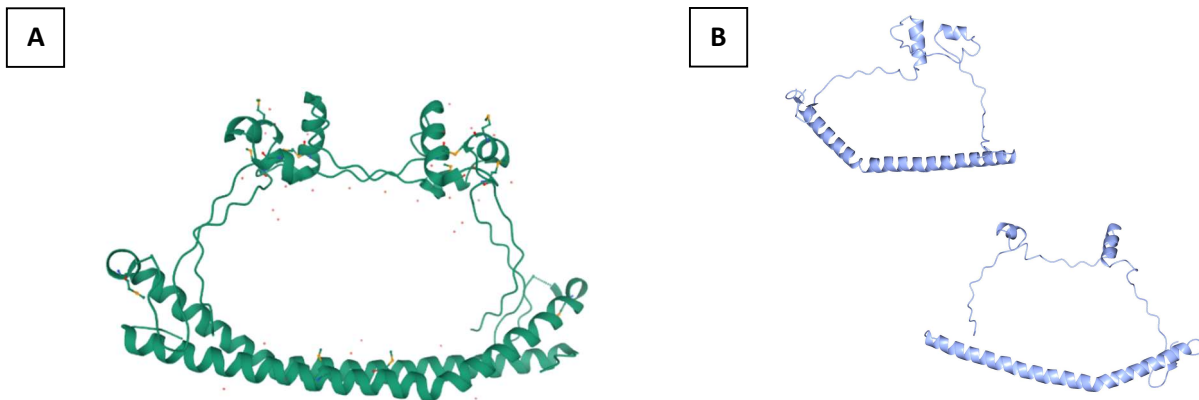
The LC-MS method was able to identify completely unique protein fragments for HiGam and HiB, therefore both proteins are transcribed and translated in *Haemophilus influenzae* Rd KW20. The evidence suggests that HiA is not produced within the cells. No other Mu phage protein present in the Mu prophage of *H. influenzae* Rd KW20 was shown to be present in the analysed cell lysate. This may suggest that the HiGam protein is produced due to a survival benefit provided to the host bacterial cell. The evidence here suggests that the HiA protein is not produced though it is possible that the detection limit was not sufficient to detect any fragments unique to HiA either due to concentration being too low or lack of separation between similar protein fragments.

In order to further investigate whether HiA is produced in the cell the detection of protein fragments specific to HiA can be improved by optimising separation and detection at the expected molecular weight region. By identifying the detection point of isotopically labelled fragments expected to be seen from HiA, the LC-MS method can analyse these specific regions with greater accuracy to identify with greater confidence whether the protein fragments are present within the cell lysate (Kirkpatrick, Gerber and Gygi, 2005; Colgrave *et al.*, 2019).

## 8. Discussion

### 8.1. The 3-Dimensional structure of MuGam

The 3-Dimensional (3D) structure of MuGam has not been solved by any method to this point, neither has any other MuGam homologue with significant sequence homology. The only structure that currently exists is for a putative host-nuclease inhibitor protein Gam from *Desulfovibrio vulgaris* (2PU2, Bonano et. al 2007). This putative host-nuclease protein (DvGam) is homologous to MuGam though the sequence homology is very low, only 26% sequence identity and 50% positive identities suggests that the similarity between these two proteins is limited. It is certainly possible for the two proteins to have structural homology, though predicting that the DvGam model could be appropriate for MuGam and its closer homologues is questionable. It has previously been asserted that the solved crystal structure for DvGam could be used as a model for MuGam (Bhattacharyya et al., 2018). The issue with using the DvGam crystal structure as a model for MuGam primarily stems from the way that the structure was solved, the 2PU2 structure consists of crystallised DvGam monomers and the dimeric structure has been inferred through modelling the two solved monomeric structures in an inverted orientation.



**Figure 53.** **A:** 2PU2 solved crystal structure for putative host-nuclease inhibitor protein Gam from *Desulfovibrio vulgaris* (Bonano et. al 2007). The crystal structure presents two monomers of the DvGam protein that are not interacting. **B:** The biological assembly of the DvGam dimer structure proposed based on the solved monomeric structure.

The dimeric structure (**Fig. 53 A**) does not take into account the possibility for changes in conformation as a result of dimerization or difference in the folding of the protein when in the presence of the other monomer. If the two monomers are translated together at the same time and fold initially into their dimeric conformation, then the structure of the monomeric protein may bear little resemblance to the dimeric protein.

Based on what we know about the function of MuGam and the MuGam homologue HiGam from *H. influenzae* Rd KW20, from previous unpublished research as well as the research presented in Bhattacharyya et al., 2018 and d'Adda di Fagagna et al., 2003, MuGam binds exclusively to linear double stranded DNA (dsDNA) in a manner similar to human Ku70/80 protein. The proposed structure for dimeric DvGam suggests that the DNA binding domain is approximately 50Å wide, more than large enough to fit two linear dsDNA strands (approx. 23Å each) if the DNA could be packed closely enough. The structure for human Ku70/80 heterodimer has a DNA binding domain of only approx. 27Å which would allow more peptide-DNA interactions that result in the significant binding affinity of the protein to linear dsDNA. The high binding affinity of MuGam and HiGam to linear dsDNA would suggest that the interactions between the protein and the DNA are strong

despite being non-specific, logically then we would expect the binding domain to allow for a strong interaction by forming multiple tight non-covalent interactions which the structure for DvGam would not provide.

### **8.1.1 The structure of MuGam and the relationship between MuGam, DvGam and Eukaryotic Ku**

MuGam was identified as a potential orthologue to human Ku70/80 heterodimer (d'Adda di Fagagna *et al.*, 2003) based on sequence homology on the binding domain of the eukaryotic heterodimer. This would suggest that there could be structural homology in this region and would not support the model presented for DvGam. The sequence homology between the human Ku70/Ku80 proteins and MuGam is not strong "17% identity and 27% similarity with human Ku80, and 13% identity and 9% similarity with human Ku70" (d'Adda di Fagagna *et al.*, 2003) and the secondary structure prediction does not support the idea that MuGam or HiGam could take a conformation that is homologous to the Ku70/80 heterodimer. The only structural homology that could be predicted using one-to-one threading to force conformation of HiGam into the Ku secondary and tertiary structures was in the non-DNA binding domain.

All the predictions that were produced suggested that HiGam and MuGam would take the secondary structure conformations presented by the 2PU2 structure for DvGam. These prediction algorithms rely heavily on existing structures with sequence homology to the query sequence, for which DvGam is the closest match. This skews the data towards the structure for DvGam, though even when looking at unrelated structures the secondary structure elements that are predicted are consistent, in particular with the long N-terminal  $\alpha$ -helix. This heavily suggests that MuGam would also take the conformation shown in the solved crystal structure for monomeric DvGam.

### **8.1.2 How secondary structure translates to 3-Dimensional structure**

Secondary structure prediction can only take place on the monomeric sequence for MuGam, there is no way to reasonably predict the differences that could occur between the conformation of monomeric and dimeric protein or any potential reasons for those differences. During dimerization it is unlikely that secondary structural conformations will alter significantly as they are formed prior to tertiary structure folding this supports the DvGam structure, but it also does not confirm it.

The only way to definitively prove the structure of MuGam is to solve the 3D structure either through the formation of dimeric crystals of MuGam or a significantly similar MuGam homologue. One of the MuGam homologue proteins that have been purified in this study may be capable of forming crystals capable of being used for X-ray crystallography to solve the 3D structure though they have not yet been screened. The other reasonable option to solve the 3D structure of a MuGam homologue dimer is to use the NMR structural analysis method. The protein is typically considered to be too large to be solved by NMR (MuGam dimer = 38.07kDa, typical upper limit for NMR protein structure solving = 20kDa) though there is evidence that NMR can be used to solve 3d structures for proteins with a much greater molecular weight of 50kDa and beyond (Frueh *et al.*, 2013). If dimeric crystal structures cannot form from any MuGam homologues with significantly similar sequence homology then NMR could be a viable option despite the increased complexity and potentially lower accuracy/resolution when compared to X-ray crystallography.

## **8.2. The relationship between MuGam and Ligase for DNA repair**

Human Ku70/80 recruits ATP-dependent DNA ligase (DNA ligase IV) to repair double strand breaks (DSBs) (Featherstone and Jackson, 1999; Lieber, 2010; Pannunzio, Watanabe and Lieber, 2018),



MuGam has been shown to increase the rate of DSB repair when overproduced with NAD<sup>+</sup>-dependent DNA ligase (LigA) in *Escherichia coli* K-12 MG1655 even when RecA was knocked out to prevent DNA repair by homologous recombination (Bhattacharyya *et al.*, 2018).

### **8.2.1 Proposed mechanism for DNA repair by NHEJ in bacteria**

Bhattacharyya *et al.* (2018) found that by using an expression vector for MuGam and NAD<sup>+</sup>-dependent ligase (LigA) instances of DNA repair increased in the presence of induced DSBs by phleomycin. They also found that by knocking out RecA, a protein essential for homologous recombination (HR), that the presence of Gam still improved DNA repair. They observed both exact accurate repairs, where the resulting sequence is the same as the original, and approximate repair where the sequence differs from the original sequence by 1-13 nucleotides.

