Structural and Functional Characterisation of Bacterial Polysaccharide Synthesis and Transport in Health and Disease

Gareth Ashworth

Submitted for the degree of Doctor of Philosophy (PhD) to the Norwich Medical School, University of East Anglia on December 2020

Acknowledgements

This thesis represents years of labour with the input, helpful and not, of dozens of scientists. I would first like to thank Professor Changjiang Dong and Dr Andrew Hemmings for providing me with the opportunity and the necessary space to perform this work. Your experience, wisdom and support has been invaluable in the directing of this project. I would also like to thank Dr Tom Clarke for providing excellent advice when my supervisors were unavailable.

I would like to thank the members of the floor two of the School of Biological Sciences for their help and support, with a special thanks to Dr Melissa Salmon, Dr Marcus Edwards and Dr Yinghong Gu for the supervision they were able to provide. Recognition is also due for the work of Dr Andrea Hall, Andrew Loveday and Jasmine Waters in maintaining the labs, an achievement which been nothing short of legendary.

A PhD is sometimes a somewhat daunting and isolating task, so I would like to thank the people who helped make this task just a little less intimidating; Dr Isabella Acquistapace, Dr Raquel Rodriguez and soon-to be Drs Hans Pfalzgraf, Lauren Mills, Hannah Wright and even Joseph Shepherd. You have helped provide an environment with community where work is not simply a job.

Last, but not least, I would like to thank my fiancé Francesca Hepplewhite for her loving devotion, support and understanding.

Abstract

The *Pseudomonas aeruginosa pel* operon is involved in pathogenic biofilm formation via synthesis of the Pel exopolysaccharide. Described here is the cloning and expression of the seven proteins of the Pel operon. A 2.6 Å resolution structure of an N-terminal 53 kDa fragment of PelA has been solved, revealing an α-1,4-N-acetylgalactosaminidase domain with an unusual $(\beta\alpha)_7$ TIM barrel fold and a second domain belonging to the glutamine amidotransferase-like superfamily which contains a degenerate active site and mediates oligomerisation of PelA into a trimeric channel.

The glycosyltransferase WaaB from *Salmonella enterica* serotype Typhmirium is involved in essential lipopolysaccharide synthesis. Key residues were identified from the previously-solved structure of UDP-bound WaaB (PDB: 5N80). These residues were mutated and investigated using a glycosyltransferase assay for activity and five were found to contribute significantly towards activity. The T273A mutant was found to enhance activity and was further investigated to determine that this does not increase substrate promiscuity. The structure of this mutant co-crystallised with UDP was determined at 2.5 Å, revealing that WaaB T273A adopts a tighter, closed conformation in comparison to the open conformation of the wild type protein, allowing the identification of F13 and S265 as potentially important to protein activity and providing evidence against membrane-associated activation of WaaB.

Lactobacillus reuteri forms symbiotic biofilms within the guts of host organisms. The glycosyltransferase GtfC from *Lactobacillus reuteri* strains 100-23 and 53608 share ~97% sequence homology but show specific activity towards UDP-glucose and UDP-N-acetylglucosamine, respectively. The crystal structure of $GtfC_{100\text{-}23}$ was determined at 2.6 Å resolution in both the apo form and in complex with UDP. Key residues involved in catalysis, catalysis-related tetramerization and acceptor and donor substrate binding have been identified for functional analysis by collaborators. Candidate residues responsible for UDP-sugar specificity have been identified as L174, S238 and F240.