### **8.2.2 The presence of ligase in species with conserved MuGam homologue genes**

The DNA repair observed in the Bhattacharyya *et al.* (2018) study focussed on the interaction between MuGam and LigA. LigA is highly conserved and ubiquitous among all bacterial species and so it is possible that bacteria could benefit from the presence of a MuGam homologue gene. It is also common for bacteria to possess the gene for ATP-dependent ligase (LigD). If MuGam homologues are able to recruit these LigD proteins to repair DNA then it suggests that it could be even more effective when both proteins are produced in the cell to increase DSB repair rate.

### **8.2.3 How *Avibacterium paragallinarum* JF2411 may be different**

It is rare for a bacterial species to lack a complete LigA gene to produce a functioning LigA protein, though the presence of the MuGam homologue in the genome of *A. paragallinarum* JF2411 could suggest that it is still providing survival benefit despite the lack of LigA. This could either mean that the MuGam homologue could improve survivability through another mechanism or that the protein can recruit the LigD protein as well. The sequence homology between the MuGam homologue protein from *A. paragallinarum* JF2411 and MuGam itself is strong (53% identities and 66% similar) which would suggest a strong structural homology, though the difference may be significant enough to result in the recruitment of a different Ligase for DNA repair.

### **8.3. The production of HiGam, HiA and HiB in *Haemophilus influenzae* Rd KW20**

The presence of the functional HiGam (MuGam homologue) gene in the Mu prophage region of *Haemophilus influenzae* Rd KW20 genome does not implicate that the gene provides any survival benefit to the host cell. The presence of the protein in the soluble fraction of the cell lysate proves that the protein is produced and therefore may perform a function in the cell that provides a survival benefit such as DNA DSB repair through the NHEJ pathway.

#### **8.3.1 The presence of HiGam within the cell under natural conditions**

Evidence that HiGam is produced in the cell was previously shown in previous research conducted by Rhian Gore (unpublished data, 2017). The equipment and method available at that time were not adequate to detect the other key proteins from the Mu prophage (HiA and HiB). Here we were able to observe evidence that HiB was produced in the cell. MuGam and MuB are known to interact during DNA transposition (Lavoie and Chaconas, 1996), the HiB protein within *H. influenzae* Rd KW20 may therefore interact with the HiGam protein within the cell, this could result in a loss in potential DNA repair function or the HiB may play some kind of role in the mechanism. To test whether HiB interferes with the binding of HiGam to linear DNA, an EMSA competition experiment could be used as was done for HI1450 and Im9-HiGam. Since HiGam and HiB should work together to bind to the

end of linear DNA it would be interesting to see if the HiB protein locks the HiGam protein to the end of the linear DNA therefore preventing it from moving along and saturating the linear DNA. If HiGam was held near the end of the linear DNA and does indeed recruit DNA ligase to repair DNA by NHEJ then the presence of HiB may improve the DNA repair capabilities associated with the potential HiGam DNA repair pathway.

Other methods to investigate protein-protein interactions could confirm the potential interaction between HiGam and HiB. In addition, this complex may be stable enough to generate protein crystals for X-ray crystallography. One method for investigating protein-protein interactions is the use of Microscale Thermophoresis (MST) (Jerabek-Willemsen *et al.*, 2011) This method is capable of measuring the motion of molecules in temperature fields and detecting changes in molecular charge, size and solvation to determine if molecules are interacting. Equally if HiGam and HI1450 do share a weak interaction then this may be observed using this method and may be another avenue to create stable complexes for X-ray crystallography.

### **8.3.2 The absence of HiA in *H. influenzae* Rd KW20 LC-MS results**

The liquid chromatography – mass spectrometry (LC-MS) results did not identify any non-specific or unique fragments that correspond to the HiA protein sequence. The absence of evidence of the HiA protein in this experiment suggests that the protein is either not produced, it is degrading within the cell or that the level of production is too low for detection using this method. The use of a more focussed search for specific fragments most likely to be MS amenable particularly with the use of labelled isotopic fragments [literature reference needed] used in conjunction with the analysis protocol to identify the potential HiA fragments in the trypsinised cell lysate of the *H. influenzae* Rd KW20 cells. The host cell may not benefit from the function that the MuA protein provides and may target it's destruction to prevent transposition events involving the Mu prophage region.

### **8.4. The success of MuGam homologue and DNA mimic protein purification**

Im9-NmGam was not successfully overproduced upon induction with IPTG under the same conditions as the other MuGam homologues. There did not appear to be any bands in the expected molecular weight range that appeared in the post-induction samples compared to the pre-induction samples. The sequencing data showed that the gene was not present on the recombinant plasmid of the transformed *E. coli* K-12 MG1655 cells. The other MuGam homologue genes were successfully overproduced in the cells. The MuGam homologues that were successfully overproduced and purified from the soluble fraction of the cell lysate so the proteins fold correctly in the bacterial host. The variety of origin and the differences between the protein primary sequences of the MuGam homologues investigated suggests that the protein may be capable of being produced in a wide variety of bacterial host. The consistent conservation of MuGam homologue genes may be a result of protein production as supported by this evidence.

The oligomeric states of the MuGam homologue proteins were not determined in this work; however, Im9-HiGam was shown to be homo-dimeric by size-exclusion chromatography (SEC) and sedimentation equilibrium analytical ultracentrifugation (seAUC) (R. Gore, F. Engelhardt and C.G. Baumann; unpublished data). To confirm the oligomeric state of the MuGam homologue proteins the samples would first be purified by SEC (Hong, Koza and Bouvier, 2012) then analysed by seAUC (Svedberg and Fåhræus, 1926; Scott and Schuck, 2005; Cole *et al.*, 2008). Using this method, the molecular weight of the protein complexes in solution can be determined from the sedimentation coefficient to establish the oligomeric state of the proteins.

Evidence of a dimeric state for one of the MuGam homologue proteins cloned and purified in this study was seen in the SDS-PAGE analysis of the Ni<sup>2+</sup> affinity chromatography fractions (section 5.2.5). In this section, the non-target protein bands identified by SDS-PAGE gel analysis for Im9-RcGam (**Fig. 27**) were identified as the correct MW to correspond to the N-terminal His6-tagged Im9 protein solubility tag and the C-terminal RcGam protein fragments from proteolysis. Since the His6-tag (required to bind tightly to the Ni<sup>2+</sup> affinity column) is present on the N-terminal Im9 protein fragment, the only way the RcGam protein fragment without a His6-tag could be present in the fractions eluted from the Ni<sup>2+</sup> affinity column is if the Im9-RcGam and RcGam protein fragment were forming a heterodimer of recombinant protein. This could be further investigated using seAUC as the Im9-RcGam homodimer would have a molecular weight which was ~9.5 kDa greater than the Im9-RcGam-RcGam fragment heterodimer (Scott and Schuck, 2005). **8.5. Analysis of MuGam homologue function using Electrophoretic mobility shift assays (EMSA)**

Every MuGam homologue that was overproduced and purified from *E. coli* K-12 MG1655 was shown to bind to linear DNA to cause a shift in the mobility through the 1.5% (w/v) agarose-TAE gel during electrophoresis. The differences in the primary sequences of the proteins did not negatively affect the expected protein function, and the EMSA evidence presented suggests that each protein is functional. The greater the concentration of the MuGam homologue protein the slower the DNA migrated through the gel until reaching a threshold to saturate the DNA with protein.

One way the DNA binding properties of the MuGam homologues could be investigated more accurately would be by reducing the length of the linear target DNA from 500bp to closer to the predicted binding footprint (estimated to be between 9-23bp, **Fig. 40-46**) of the MuGam homologue proteins. This would reduce the number of dimeric units bound to each linear DNA fragment and therefore reduce variation in measured molecular weight using the EMSA method. Detection of the short DNA fragments at the lower DNA concentrations required here is more challenging when using DNA staining with intercalating fluorescent dyes after gel electrophoresis. To circumvent this problem, short synthetically produced DNA strands containing a fluorescent or radio-isotopic label can be utilised to improve the detection of the shorter linear DNA molecules within the gel (Hellman and Fried, 2007).

Fluorescence anisotropy can be used to analyse the binding affinity of the MuGam homologue proteins with linear DNA (Anderson et al., 2008). By covalently coupling an appropriate fluorophore to the target double-stranded DNA oligonucleotide, the anisotropy of the fluorescence emission will increase upon MuGam homologue binding to the DNA. Measuring this change in fluorescence anisotropy as a function of MuGam homologue protein concentration allows determination of the equilibrium dissociation constant ( $K_d$ ) as well as the number of protein dimers bound to the DNA fragment (Arosio et al., 2004). This technique has been used to study the binding of the eukaryotic Ku70/Ku80 complex to linear DNA and would provide an opportunity for comparison with these predicted structural homologues.

### **8.5.1 How molecular weight (MW) relates to binding footprint**

The MW of the protein does not directly correlate with the binding footprint and the structure and residues of the individual protein likely has a greater impact on the binding footprint. The 3-dimensional structures of each MuGam homologue may be subtly different enough to influence the binding footprint at saturation. The potential differences in 3D structure and residue charge and size at key positions most likely do not significantly impact the overall structure of the proteins as they have significant sequence homology and show the same linear DNA binding function.

### 8.5.3 HI1450 as a competitive inhibitor to Im9-HiGam

The competitive binding EMSA results did not present any evidence that HI1450 binds to Im9-HiGam to prevent binding to linear DNA. Either the two proteins do not interact, or the affinity and strength of the interaction is significantly lower than the interaction between Im9HiGam and linear DNA. The result suggests that in the *H. influenzae* Rd KW20 cells the HI1450 protein will not inhibit HiGam function.

### 8.6. The potential for MuGam homologues to function in a DNA repair mechanism

To further investigate whether MuGam homologues universally provide a survival benefit to the host cell different homologues should be tested in different cell strains. Battacharyya et. al 2018 found that overproducing MuGam in *E. coli* K-12 MG1655 cells that were treated with phleomycin to induce DSBs in the genomic DNA. The genome of this strain contains at least 14 different prophage regions from different phage origins based on a genome search through PHAge Search Tool (PHAST). If there are proteins that are produced in the cell from these prophage genes much like HiGam and HiB are shown to be produced, then these proteins may also influence DNA DSB repair either on their own or in conjunction with MuGam inside the cell. In order to better investigate the proposed model that MuGam recruits DNA ligase LigA to DSB sites to induce repair similarly to the mechanism of NHEJ by eukaryotic Ku heterodimer then the survival rate should be tested using a strain with a genome free from other prophage regions. This could be investigated using *E. coli* MDS42 strain [literature reference needed] which has all of the prophage regions deleted to simplify the genome by removing non-essential genes. If MuGam still shows the same survival and DSB repair impact on the cell then we can be more confident in the proposed mechanism.

One other important factor to investigate is the type of DSB inducing agent used in the survival experiments, Battacharyya et. al 2018 used phleomycin to induce DSB at a concentration sufficient to cause 50% mortality without the presence of MuGam to aid repair. In unpublished research by Holly Clynes (2018) there was no significant survival benefit seen in *E. coli* K-12 MG1655 cells when DSBs were induced using ciprofloxacin to introduce a 50% mortality rate. The difference in the result of these two investigations may primarily be due to the difference in the DSB inducing agent used. One key difference is the mechanism by which the agents produce DSBs. Ciprofloxacin inhibits DNA topoisomerase and DNA gyrase proteins to induce DSBs by preventing the proteins from releasing from the DNA as the protein becomes covalently linked with the 5' end of the DNA (LeBel, 1988; Evans-Roberts *et al.*, 2016; Tamayo *et al.*, 2009). This therefore would severely impact the potential repair mechanism provided by MuGam homologues as it would prevent binding to the end of the DNA. Phleomycin uses a free-radical mediated mechanism to introduce the DSBs which leaves the end of the linearised DNA free (Sleigh, 1976). MuGam would then be able to bind to the end and potentially mediate DNA repair.

It would be important to test different DSB inducing agents with different MuGam homologue proteins in different strains of bacteria to show that the proteins are consistently conserved due to their survival benefit to the host cell. To then further investigate the mechanism, interactions between MuGam and LigA should be confirmed during the process of DNA repair. If proven this mechanism could become a new anti-microbial resistance target.

## 9. References

- Akroyd, J. and Symonds, N. (1986) 'Localization of the gam gene of bacteriophage Mu and characterisation of the gene product', *Gene*. Elsevier, 49(2), pp. 273–282. doi: 10.1016/0378-1119(86)90288-X.
- Aravind, L. and Koonin, E. V (2001) 'Prokaryotic homologs of the eukaryotic DNA-end-binding protein Ku, novel domains in the Ku protein and prediction of a prokaryotic double-strand break repair system.', *Genome research*. Cold Spring Harbor Laboratory Press, 11(8), pp. 1365–74. doi: 10.1101/gr.181001.
- Bhattacharyya, S. *et al.* (2018) 'Phage Mu Gam protein promotes NHEJ in concert with Escherichia coli ligase.', *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, 115(50), pp. E11614–E11622. doi: 10.1073/pnas.1816606115.
- Brissett, N. C. and Doherty, A. J. (2009) 'Repairing DNA double-strand breaks by the prokaryotic non-homologous end-joining pathway.', *Biochemical Society transactions*. Portland Press Limited, 37(Pt 3), pp. 539–45. doi: 10.1042/BST0370539.
- Clark, D. P., Pazdernik, N. J. and McGehee, M. R. (2019) 'Mobile DNA', in *Molecular Biology*. Elsevier, pp. 793–829. doi: 10.1016/b978-0-12-813288-3.00025-2.
- Colgrave, M. L. *et al.* (2019) 'Quantitation of seven transmembrane proteins from the DHA biosynthesis pathway in genetically engineered canola by targeted mass spectrometry', *Food and Chemical Toxicology*. Pergamon, 126, pp. 313–321. doi: 10.1016/J.FCT.2019.02.035.
- d'Adda di Fagagna, F. *et al.* (2003) 'The Gam protein of bacteriophage Mu is an orthologue of eukaryotic Ku.', *EMBO reports*. European Molecular Biology Organization, 4(1), pp. 47–52. doi: 10.1038/sj.embor.embor709.
- Drozdetskiy, A. *et al.* (2015) 'JPred4: A protein secondary structure prediction server', *Nucleic Acids Research*. Oxford University Press, 43(W1), pp. W389–W394. doi: 10.1093/nar/gkv332.
- Evans-Roberts, K. M. *et al.* (2016) 'DNA gyrase is the target for the quinolone drug ciprofloxacin in Arabidopsis thaliana', *Journal of Biological Chemistry*. American Society for Biochemistry and Molecular Biology Inc., 291(7), pp. 3136–3144. doi: 10.1074/jbc.M115.689554.
- Fattah, F. *et al.* (2010) 'Ku Regulates the Non-Homologous End Joining Pathway Choice of DNA Double-Strand Break Repair in Human Somatic Cells', *PLoS Genetics*. Edited by C. E. Pearson. Public Library of Science, 6(2), p. e1000855. doi: 10.1371/journal.pgen.1000855.
- Featherstone, C. and Jackson, S. P. (1999) 'Ku, a DNA repair protein with multiple cellular functions?', *Mutation Research/DNA Repair*. Elsevier, 434(1), pp. 3–15. doi: 10.1016/S0921-8777(99)00006-3.
- Ferrières, L. *et al.* (2010) 'Silent mischief: Bacteriophage Mu insertions contaminate products of Escherichia coli random mutagenesis performed using suicidal transposon delivery plasmids mobilized by broad-host-range RP4 conjugative machinery', *Journal of Bacteriology*. American Society for Microbiology (ASM), 192(24), pp. 6418–6427. doi: 10.1128/JB.00621-10.
- Frueh, D. P. *et al.* (2013) 'NMR methods for structural studies of large monomeric and multimeric proteins', *Current Opinion in Structural Biology*. NIH Public Access, pp. 734–739. doi: 10.1016/j.sbi.2013.06.016.
- Haapa-Paananen, S., Rita, H. and Savilahti, H. (2002) 'DNA transposition of bacteriophage Mu. A quantitative analysis of target site selection in vitro', *Journal of Biological Chemistry*. JBC Papers in Press, 277(4), pp. 2843–2851. doi: 10.1074/jbc.M108044200.

## References

- Han, S., Chang, J. S. and Griffor, M. (2009) 'Structure of the adenylation domain of NAD(+)-dependent DNA ligase from *Staphylococcus aureus*.', *Acta crystallographica. Section F, Structural biology and crystallization communications*. International Union of Crystallography, 65(Pt 11), pp. 1078–82. doi: 10.1107/S1744309109036872.
- Harshey, R. M. (2012) 'The Mu story: how a maverick phage moved the field forward', *Mobile DNA*. Springer Nature, 3(1), p. 21. doi: 10.1186/1759-8753-3-21.
- Jerabek-Willemsen, M. *et al.* (2011) 'Molecular interaction studies using microscale thermophoresis', *Assay and Drug Development Technologies*. Mary Ann Liebert, Inc., pp. 342–353. doi: 10.1089/adt.2011.0380.
- Kelley, L. A. *et al.* (2015) 'The Phyre2 web portal for protein modeling, prediction and analysis', *Nature Protocols*. Nature Publishing Group, 10(6), pp. 845–858. doi: 10.1038/nprot.2015.053.
- Kirkpatrick, D. S., Gerber, S. A. and Gygi, S. P. (2005) 'The absolute quantification strategy: A general procedure for the quantification of proteins and post-translational modifications', *Methods*. Academic Press Inc., 35(3 SPEC.ISS.), pp. 265–273. doi: 10.1016/j.ymeth.2004.08.018.
- Koskiniemi, S. *et al.* (2012) 'Selection-Driven Gene Loss in Bacteria', *PLoS Genet*, 8(6), p. 1002787. doi: 10.1371/journal.pgen.1002787.
- Lavoie, B. D. and Chaconas, G. (1996) 'Transposition of phage Mu DNA', *Current Topics in Microbiology and Immunology*, pp. 83–102. doi: 10.1128/9781555817954.ch17.
- LeBel, M. (1988) 'Ciprofloxacin: Chemistry, Mechanism of Action, Resistance, Antimicrobial Spectrum, Pharmacokinetics, Clinical Trials, and Adverse Reactions', *Pharmacotherapy: The Journal of Human Pharmacology and Drug Therapy*. Pharmacotherapy, 8(1), pp. 3–30. doi: 10.1002/j.1875-9114.1988.tb04058.x.
- Lee, J. Y. *et al.* (2000) 'Crystal structure of NAD(+)-dependent DNA ligase: modular architecture and functional implications.', *The EMBO journal*. European Molecular Biology Organization, 19(5), pp. 1119–29. doi: 10.1093/emboj/19.5.1119.
- Lercher, L. *et al.* (2014) 'Structural insights into how 5-hydroxymethylation influences transcription factor binding', *Chemical Communications*. Chem Commun (Camb), 50(15), pp. 1794–1796. doi: 10.1039/c3cc48151d.
- Lieber, M. R. (2010) 'The Mechanism of Double-Strand DNA Break Repair by the Nonhomologous DNA End-Joining Pathway', *Annual Review of Biochemistry*. Annual Reviews, 79(1), pp. 181–211. doi: 10.1146/annurev.biochem.052308.093131.
- Malhotra, M. and Sahal, D. (1996) 'Anomalous mobility of sulfitolysed proteins in SDS-PAGE Analysis and applications', *International Journal of Peptide and Protein Research*, 48(3), pp. 240–248. doi: 10.1111/j.1399-3011.1996.tb00837.x.
- Miruuchi, M. and Mizuuchi, K. (1989) *Efficient Mu Transposition Requires Interaction of Transposase with a DNA Sequence at the Mu Operator: Implications for Regulation, Cell*.
- Pannunzio, N. R., Watanabe, G. and Lieber, M. R. (2018) 'Nonhomologous DNA end-joining for repair of DNA double-strand breaks', *Journal of Biological Chemistry*. American Society for Biochemistry and Molecular Biology Inc., pp. 10512–10523. doi: 10.1074/jbc.TM117.000374.
- Parsons, L. M., Liu, F. and Orban, J. (2009) 'HU- $\alpha$  binds to the putative double-stranded DNA mimic HI1450 from *Haemophilus influenzae*', *Protein Science*. Wiley, 14(6), pp. 1684–1687. doi: 10.1110/ps.041275705.
- Parsons, L. M., Yeh, D. C. and Orban, J. (2004) 'Solution Structure of the Highly Acidic Protein HI1450