Contents

List of Tables

List of Figures

Abbreviations

αX:	α-helix number X
aSEC:	Analytical size exclusion chromatography
ABR:	Antibiotic resistant
ARM:	Armadillo-like helical
βX:	β-strand number X
bp:	Base pair
c-di-GMP:	Bis-(3′–5′)-cyclic dimeric guanosine monophosphate
CMP:	Cytidine monophosphate
DNA:	Deoxyribonucleic acid
ECM:	Extracellular matrix
eDNA:	Extracellular deoxyribonucleic acid
EPS/CPS:	Extracellular and capsular polysaccharide
Gal:	D-galactose
GalNAc:	N-acetyl-galactosamine
GATase:	Glutamine amidotransferase
GH:	Glycoside hydrolase
Glc:	D-glucose
GlcNAc:	N-acetyl-glucosamine
Gtf or GT:	Glycosyltransferase
IM:	Inner membrane
IMAC:	Immobilised metal affinity chromatography
Kdo:	α-3-deoxy-d-manno-oct-2-ulosonic acid
LPS:	Lipopolysaccharide
MATE:	Multi-Antimicrobial Extrusion (transporter)
MOP:	Multidrug/oligosaccharidyl-lipid/polysaccharide
OLF:	Oligosaccharidyl-lipid flippase
OM:	Outer membrane
OMP:	Outer membrane protein
PDB:	Protein data bank
PST:	Polysaccharide transporter

SSRP:	Serine-rich repeat protein
SS1:	Sucrose synthase-1
TPR:	Tetratricopeptide repeat
UDP:	Uridine diphosphate
UDP-GA:	Uridine diphosphate-glucuronic acid
UDP-Gal:	Uridine diphosphate-galactose
UDP-GalNAc:	Uridine diphosphate-N-acetylgalactosamine
UDP-Glc:	Uridine diphosphate-glucose
UDP-GlcNAc:	Uridine diphosphate-N-acetylglucosamine

Chapter 1 – Introduction

1.1.1 Antibiotic resistance

Antibiotic resistance (ABR) has become acknowledged to be one of the greatest challenges to face public health in our generation, as shown by the voting of the British public to designate finding means to combat this growing concern as the target of the £8 million Longitude Prize[1]. Antibiotics, as a subset of antimicrobials, are used in the treatment of bacterial infections, varying from mild to life threatening, in humans, pets and livestock. They are natural or synthetic compounds with bactericidal properties. Before the discovery and commercialisation of antibiotics medical practitioners were forced to treat dangerous infections through damaging procedures such as amputation and life expectancy was 20-40 years lower than it is today. Bacterial pandemics have become a thing of the past, with a 2017 outbreak of *Yersinia pestis* in Madagascar resulting in an 8% fatality rate, thanks to appropriate treatment with antibiotics, compared to the historic plagues that devastated two-thirds of the population of Europe[2,3]. It therefore behoves us to take threats to the effectiveness of antibiotics seriously.

ABR is the ability for a microbe to resist treatment with an antibiotic. The emergence of antibiotic resistance is a global phenomenon that has occurred due to the overuse of antibiotics, both in the treatment of infections and as a preventative measure in the livestock industry[4]. Common examples of misuse are the use of antibiotics to treat viral infections such as cold and flu, for which they are ineffective, and the addition of antibiotics to cattle feed. This overabundance of antibiotics has generated a strong selection pressure for the evolution of mechanisms of antibiotic resistance. While resistance to a single antibiotic can be worked around by use of a different antibiotic the new treatment will also provide a strong selection pressure and has led to the evolution of multi-drug resistant bacteria that become increasingly difficult to treat effectively. Further exacerbating this issue is that resistance is frequently extended to entire families of antibiotics and that resistance genes are often found on plasmids, which can be transferred to other bacteria to propagate resistance in other species.

This thesis operates under the umbrella of tackling the problem of ABR by improving our understanding of the structural biology of proteins involved in three contributing mechanisms; biofilm formation by *Pseudomonas aeruginosa*, outer membrane biogenesis in *Salmonella enterica* and through gut colonisation by the bacteria *Lactobacillus reuteri*, which acts as a commensal competitor to pathogens, such as *Clostridium difficile* and *S. enterica,* whose invasion is often dependent upon ABR.