## References

- from Haemophilus influenzae, a Putative Double-Stranded DNA Mimic', *Proteins: Structure, Function and Genetics*, 54(3), pp. 375–383. doi: 10.1002/prot.10607.
- Pastwa, E. and Błasiak, J. (no date) *Non-homologous DNA end joining*.
- Putnam, C. D. and Tainer, J. A. (2005) 'Protein mimicry of DNA and pathway regulation', in *DNA Repair*, pp. 1410–1420. doi: 10.1016/j.dnarep.2005.08.007.
- Rosano, G. L. and Ceccarelli, E. A. (2014) 'Recombinant protein expression in Escherichia coli: Advances and challenges', *Frontiers in Microbiology*. Frontiers Research Foundation. doi: 10.3389/fmicb.2014.00172.
- Ruggeri, F. S. *et al.* (2019) 'Atomic force microscopy for single molecule characterisation of protein aggregation', *Archives of Biochemistry and Biophysics*. Academic Press Inc., pp. 134–148. doi: 10.1016/j.abb.2019.02.001.
- Schwede, T. *et al.* (2003) 'SWISS-MODEL: An automated protein homology-modeling server', *Nucleic Acids Research*. Oxford University Press, 31(13), pp. 3381–3385. doi: 10.1093/nar/gkg520.
- Singleton, M. R. *et al.* (1999) 'Structure of the adenylation domain of an NAD<sup>+</sup>-dependent DNA ligase', *Structure*. Cell Press, 7(1), pp. 35–42. doi: 10.1016/S0969-2126(99)80007-0.
- Sleigh, M. J. (1976) 'The mechanism of DNA breakage by phleomycin in vitro', *Nucleic Acids Research*. Oxford University Press, 3(4), pp. 891–901. doi: 10.1093/nar/3.4.891.
- Söding, J., Biegert, A. and Lupas, A. N. (2005) 'The HHpred interactive server for protein homology detection and structure prediction', *Nucleic Acids Research*. Oxford University Press, 33(SUPPL. 2), p. W244. doi: 10.1093/nar/gki408.
- Tamayo, M. *et al.* (2009) 'Rapid assessment of the effect of ciprofloxacin on chromosomal DNA from Escherichia coli using an in situ DNA fragmentation assay', *BMC Microbiology*. BioMed Central, 9, p. 69. doi: 10.1186/1471-2180-9-69.
- Tataje-Lavanda, L. *et al.* (2019) 'Genomic Islands in the Full-Genome Sequence of an NAD-Hemin-Independent Avibacterium paragallinarum Strain Isolated from Peru'. doi: 10.1128/MRA.
- Unciuleac, M.-C., Goldgur, Y. and Shuman, S. (2017) 'Two-metal versus one-metal mechanisms of lysine adenylation by ATP-dependent and NAD<sup>+</sup>-dependent polynucleotide ligases.', *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, 114(10), pp. 2592–2597. doi: 10.1073/pnas.1619220114.
- Walker, J. R., Corpina, R. A. and Goldberg, J. (2001) 'Structure of the Ku heterodimer bound to DNA and its implications for double-strand break repair', *Nature*. Nature Publishing Group, 412(6847), pp. 607–614. doi: 10.1038/35088000.
- Wei, H. *et al.* (2008) 'The Fidelity Index provides a systematic quantitation of star activity of DNA restriction endonucleases', *Nucleic Acids Research*, 36(9). doi: 10.1093/nar/gkn182.
- Williams, J. G. and Radding, C. M. (1981) 'Partial purification and properties of an exonuclease inhibitor induced by bacteriophage Mu-1.', *Journal of Virology*, 39(2).
- David J. Scott and Peter Schuck (2005) 'A Brief Introduction to the Analytical Ultracentrifugation of Proteins for Beginners', in *Analytical Ultracentrifugation*. Royal Society of Chemistry, pp. 1–25. doi: 10.1039/9781847552617-00001.
- Anderson, B. J. *et al.* (2008) 'Chapter 12 Using Fluorophore-Labeled Oligonucleotides to Measure Affinities of Protein-DNA Interactions', *Methods in Enzymology*. NIH Public Access, pp. 253–272.

## References

- Arosio, D. *et al.* (2004) 'Fluorescence anisotropy studies on the Ku-DNA interaction: Anion and cation effects', *Journal of Biological Chemistry*. *J Biol Chem*, 279(41), pp. 42826–42835.
- Cole, J. L. *et al.* (2008) 'Analytical Ultracentrifugation: Sedimentation Velocity and Sedimentation Equilibrium', *Methods in Cell Biology*. *Methods Cell Biol*, pp. 143–179.
- Hellman, L. M. and Fried, M. G. (2007) 'Electrophoretic mobility shift assay (EMSA) for detecting protein-nucleic acid interactions', *Nature Protocols*. NIH Public Access, 2(8), pp. 1849–1861.
- Hong, P., Koza, S. and Bouvier, E. S. P. (2012) 'A review size-exclusion chromatography for the analysis of protein biotherapeutics and their aggregates', *Journal of Liquid Chromatography and Related Technologies*. Taylor & Francis, pp. 2923–2950.
- Schuck, P. (2007) 'Sedimentation Equilibrium Analytical Ultracentrifugation for Multicomponent Protein Interactions', in *Protein Interactions*. Springer US, pp. 289–316.
- Shee, C. *et al.* (2013) 'Engineered proteins detect spontaneous DNA breakage in human and bacterial cells.', *eLife*. eLife Sciences Publications, Ltd, 2, p. e01222.
- Svedberg, T. and Fåhræus, R. (1926) 'A new method for the determination of the molecular weight of the proteins', *Journal of the American Chemical Society*. American Chemical Society, 48(2), pp. 430–438.
- Canchaya, C. *et al.* (2003) 'Prophage Genomics', *Microbiology and Molecular Biology Reviews*. American Society for Microbiology, 67(2), pp. 238–276. doi: 10.1128/mubr.67.2.238-276.2003.
- Lawrence, J. G. and Roth, J. R. (2014) 'Genomic Flux: Genome Evolution by Gene Loss and Acquisition', in *Organization of the Prokaryotic Genome*. Washington, DC, USA: ASM Press, pp. 263–289. doi: 10.1128/9781555818180.ch15.