1.1.2 Antibiotic resistance and biofilms

Biofilms are bacterial communities that adhere to an abiotic or biotic surface that surround themselves in a self-made extracellular matrix. Formation of a biofilm provides an inherent resistance to antimicrobial

agents and is the common root cause of many chronic or persistent bacterial infections[5].The switch from planktonic, free-floating, cells to a biofilm phenotype has been shown to cause a thousand-fold increase in resistance to tobramycin in *P. aeruginosa* infections of urinary catheters[6] and the biofilm phenotype is present in chronic antibiotic resistant *P. aeruginosa* infection of the lungs of patients with cystic fibrosis[7]. While the focus of this thesis is on *P. aeruginosa* biofilms, it is important to note that biofilm formation is ubiquitous throughout pathogenic bacteria.

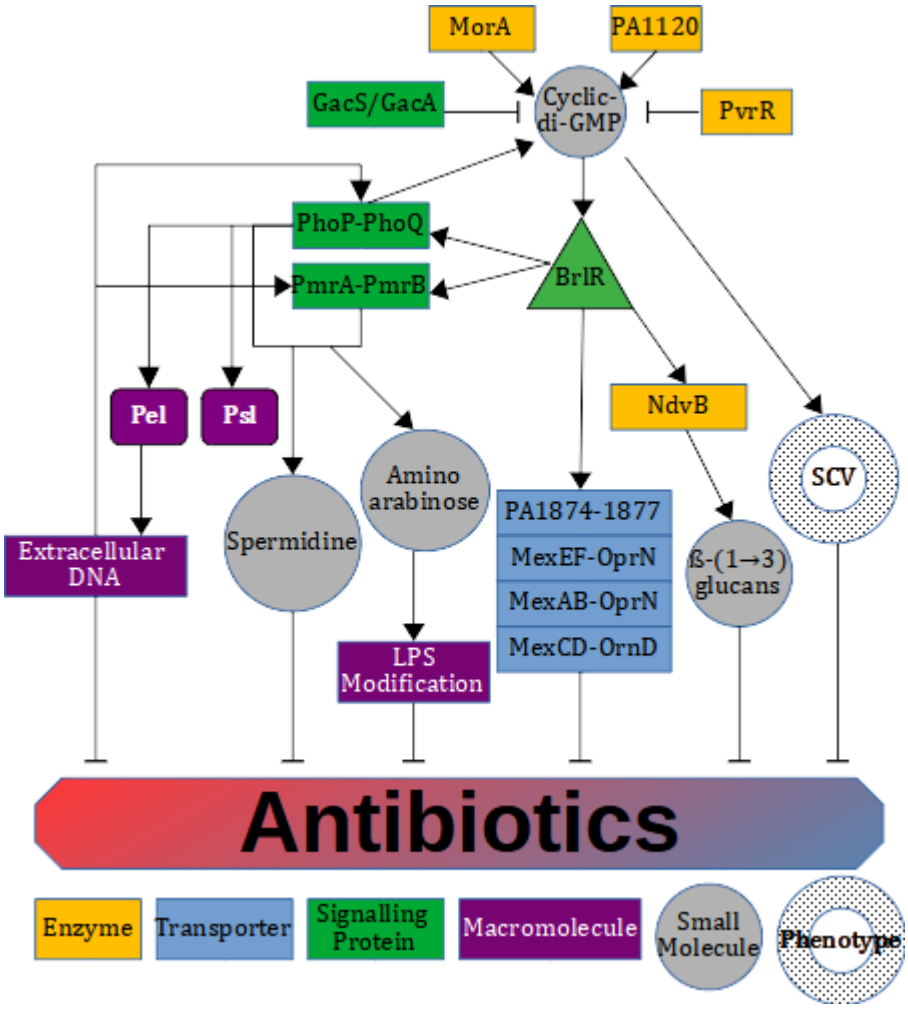

Figure 1 Flow schematic of biofilm associated antibiotic resistance mechanisms. Cyclic-di-GMP acts as the central secondary messenger stimulating biofilm formation. Its formation is controlled by a balance of phosphodiesterases, which break the compound down, and diguanylate cyclases, which synthesise it, under the control of two component systems. Cyclic-di-GMP drives gene expression of multiple two component systems and regulators, generating positive feedback and activating the expression or synthesis of antibiotic resistance effectors. Exopolysaccharides (Pel/Psl) sequester antibiotics, either directly or indirectly through extracellular DNA localisation, to prevent them from reaching the cell. Aminoarabinose modification of LPS and spermidine reduce the permeability of the outer membrane to cationic antimicrobials. Multi-drug efflux transporters (e.g. PA1874-1877, MexEF-OprN, MexAB-OprN, MexCD-OrnD) directly transport antibiotics from the cytoplasm to the cell exterior. ß-(1→3)-glucans act as chaperones for intercellular communication promoting antibiotic resistant gene expression in neighbouring cells. The small colony variant (SCV) phenotype is a phase variant associated slow growth, enhanced capacity for biofilm formation via hyperpiliation and increased antibiotic resistance.

The ways in which biofilms enhance ABR are manifold, complementary and overlapping (see Figure 1). The extracellular matrix has been shown to cause protection from the antibiotic tobramycin, but not other antibiotics such as ciprofloxacin, through diffusion/penetration limitation[8]. In particular, this example has been linked to the presence of negatively charged molecules such as extracellular DNA being sequestered within the biofilm and interfering with the diffusion of the positively charged tobramycin as this barrier was overcome when growth media was supplemented with cations. Exopolysaccharides, polysaccharide molecules exported into the extracellular matrix, are another key determinant of increased antibiotic resistance against a variety of antibiotics[9] and will be a key topic in this thesis. A precise mechanism of action for the sequestering of antibiotics by polysaccharides has yet to be determined, although the *P. aeruginosa* exopolysaccharide Pel has been shown to control the localisation of extracellular DNA within the biofilm[10] and so may be involved in the previously identified control of tobramycin. The condensed and localised form of a biofilm allows for the secretion of β-lactamase enzymes, which can degrade antimicrobials and prevent them from reaching their target, and their maintenance within the local microenvironment of the extracellular matrix. Chromosomally encoded AmpC is a clinically relevant determinant of resistance to β-lactam antibiotics in *P. aeruginosa* and has been shown to minimise penetration of the biofilm by such antibiotics when treated at low concentrations[11].

Extracellular DNA, in addition to its previously mentioned role in diffusion limitation of cationic antibiotics, is implicated in driving gene expression of resistant determinants by decreasing the availability of Mg^{2+} through chelation and by generating acidic micro-environments. These signals activate the two-component PmrA-PmrB and PhoP-PhoQ systems which produce redundant responses that increase ABR to cationic antimicrobial peptides and polymyxin B[12,13] via the addition of an aminoarabinose to the lipid A of lipopolysaccharide (LPS) and the production of spermidine[14]. These responses serve to increase the stability of the outer membrane by reducing its permeability to cationic antimicrobials[15]. Additionally, extracellular DNA can be generated by a wide variety of sources and its concentration in biofilms promotes the horizontal gene transfer of potential resistance genes[16].

Antimicrobials are generally more effective against fast-growing bacteria rather than those in a stationary phase of growth[17] due to the need to inhibit an active process such as cell wall synthesis or ribosomal translation. Biofilms therefore pose a challenge to effective clearing of bacteria as the structure of the colony generates a gradient of nutrient and oxygen availability between the exposed edges of the colony and the enclosed interior[18] that generates a more nutrient-deficient, hypoxic environment that favours a reduced growth rate[19]. Furthermore, anaerobic growth environments have been found to support higher antimicrobial tolerance[20]. Some evidence suggests that this may be due to the upregulation of multidrug efflux transporters MexEF-OprN[21] and MexCD-OprD[22], which are

involved in the removal of antimicrobial compounds from the cell, during hypoxia. Another proposal is that some antibiotics, such as ciprofloxacin, also enhance the metabolism of the citric acid cycle and that the increased production of reactive oxygen species contributes towards cell death; if this is the case then cells in a hypoxic environment would be protected due to decreased availability of oxygen. Evidence of increase levels of reactive oxygen species in response to antibiotic treatment has been shown[23], although a direct link to increased metabolism remains lacking.