## 10. Abbreviations

- Amp: Ampicillin
- AUC: Analytical Ultracentrifugation
- BHI: Briar Heart Infusion
- EDTA: Ethylenediaminetetraacetic acid
- EMSA: Electrophoretic mobility shift assay
- FPLC: Fast Protein Liquid Chromatography
- GFP: Green Fluorescent Protein
- $\beta$ ME: Beta-Mercaptoethanol
- bp: Base Pairs
- Da: Daltons
- DNA: Deoxyribonucleic Acid
- DSBs: Double Stranded Breaks
- dsDNA: Double stranded DNA
- DTT: Dithiothreitol
- HR: Homologous Recombination
- Kd: Dissociation Equilibrium Constant
- kDa: kilo Dalton
- LB: Lysogeny Broth
- LC-MS: Liquid Chromatography-Mass Spectrometry
- MWCO: Molecular Weight Cut Off
- NAD: Nicotinamide Adenine Dinucleotide
- NHEJ: Non-Homologous End Joining
- OD: Optical Density
- OD<sub>600</sub>: Optical Density at 600nm
- ORFs: Open Reading Frames
- PBS: Phosphate-Buffered Saline
- PCR: Polymerase Chain Reaction
- PHAST: PHAge Search Tool
- PMSF: Phenylmethylsulfonyl fluoride
- Rf: Relative mobility
- RNA: Ribonucleic Acid
- rRNA: ribosomal RNA
- RT-PCR: Real time PCR
- SDS-PAGE: Sodium Dodecyl Sulfate Polyacrylamide Gel Electrophoresis
- SEC: Size Exclusion Chromatography
- TAE: Tris-Acetate EDTA
- Tris: Tris (hydroxyl methyl) amino methane
- w/v: weight by volume

### 10.1 Bacterial strains and proteins

- *A. paragallinarum*: *Avibacterium paragallinarum*
  - *D. vulgaris*: *Desulfovibrio vulgaris*
  - *E. coli*: *Escherichia coli*
  - *E. coli* MDS42: *Escherichia coli* MDS42 LowMut  $\Delta$ recA
  - *H. influenzae*: *Haemophilus influenzae*
  - *K. pneumoniae*: *Klebsiella pneumoniae*
  - *N. meningitidis*: *Neisseria meningitidis*
  - *P. aeruginosa*: *Pseudomonas aeruginosa*
  - *R. capsulatus*: *Rhodobacter capsulatus*
  - *S. enterica*: *Salmonella enterica*
  - *S. sonnei*: *Shigella sonnei*
- 
- HiGam: *Haemophilus influenzae* Gam Protein
  - MuGam: Mu Bacteriophage Gam Protein
  - DvGam: *Desulfovibrio vulgaris* Gam Protein
  - KpGam: *Klebsiella pneumoniae* Gam Protein
  - ApGam: Gam Protein
  - SsGam: Gam Protein
  - SeGam: Gam Protein
  - EcGam: *Escherichia coli* Gam Protein
  - NmGam: *Neisseria meningitidis* Gam Protein
  - PaGam: *Pseudomonas aeruginosa* Gam protein
  - RcGam: *Rhodobacter capsulatus* Gam protein
  - HI1450: *Haemophilus influenzae* DNA mimic protein
  - Im9-HiGam: *Haemophilus influenzae* Gam Protein bound to Colicin E9 immunity protein on N-terminus

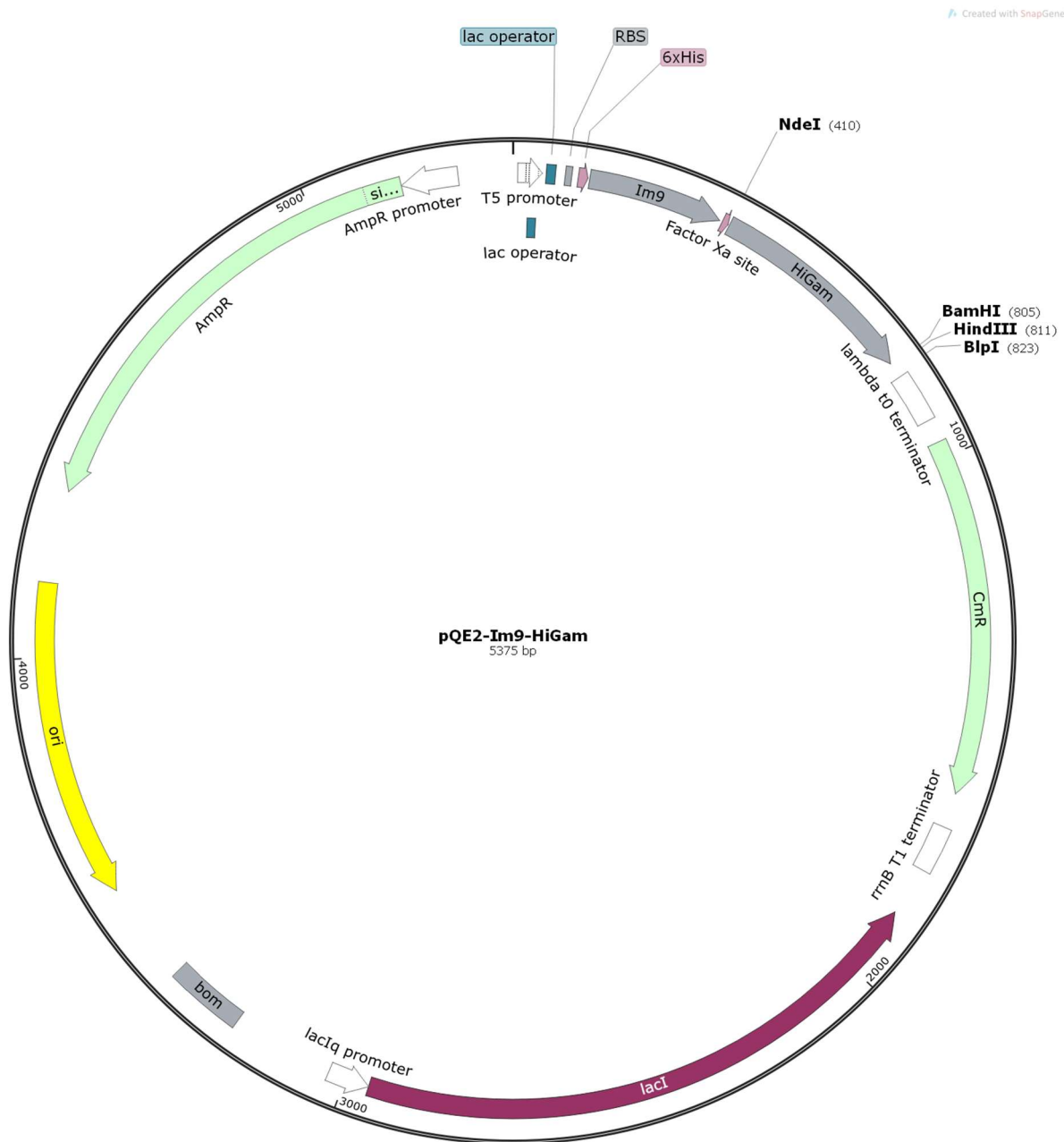
## 11. Appendix

### 11.1 pQE2-Im9-HiGam plasmid sequence

In blue is the sequence for the Im9 gene and in green is the genetic nucleic acid sequence for the HiGam gene.

```
CTCGAGAAATCATAAAAAATTTATTTGCTTTGTGAGCGGATAACAATTATAATAGATTCAATTGTGAG
CGGATAACAATTCACACAGAATTCATTAAGAGGAGAAATTA ACTATGAAACATCACCATCACCAT
CACAATATGGA ACTGAAGCATAGCATTAGTGATTATACAGAAGCTGAATTTTTACAGCTTGTAACAA
CAATTTGTAATGCGGACACTTCCAGTGAAGAAGAACTGGTTAAATTGGTTACACACTTTGAGGAAAT
GACTGAGCACCCTAGTGGTAGTGATTTAATATATTACCCAAAAGAAGGTGATGATGACTCACCTTCA
GGTATTGTAACACAGTAAAACAATGGCGAGCCGCTAACGGTAAGTCAGGATTTAAACAGATCGAA
GGTCGTCATATGAACGATAAAATTGGCGCGACCAGCGAACATTATGCGCCGAAACTGAAAGCGCTG
AAAGAAGAAATTGAACCGCTGCAGAAAGCGGTGCAGGAATATTGCGAAGCGAACCCGCGATGAACT
GACCGAATTTGGCAAACCAAACCGCGAACTTTGTGACCGGCGAAGTGCAGTGGCGCCAGCGCCC
GCCGAGCGTGGCGATTGCGCGCGCGGAAGCGGTGATGGAATTTCTGCAGCGCATGGGCTTTGATC
GCTTTATTCGCACCCGCCAGGAAATTAACAAAGAAGCGCTGCTGAACGAACCGGAAGTGGCGAAAG
GCATTGCGGGCGTGACCATTAACAGGGCCTGGAAGATTTTGTGATTAAACCGTTTGAACAGGATG
CGCGCTAAGGATCCAAGCTTAATTAGCTGAGCTTGGACTCCTGTTGATAGATCCAGTAATGACCTCA
GAACTCCATCTGGATTTGTT CAGAACGCTCGGTTGCCGCCGGCGTTTTTTATTGGTGAGAATCCAA
GCTAGCTTGGCGAGATTTTCAGGAGCTAAGGAAGCTAAAATGGAGAAAAAAATCACTGGATATAACC
ACCGTTGATATATCCCAATGGCATCGTAAAGAACATTTTGAGGCATTTTCAGTCAGTTGCTCAATGTAC
CTATAACCAGACCGTTCAGCTGGATATTACGGCCTTTTTTAAGACCGTAAAGAAAAATAAGCACAAG
TTTTATCCGGCCTTTATTACATTCTTGCCCGCTGATGAATGCTCATCCGGAATTTTCGTATGGCAATG
AAAGACGGTGAGCTGGTGATATGGGATAGTGTTACCCTTGTTACACCGTTTTCCATGAGCAA ACTG
AAACGTTTTCATCGCTCTGGAGTGAATACCACGACGATTTCCGGCAGTTTCTACACATATATTCGCAA
GATGTGGCGTGTTACGGTGAAAACCTGGCCTATTTCCCTAAAGGGTTTATTGAGAATATGTTTTTCGT
CTCAGCCAATCCCTGGGTGAGTTTACCAGTTTTGATTTAAACGTGGCCAATATGGACA ACTTCTTCG
CCCCGTTTTTACCATGGGCAAATATTATACGCAAGGCGACAAGGTGCTGATGCCGCTGGCGATTCA
GGTTCATCATGCCGTTTGTGATGGCTTCCATGTCGGCAGAATGCTTAATGAATTACAACAGTACTGC
GATGAGTGGCAGGGCGGGGCGTAATTTTTTTAAGGCAGTTATTGGTGCCCTTAAACGCCTGGGGTA
ATGACTCTCTAGCTTGAGGCATCAAATAAAACGAAAGGCTCAGTCGAAAGACTGGGCCTTTGTTTTT
ATCTGTTGTTTGTGCGGTGAACGCTCTCCTGAGTAGGACAAATCCGCCCTCTAGATTACGTGCAGTCG
ATGATAAGCTGTCAAACATGAGAATTGTGCCTAATGAGTGAGCTAACTTACATTAATTGCGTTGCGC
TCACTGCCCGCTTTCCAGTCGGGAAACCTGTCGTGCCAGCTGCATTAATGAATCGGCCAACGCGCGG
GGAGAGGCGGTTTGC GTATTGGGCGCCAGGGTGGTTTTTTCTTTTACCAGTGAGACGGGCAACAGC
TGATTGCCCTTACC GCCTGGCCCTGAGAGAGTTGCAGCAAGCGGTCCACGCTGGTTTGCCCCAGCA
GGCGAAAATCCTGTTTGTGATGGTGGTTAACGGCGGGATATAACATGAGCTGTCTTCGGTATCGTCGTA
TCCC ACTACCGAGATATCCGCACCAACGCGCAGCCCGGACTCGGTAATGGCGCGCATTGCGCCCAGC
GCCATCTGATCGTTGGCAACCAGCATCGCAGTGGGAACGATGCCCTCATT CAGCATTTGCATGGTTT
GTTGAAAACCGGACATGGCACTCCAGTCGCCTTCCGTTCCGCTATCGGCTGAATTTGATTGCGAGT
GAGATATTTATGCCAGCCAGCCAGACGACGACGCGCCGAGACAGAACTTAATGGGCCCCGCTAACAG
CGCGATTTGCTGGTGACCCAATGCGACCAGATGCTCCACGCCAGTCGCGTACCGTCTTCATGGGAG
AAAATAACTGTTGATGGGTGTCTGGTCAGAGACATCAAGAAATAACGCCGGAACATTAGTG CAG
```