The heterogenous biofilm environment encourages subsets of cells in more limiting conditions to reduce their metabolism. As previously mentioned, such cells are more resistant to the subversion of cellular processes by antibiotics and are termed "persister" cells. In addition to dormant cells, persister cells may also be cells under the effects of toxin-antitoxin modules which protect the cell from antibiotics by preventing or limiting the activity of the cellular components that the antibiotics act upon[24]. Persister cells are estimated to make up to 1% of cell populations[24] and their enhanced ABR accounts for the recalcitrance of infections to a significant extent in biofilms[25].

The NdvB protein is a membrane glucosyltransferase responsible for the production of cyclic β-(1→3)-glucans that are localised in the periplasmic and extracellular spaces[26,27]. The glucans produced are composed of 12 to 15 glucose units with a 50% incidence of anionic 1-phosphoglycercol substitution at the *O*-6 position, which would be able to sequester positively charged antibiotics outside of the cell[27]. The size of the glycan produced suggests that diffusion limitation is unlikely to be its sole purpose, given that smaller molecules would also be effective at such at task at a lower cost to the cell, and indeed, it has been proposed that NdvB glycans act as chaperones that support the hydrophobic core of intercellular signalling molecules[28]. This reveals another mechanism by which biofilms enhance ABR; intercellular signalling allows signal transduction throughout the colony leading to changes in gene expression that support antibiotic readiness. In the case of NdvB, signalling has been shown to increase alcohol oxidation[28] which increased ABR through an unconfirmed mechanism, possibly relating to the management of reactive oxygen species[9]. A ΔndvB biofilm is 8-16 times more susceptible to tobramycin[26] while deletion of the downstream alcohol oxidation pathway led to only a 4 times more susceptible phenotype[28], highlighting that the NdvB protein product is multifunctional and may contribute to ABR in ways yet to be identified.

The BrlR (biofilm resistance locus regulator) protein is a Mer-like transcriptional activator of multiple biofilm-related ABR processes[29]. Upregulation of the *ndvB*, the two-component PmrA-PmrB system, the ABC transporter operon PA1874-1877, and the multidrug efflux transporters *mexAB-oprM* and *mexEF-oprN* is caused in direct response to this protein and unsurprisingly correlates with increased ABR[29–32]. BrlR is upregulated in biofilms but not in planktonic cells[29]. The binding of BrlR is enhanced in the presence of cyclic-di-GMP, a secondary messenger associated with biofilm

formation and maintenance[33], and the *brlR* promoter region contains a BrlR binding site which allows autoregulation and the generation of a positive feedback loop[34]. Interestingly, BrlR also functions as a transcriptional repressor of the PhoPQ two-component system which is important in colistin resistance[31], highlighting selective interplay between resistance mechanisms.

RapA is a SWI1/SNF2 family helicase-like protein which activates RNA transcription by stimulating RNA polymerase recycling[35]. This protein has been found to improve resistance to penicillin G, norfloxacin, chloramphenicol and gentamicin in biofilms but not planktonic cells[26,36]. In *E.coli*, a ΔrapA strain was found to have reduced expression of 22 proteins in comparison to the wild type, including two putative multidrug efflux transporters *yhcQ* and *yejG*, and to produce a smaller portion of exopolysaccharides[36]. This suggests that RNA transcription activation by RapA is selective towards the transcription of these products and that the combined action of the efflux transporters and exopolysaccharides are responsible for ABR.

P. aeruginosa is well known for having numerous redundant multidrug efflux pumps which contribute to ABR by removing antibiotics from the cell. There are five major families of efflux transporter, classified based upon sequence, the energy source used to facilitate transport and the substrate transported[37]. The five classes are ATP-binding cassette (ABC), small multi-drug resistance (SMR), major facilitator family (MFS), resistance-nodule-cell division (RND) and multidrug and toxic compound extrusion (MATE) families[38]. ABC transporters utilise ATP hydrolysis for active transport while the other transporters utilise energy from hydrogen or sodium gradients[39]. Many of these transporters are active in both the planktonic and biofilm states[40]. Others, such as the PA1874-1877 operon, MexAB-OprM, MexCD-OprD and MexEF-OprN, are preferentially expressed in the biofilm[22,30,41], either under direct control of a biofilm regulator or as a stress response to the limiting conditions within a biofilm[30,32,42]. PA1874-1877 belongs to the ABC transporter family[32] while Mex(-Opr) proteins belong to the RND family[37]. Both of these families operate as tripartite transporters including inner and outer membrane components linked by a periplasmic component[38], as is required to transport from the cytoplasm to cell exterior.

The biofilm lifestyle encourages a state of hypermutation which correlate with increased oxidative stress and deficiencies in DNA oxidative repair systems[43]. This increased level of mutation is self-perpetuating, as accumulated mutations further limit mechanisms preventing mutation, and is associated with an enhanced adaptive response to antibiotics[44]. This is the unsurprising result of applying a strong selection pressure to a strain with greater genetic diversity. In fact, adaptation of naturally soil-dwelling *P. aeruginosa* to allow chronic infection of the lungs in patients with cystic fibrosis is routinely associated with a mutation causing the overexpression of the biofilm exopolysaccharide alginate and generating a mucoid phenotype that is resistant to both antibiotics and the immune system[45].

The process of phase variation in bacterial colonies is the adoption of heterogenous phenotypes throughout the colony to allow rapid adaptation to change in environmental conditions. The mucoid phenotype can be considered an example of this as it is unnecessary for the entire colony to be overexpressing alginate to benefit from its protective effect but it remains beneficial for a subpopulation to be engaging in gene expression that would provide an adaptive advantage where the mucoid phenotype fails to offer full protection. An example of this is the small colony variants (SCV) of *P. aeruginosa* that show slow growth, enhanced capacity for biofilm formation via hyperpiliation and increased antibiotic resistance that correlate with poor lung function and inhaled antibiotic therapy[46–48]. Having such a subpopulation provides a selective advantage as it provides a means of rapid expansion following antibiotic challenge to re-establish the protective effects of the biofilm and then re-diversify into the heterogenous subpopulations that previously provided a protective effect, such as the mucoid variant. The process of conversion between SCV type and wild type cells is reversible and under control of the EAL domain containing phosphodiesterase phenotype variant regulator (PvrR) which breaks down the biofilm-accumulating secondary messenger c-di-GMP[49] and shifts equilibrium in favour of the wild type[50]. GGDEF domain containing protein such as PA1120 and MorA, which also contains a hydrolytic EAL domain, contribute towards enhancement of c-di-GMP signals and are necessary for formation of the SCV phenotype[49,51]. The SCV phenotype is also inhibited by the GacS/GacA two component system[52] which is associated with switch from acute to chronic infections[53].

1.1.3 *Pseudomonas aeruginosa* and pathology

Pseudomonas aeruginosa is a Gram negative rod-shaped facultative anaerobic bacterium with clinical relevance as a pathogen of both plants and animals. In humans, the bacterium is associated with opportunistic infections, otitis media, chronic bacterial prostatitis, cystic fibrosis and nosocomial infections (from inserted medical devices), including catheters and contact lenses[54]. *P. aeruginosa* has multiple modes of infection in man, varying from acute to chronic, local to systemic and benign to life threatening. In recent decades it has become one of the most common causative agents of high morbidity and mortality nosocomial infections due to its extensive ABR[55]. Infections associated with pneumonia, sepsis, cystic fibrosis and chronic obstructive pulmonary disease result in high morbidity and bleak prognosis[56]. Even infections that are not life-threatening often have a severe impact upon quality of life, such as chronic pneumonia in cystic fibrosis or impaired eyesight related to keratitis of corneal infection[57].