GCAGCTTCCACAGCAATGGCATCCTGGTCATCCAGCGGATAGTTAATGATCAGCCACTGACGCGTT  
GCGCGAGAAGATTGTGCACCGCCGCTTTACAGGCTTCGACGCCGCTTCGTTCTACCATCGACACCAC  
CACGCTGGCACCCAGTTGATCGGGCGGAGATTTAATCGCCGCGACAATTTGCGACGGCGCGTGCAG  
GGCCAGACTGGAGGTGGCAACGCCAATCAGCAACGACTGTTTGCCCGCCAGTTGTTGTGCCACGCG  
GTTGGGAATGTAATTCAGCTCCGCCATCGCCGCTTCCACTTTTTCCCGCGTTTTTCGAGAAACGTGGC  
TGGCCTGGTTCACCACGCGGGAACGGTCTGATAAGAGACACCGGCATACTCTGCGACATCGTATA  
ACGTTACTGGTTTCACATTCACCACCCTGAATTGACTCTTCCGGGCGCTATCATGCCATACCGCGA  
AAGTTTTGCACCATTCGATGGTGTGCGAATTTCCGGCAGCGTTGGGTCCTGGCCACGGGTGCGCA  
TGATCTAGAGCTGCCTCGCGGTTTTCGGTGATGACGGTGAAAACCTCTGACACATGCAGCTCCCGGA  
GACGGTCACAGCTTGTCTGTAAGCGGATGCCGGGAGCAGACAAGCCCGTCAGGGCGCGTCAGCGG  
GTGTTGGCGGGTGTGCGGGGCGCAGCCATGACCCAGTCACGTAGCGATAGCGGAGTGTATACTGGCT  
TAACTATGCGGCATCAGAGCAGATTGACTGAGAGTGACCACATGCGGTGTGAAATACCGCACAG  
ATGCGTAAGGAGAAAATACCGCATCAGGCGCTTCCGCTTCTCGCTCACTGACTCGCTGCGCTCG  
GTCGTTCCGCTGCGGGCAGCGGTATCAGCTCACTCAAAGGCGGTAATACGGTTATCCACAGAATCA  
GGGATAACGCAGGAAAGAACATGTGAGCAAAAGGCCAGCAAAAGGCCAGGAACCGTAAAAAGG  
CCGCGTTGCTGGCGTTTTTCCATAGGCTCCGCCCCCTGACGAGCATCACAAAATCGACGCTCAAG  
TCAGAGGTGGCGAAACCCGACAGGACTATAAAGATACCAGGCGTTTTCCCTGGAAGCTCCCTCGT  
GCGCTCTCCTGTTCCGACCCTGCCGTTACCGGATACCTGTCCGCTTTCTCCCTTCGGGAAGCGTGG  
CGTTTTCTCATAGCTCACGCTGTAGGTATCTCAGTTCGGTGTAGGTGTTTCGCTCCAAGCTGGGCTGT  
GTGCACGAACCCCCGTTAGCCCGACCGCTGCGCCTTATCCGGTAACTATCGTCTTGAGTCCAACCC  
GGTAAGACACGACTTATCGCCACTGGCAGCAGCCACTGGTAACAGGATTAGCAGAGCGAGGTATGT  
AGGCGGTGCTACAGAGTTCTGAAGTGGTGGCCTAACTACGGCTACACTAGAAGGACAGTATTTGG  
TATCTGCGCTCTGCTGAAGCCAGTTACCTTCGAAAAAGAGTTGGTAGCTCTTGATCCGGCAAACAA  
ACCACCGCTGGTAGCGGTGGTTTTTTTTGTTTGCAAGCAGCAGATTACGCGCAGAAAAAAGGATCTC  
AAGAAGATCCTTTGATCTTTTCTACGGGGTCTGACGCTCAGTGGAACGAAAACCTCACGTTAAGGGAT  
TTTGGTCATGAGATTATCAAAAAGGATCTTCACCTAGATCCTTTTAAATTAATAAATGAAGTTTTAAAT  
CAATCTAAAGTATATATGAGTAACTTGGTCTGACAGTTACCAATGCTTAATCAGTGAGGCACCTATC  
TCAGCGATCTGTCTATTTTCGTTTCATCCATAGTTGCCTGACTCCCCGTCGTGTAGATAACTACGATACG  
GGAGGGCTTACCATCTGGCCCCAGTGCTGCAATGATACCGCGAGACCCACGCTCACCGGCTCCAGA  
TTTATCAGCAATAAACCAGCCAGCCGGAAGGGCCGAGCGCAGAAGTGGTCCTGCAACTTTATCCGC  
CTCCATCCAGTCTATTAATTGTTGCCGGGAAGCTAGAGTAAGTAGTTCGCCAGTTAATAGTTTGC  
AACGTTGTTGCCATTGCTACAGGCATCGTGGTGTACGCTCGTTCGTTGGTATGGCTTATTAGCTC  
CGGTTCCCAACGATCAAGGCGAGTTACATGATCCCCATGTTGTGCAAAAAGCGGTTAGCTCCTTC  
GGTCTCCGATCGTTGTCAGAAGTAAGTTGGCCGAGTGTATCACTCATGGTTATGGCAGCACTGC  
ATAATTCTTACTGTCATGCCATCCGTAAGATGCTTTTTCTGTGACTGGTGAGTACTCAACCAAGTCA  
TTCTGAGAATAGTGATGCGGCGACCGAGTTGCTCTTGCCCGGCGTCAATACGGGATAATACCGCGC  
CACATAGCAGAACTTTAAAAGTGCTCATCATTGGAAAACGTTCTTCGGGGCGAAAACCTCTCAAGGAT  
CTTACCGCTGTTGAGATCCAGTTCGATGTAACCCACTCGTGCACCCAACTGATCTTCAGCATCTTTTAC  
TTTACCAGCGTTTCTGGGTGAGCAAAAACAGGAAGGCAAAATGCCGCAAAAAGGGGAATAAGGG  
CGACACGGAAATGTTGAATACTCATACTTCTTTTTCAATATTATTGAAGCATTATCAGGGTTATT  
GTCTCATGAGCGGATACATATTTGAATGTATTTAGAAAAATAAACAAATAGGGGTTCCGCGCACATT  
TCCCGAAAAGTGCCACCTGACGTCTAAGAAACCATTATTATCATGACATTAACCTATAAAAATAGGC  
GTATCACGAGGCCCTTTCGTCTTCAC



**Figure 54.** Schematic view of the pQE2-Im9-HiGam. The sequence above is used to create the schematic view, with the Im9-HiGam gene sequences, NdeI and BamHI restriction sites labelled to show the sites of digestion for MuGam homologue gene insertion.