On infection, the pathogen secretes a wide variety of virulence infectors that are responsible for damage to the host. Secreted proteases disrupt a wide variety of host processes relating to immune regulation, such as cell surface receptors, chemokines and cytokines, to prevent an effective immune response[58–62]. Also targeted are host extracellular matrix components and clotting factors, leading to haemorrhagic lesions and tissue necrosis[63–66]. Cell lysis or invasion is also facilitated via numerous

secreted toxins with various modes of action, including disruptions to the membrane, protein synthesis and trafficking as well as subversion of host apoptotic pathways[67].

The severity of disease caused by *P. aeruginosa* varies from mild to life threatening depending upon circumstances and the effectiveness of treatment; nosocomial infections can most frequently be treated by removing the device but tissue infections rely upon antibiotics[68]. One of the greatest challenges in treating this pathogen is emerging antibiotic resistance, for which it has been classified as a critical target for the development of new antibiotics[69], that is enhanced through the formation of complex three dimensional colonies known as biofilms[70], the disruption of which has been shown to limit antibiotic resistance and enhance killing by immune cells[71,72]. The originators of biofilms in *P.aeruginosa* are therefore attractive targets for development of novel therapies for the bacteria.

1.1.4 Biofilms in *Pseudomonas aeruginosa*

A *P. aeruginosa* biofilm is formed of heterogeneous microcolonies that band together in dynamic mushroom-shaped towers (see Figure 2). This morphology allows varying water and nutrient access to the different microcolonies, creating a heterogeneous metabolic environment, and aids their dissemination throughout the host should the 'stalk' break under the combined sheer stress of the environment and mass of the 'cap'. Rather than homologous layers of cells, a biofilm's cell content is a mere 15% by volume, with the remaining 85% of the colony volume being occupied by extracellular matrix (ECM)[54] composed of a complex mixture of proteins, nucleic acids and lipids enmeshed on a bed of polysaccharides. This modifies the local environmental conditions to be more favourable and interferes with the diffusion of complex molecules such as antibodies and larger antibiotics[73] by providing alternative binding molecules such as extracellular DNA[68]. The ECM also functions as an extracellular digestive system for the colony as secreted enzymes involved in nutrient acquisition, or the breakdown of antibiotics, can similarly be locally concentrated by limited diffusion potential, as well as modified and stabilised[73], by ECM components which also function as a web for nutrient capture through electrostatic potentials. The heterogeneous availability of water, oxygen and nutrients that this causes generates an equally heterogeneous metabolic environment that causes cells on the edge to grow rapidly, encouraging pathology, while internalised cells are forced into a state of low metabolism, known as persister cells[68], that has protective effects against antibiotics that target processes involved in cell growth such as cell wall synthesis[74].