**11.2 Gam homologue gene sequences optimised for transcription and translation in *E. coli* K-12 MG1655 as confirmed by sequencing by Eurofins Scientific Genomics, with translated amino acid sequences.**

***Klebsiella pneumoniae* CCIC01000013 (WP\_040211999.1)**

ATGGCAAAGCAAAAAAGCGACTGAAAGCAACCGCAGCCCTGTATGTTGCCAGACTAAAGAGGA  
AGTGATCACCGGTATCAAGCTGCTGGGCGATATCCAGCGTGAGCTGATCCGTGTCAAACCGAAAT  
GAACGATTCCATTGCTGAAATCACCGCATCTCACAGCCCTGCAATCGAAGGCCTGAAAGCTAAAATG  
GAAGAACTGCAGAACGGTATCCAACTTGGTGCGAAGCTCATCGCGATGAACTGACGAACAACGGT  
AAAGTTAAATTTGCGAACCTGACGACTGGCGAAGTTCAGTGGCGCAACCGTCCACCATCTGTTAGCA  
TCCGTGGCGCAGACGCTGTGATCGACTTTCTGAAGCGTCTGGGTCTGAAACGCTTTATCCGTACCAA  
AGACGAACTGAACAAAGAAGCGATGCTGAACGAAAAAGATGCAGTTAAAAATATCCCGGGCATCAC  
CATCAAACGGACGTTGAAGATTTTAGCATTATTCCGTTTGAGCAGGAAGTACAGTAA

**MAKAKKRLKATAALYVAQTKKEEVITGIKLLGDIQRELIRVETEMNDSIAEITASHSPAIEGLKAKMEELQ  
NGIQTWCEAHRDELNNGKVKFANLTTGEVQWRNRPPSVSIRGADAVIDFLKRLGLKRFIRTKDELNK  
EAMLNEKDAVKNIPGITIKTDVEDFSIIPFEQEVQ**

***Avibacterium paragallinarum* JF4211 (WP\_017806105.1)**

ATGTCCAAAACCAAACCTGAAGTCTGATACTATCCGTTACCAGACTCGTGAAGAAGTTGAGATCGCAA  
TCAAAGACATCGGCGATCTGCAGCGTGAAGTGCAGCGCCTGGCTACCCATCAGAATGATGAACTGG  
CGGCCATCACCGAAAAATACGCTCCGAAAATCACCGCTCTGCAGGAACAGATGAAACCGCTGCAGA  
AGGCGATCGAAGTTTGGTGCGAGGCGAACCGCGCGGAGCTGACCCAGAACGGTAAAACCTAAAACCT  
GGTTCTTTTAACACCGGCGAGGTCCAGTGGCGTCAGCGCCCACCGTCCGTATCCATTCGCAAAGCAG  
ACGAAGTACTGGCGCGCCTGCGTGCACTGGGCCTGACTCAGTTCATTCTGTACCAAAGAGGAGCCGA  
ACAAAGAAGCTATGCTGGCTGAACCGAATATCGCCAGCACGGTGACTGGCATTACTATCAAACCGG  
CGGTGGAAGATTTCTGTTATCAAGCCGTTCAACAGGAAGTTTAA

**MSKTKLKSDTIRYQTREEVEIAIKDIGDLQRELQRLATHQNDELAITEKYAPKITALQE QMKPLQKAIE  
VWCEANRAELTQNGKTKTGSFNTGEVQWRQRPPSVSIRKADEVLARLRALGLTQFIRTKKEEPNKEAM  
LAEPNIASTVTGITIKTAVEDFVIKPFQEV**

***Shigella sonnei* CDPH\_C88 contig9 (WP\_001107937.1)**

ATGGCTAAACCCGCAAACGTATTGTAATGCTGCGGCAGCTTACGTGCCGCAGTCTCGTGATGCGG  
TGGTTTGCACATCCGTGCGATTGGCGATCTGCAGCGTGAGGCTGCACGTCTGGAAACTGAAATGA  
ACGACGCGATCGCGGAAATTACCGAAAAATACGCAAGCCAGATTGCGCCGCTGAAGACCTCTATCG  
AAACCCTGAGCAAAGGTGTACAGGGCTGGTGCAGGCTAACCGTGATGAACTGACTAACGGCGGT  
AAAGTCAAACCGCTAACCTGGTCACGGGCGACGTTTTCTTGGCGTCAGCGTCCGCCTAGCGTGTCCA  
TTCGTGGCGTGGATGCGGTAATGGAAACCTGGAACGTCTGGGCCTGCAGCGTTTTATCCGCACTA  
AACAGGAAATCAACAAGGAGGCGATCCTGCTGGAACCGAAAGCCGTAGCGGGTGTGCTGGTATTA  
CGGTAAAATCCGGCATTGAAGATTTCTCTATTATCCCTTTGAACAGGAAGCCGGTATTTAA

**MAKPAKRIRNAAAAYVPQSRDAVVCDIRRIGDLQREARLETENDAIAEITEKYASQIAPLKTSIETLS  
KGVQGWCEANRDELNNGKVKTNLVTGDVSWRQRPPSVSIRGVDVAVMETLERLGLQRFIRTKQEI  
NKEAILLEPKAVAGVAGITVKSIEDFSIIPFEQEAGI**

**Salmonella enterica (WP\_044127832.1)**

ATGGCCAAACAGGTTAAACGCATTCTGTTCTGCAGCCGCGGCTTACGTGCCTCAAAGTCGCGACGCG  
GTCGTGTGTGATATTCGCCGTATTGGCGATCTGCAACGTGAAGCAGCCCGCCTGGAGACGGAGATG  
AATGATGCGATTGCGGAAATTACAGAAAAATACGCTTCCCAGATTGCGCCGCTGAAGACTTCAATTG  
AAACCCTGAGCAAAGGAGTACAGGGATGGTGCGAAGCGAACCCTGATGAACTCACAATGGTGGC  
AAGGTAAGAGCGCAAATTTAGTTACCGGGGATGTGCAATGGCGTCAACGTCCGCCGAGTGTCTCT  
ATTCGCGGGGTTCGATGCGGTTATGGAAACCTTGGAGCGTCTGGGCTTACAACGCTTTATCCGTACAA  
AACAGGAAATTAATAAAGAAGCGATTCTGTTAGAACCTAAAGCAGTAGCGGGTGTAGCTGGCATT  
CGTTAAATCGGGCATTGAAGACTTTAGCATTATCCCGTTTGAACAGGAGGCAGGGATTTAA

**MAKQVKRIRSAAAAYVPQSRDAVVCDIRRIDLQREARLETENMDAIAEITEKYASQIAPLKTSIETLS  
KGVQGWCEANRDELNNGGKVKSANLVTGDVQWRQRPPSVSIRGVDVAVMETLERLRLGLQRFIRTKQEI  
NKEAILLEPKAVAGVAGITVKSIEDFSIIPFEQEAGI**

**Escherichia coli 0157:H7 str. Sakai (WP\_001129553.1)**

ATGGCAAAACGCGTTACAAAACCTGAAAGCTGCGGCTGAAGCTGCTCCGCAGACTCGTGAAGAAGTT  
TCTCGTGACATCCGTACGCTGGGCGACATCCAGCGTGAAGCTCTGCGTCTGGAAACGGCTATGAAC  
GACGAAGTGGCCGAAATTACTGCCCGCTACACCCCGCAGATCGAAAATCTGAAGAAAGAGATCAAA  
GTCCTGTTCAAAGGTATCCAGGACTGGTGCAAAACCAACCGCGATGAACTGACCAACGGTGGCAAA  
ACCAAAACGGCTAACCTGACGACTGGTACCGTCAGCTGGCGTCTGGGTAACCCATCTTGTAGCGTAT  
CTCGCGATGTTGAAGGTGTGATCGAGATGCTGCGCCGTATGGGTCTGGAACGTTTCATCCGTACGA  
AAGAAGAAGTTAACAAAGAGAGCCGTTCTGGCTGAACCGGATGCTGTAAAAGGTATCGCAGGTATCA  
AAGTTAACAAAGGGTGTGAGTCTTTCTATGTGGAGCCGTTTGAACAGGACGCCGGACTGAATAAAT  
AA

**MAKRVTKLKAAAEAAPQTREEVSRDIRTLGDIQREALRLETAMNDEVAEITARYTPQIENLKKEIKVLFK  
GIQDWCKTNRDELNNGGKTKTANLTTGTVSWRLGNPSCSVSRDVEGVIEMLRRMGLERFIRTKKEEVN  
KEAVLAEPDAVKGIAGIKVNKGAESFYVEPFQDAGLNK**

**Neisseria meningitidis 768\_NMEN 453\_36901\_401297\_365+, 259+, 103+  
(WP\_049323767.1)**

ATGGCAAAAACCCGTATCAAACAACCGGCTATTGAGGCAGCCAGGACAAAGCAGAAGTTACGGCC  
TTCATTCGTAAAATTGGCGATCTGCAGCGTGAAGTTAAACGTCTGGAAACGGAAGCGGGCGATAAA  
AAAGCAGTTATTGAAGAAGAATACGCGGCGAAAGCGGCACCGATGTGTGCGGAAATCATGTCCCTG  
ACGGAACGTGTGGCAGCCTACTGCGAAGCTCACAAGGACGAACTGACCGAAAACGGTAAACTAA  
AACCGTTGATTTCACTACCGGTCTGATCAAATGGCGTATTCGTCCGCCGTCTGTAAAAGTTACCGGTG  
TGGCAGCCGTAAGGAAATGGCTGTCTGAAAAATCCGCATTGCGGGAATTCGTACGTACCAAGAAAG  
AAATCGACAAAGACGCCATCCTGAACCAGAAAGAACGCTTCTCTGACGGCCAGGTTCCGGGTATCA  
AGATCGTGTCTGGTCTGGAAGACTTTGTTATTGAGCCAACCGAGCAGGAGTTGGCGTAA

**MAKTRIKQPAIEAAQDKAEVTA FIRKIGDLQREVKRLETEAGDKKAVIEEEYA AKAAPMCAEIMSLTER  
VAAYCEAHKDEL TENGKTKTV DFTTGLIKWRIRPPSVKVTGVAAVLEWLSEKSAFAEFVRTKKEIDKDA  
ILNQKERFSDGGQVPGIKIVSGLEDFVIEPTEQELA**

### **11.3 Genetic and amino acid sequences for DNA Ligase protein fragments from *A. paragallinarum* JF4211.**

Protein fragment 1 of LigA from *A. paragallinarum* JF4211 (Purple indicates the conserved residues as seen in **Fig. 5-6 (3. Results - DNA Double strand break repair in Bacteria containing MuGam homologue genes)**).

MIEQSPAQQIEQLRQTLRHHEYQYHVLDNQQIPDAEYDRLFNQLKALEQAYPEFADPHSPTQRVGAKPL  
AGFAQVTHEIPMLSLDNAFSDDEF AAFVKRIQDRLGILPDPLTFCCPEKLDGLAVSILYENGVLTQAATRG  
DGTTGEDITANIRTIRNIPLQLLDNPPDRLEVRGEVFMPOAGFEKLN ETALAKGEKTFANPRNAAAGSL  
RQLDPKITRQRP[]VNAQCLWDWCGRRYAVTQHAFRTIAMAEIPRYPRE

[] = [Identified position of insertion and frame shift]

Protein fragment 2 fragment *A. paragallinarum* JF4211

MAEKRPHLGYDIDGTVLKINDIALQDRLGFISKAPRWAIAYKFP AQEELTILNDVEFQVGRGTGAITPVAKLE  
PVFVAGVTVSNATLHNGDEIERLNIAGDVIIRRAGDVIPQIIGVVADRRPADAKPIIFPTRCPVCGSLITRI  
EGEAVARCTGGLFCEAQRKEALKHFVSRKAMDIDGVGAKLIEQLVDRELIHTPADLFKLDEVTLMLRLERM  
GPKSAENALNSLEKAKKTTLARFIFALGIRDVGEATALNLANHFKNLDAVKQATIEQLQEVDPDVGEEVAN  
RIYVFWREAHNVEVVEDLLAQGIHWDDVEIKEVSENPLKGKTVVLTGTLTQMTRNDAKALLQQLGSKVT  
GSISAKTDYLIAGDNAGSKLNKALELNVKILTEDEFLIVQKSIA

*A. paragallinarum* JF4211 LigA protein when the suspected insertion is removed:

MIEQSPAQQIEQLRQTLRHHEYQYHVLDNQQIPDAEYDRLFNQLKALEQAYPEFADPHSPTQRVGAKPL  
AGFAQVTHEIPMLSLDNAFSDDEF AAFVKRIQDRLGILPDPLTFCCPEKLDGLAVSILYENGVLTQAATRG  
DGTTGEDITANIRTIRNIPLQLLDNPPDRLEVRGEVFMPOAGFEKLN ETALAKGEKTFANPRNAAAGSL  
RQLDPKITRQRPLMLNAYGIGVADGMQLPNTHFARLQWLKSLGIPVNKEIELCNGVENVKFYHTMAEK  
RPHLGYDIDGTVLKINDIALQDRLGFISKAPRWAIAYKFP AQEELTILNDVEFQVGRGTGAITPVAKLEPVFV  
AGVTVSNATLHNGDEIERLNIAGDVIIRRAGDVIPQIIGVVADRRPADAKPIIFPTRCPVCGSLITRIEGEA  
VARCTGGLFCEAQRKEALKHFVSRKAMDIDGVGAKLIEQLVDRELIHTPADLFKLDEVTLMLRLERMGPKS  
AENALNSLEKAKKTTLARFIFALGIRDVGEATALNLANHFKNLDAVKQATIEQLQEVDPDVGEEVANRIYV  
FWREAHNVEVVEDLLAQGIHWDDVEIKEVSENPLKGKTVVLTGTLTQMTRNDAKALLQQLGSKVTGSISA  
KTDYLIAGDNAGSKLNKALELNVKILTEDEFLIVQKSIA

### **11.4 Amino acid sequences for key proteins, Ku70, Ku80, MuGam, HiGam and DvGam.**

Human Ku70:

MSGWESYYKTEGDEEAEEQEENLEASGDYKYSGRDSLIFLVDASKAMFESQSEDELTPFDMSIQCIQSV  
YISKIISDRDLLAVVFGTEKDKNSVNFKNIVLQELDNPGAKRILELDQFKGQQGQKRFQDMMGHGSD  
YSLSEVLWVCANLFSVDVQFKMSHKRIMLFTNEDNPHGNDSAKASRARTKAGDLRDTGIFLDMHLKPK



## Appendix

GGFDISLFYRDIISIAEDEDLRVHFEESKLEDLLRKVRAKETRKRALSRLLKLNKDIVISVGIYNLVQKALKP  
PPIKLYRETNEPVKTKTRTFNTSTGGLLLPSDTKRSQIYGSRQIILEKEETEELKRFDDPGLMLMGFKPLVLLK  
KHHYLRPSLFVYPEESLVIGSSTLFSALLIKCLEKEVAALCRYTPRRNIPPYFVALVPQEEELDDQKIQVTPPG  
FQLVFLPFADDKRKMPFTEKIMATPEQVGKMKAIVEKLRFTYRSDSFENPVLQQHFRNLEALALDLMEPE  
QAVDLTLPKVEAMNKRLGSLVDEFKELVYPPDYNPEGKVTKRKHDNEGSGSKRPKVEYSEELKTHISKG  
TLGKFTVPMLKEACRAYGLKSGLKKQELLEALTKEHFQD

(red: N and C terminal regions, black: amino acids 262-446 DNA binding region used to create the truncated model for dimeric MuGam (d'Adda di Fagagna *et al.*, 2003)).

Human Ku80:

MVRSGNKAAVVLCMDVGFTMSNSIPGIESPFEQAKKVITMFVQRQVFAENKDEIALVLFGTGTDNPLS  
GGDQYQNIIVHRHMLMPDFDLLEDIESKIQPGSQQADFLDALIVSMDVIQHETIGKKFEKRHIEIFDLSR  
FSKSQLDIIHSLKKCDISLQFFLPFLGKEDGSGDRGDGPFRLGGHGPSFPLKGITQQKEGLEIVKMMVIS  
LEGEDGLDEIYSFSESLRKLKCVFKKIERHSIHWPCLRTIGSNLSIRIAAYKSILQERVKKTWTVVDAKTLKEDI  
QKETVYCLNDDDETEVLKEDIQGFYRGSIVPFSKVDEEQMKYKSEGKCFVSLGFCSSQVQRRFFMGN  
QVLKVFVAARDDEAAVALSSLIHALDDLDMVAIVRYAYDKRANPQVGVAFPHIKHNYECLVYVQLPFME  
DLRQYMFSSLKNSKYPTEAQLNAVDALIDSM SLAKKDEKTDLTLEDLFPPTKIPNPRFQRLFQCLLHRAL  
HPREPLPPIQQHIWNMLNPPAEVTTKSQIPLSKIKTLFPLIEAKKQVTAQEIQDNHEDGPTAK

(red: N and C terminal regions, black: amino acids 257-433 DNA binding region used to create the truncated model for dimeric MuGam (d'Adda di Fagagna *et al.*, 2003)).

HiGam polypeptide amino acid sequence

MATKVKSQAKLRFVSVQVQSAIKEIGDLSREHTRLATEMNDKIGATSEHYAPKLKALKEEIEPLQKAVQE  
YCEANRDELTEFGKTKTANFVTGEVQWRQRPPSVAIRGAEAVMEFLQRMGFDRFIRTRQEINKEALLNE  
PEVAKGIAGVTIKQGLEDFVIKPFQDAR

MuGam polypeptide amino acid sequence:

MAKPAKRIKSAAYVPQNRDAVITDIKRIGDLQREASRLETEMNDIAIEITEKFAARIAPIKTDIETLSKGV  
QGWCEANRDELNTGGKVKTANLVTGDVSWRVRPPSVSIRGMDAVMETLERLGLQRFIRTKQEINKEAI  
LLEPKAVAGVAGITVKSGIEDFSIIPFEQEAGI

DvGam polypeptide amino acid sequence from crystallographic 3D structure (PDB: 2P2U):

SLSRRKPNPVIVADIRQAEGALAEIATIDRKVGEIEAQMNEAIDAAKARASQKSAPLLARRKELEDGVATF  
ATLNKTEMFKDRKSLDLGFGTIGFRLSTQIVQMSKITKDMTLERLRQFGISEGIRIKEDVNKEAMQGWPD  
ERLEMVGLKRRTTDAFYIEINREEVADTAA