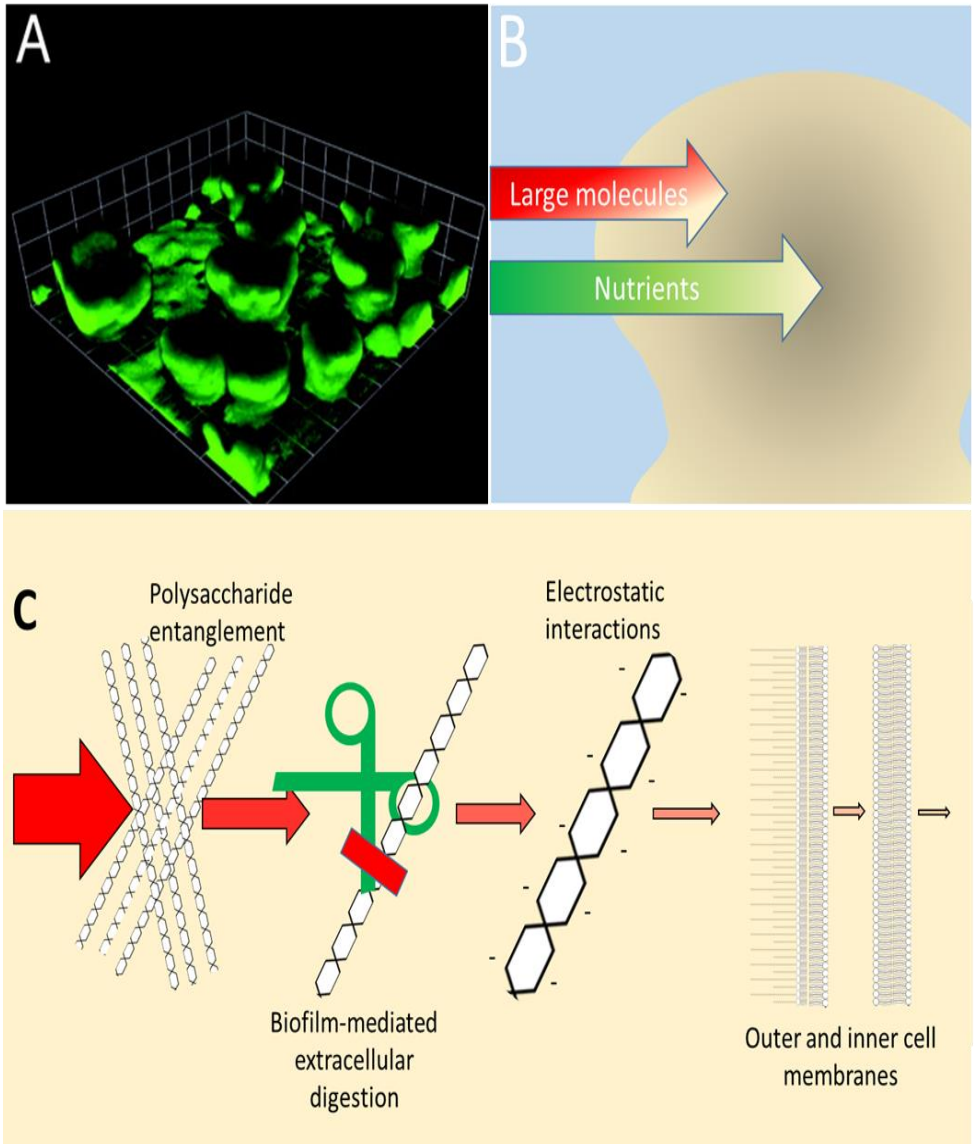

Figure 2: The importance of biofilms in Pseudomonas aeruginosa. ***A: Confocal laser microscopy image*** *showing the archetypical mushroom-oid towers of a mature Pseudomonas aeruginosa biofilm (Figure from Donlan, R.M. & Costerton, J.W. 2002).* ***B: Diffusion gradients from concentrated (dark) to dilute (light) into a biofilm****. The diffusion of large molecules such as antibiotics and antibodies is more limited than smaller molecules such as water, oxygen and metals; this effectively decreases the concentration of antimicrobials reaching cells which decreases their efficiency and encourages adaptation. The lower nutrient concentrations reaching the centre of the biofilm is sufficient to maintain the bacteria but is limited enough that metabolisms are slowed, and cells "hibernate". These persister cells can "reawaken" in response to increased nutrient access, such as following damage or clearance of the colony, and provide a new source of infection.* ***C: Challenges posed by biofilms to antimicrobials****. Biofilm exopolysaccharides form gelatinous meshes that trap or hinder the movement of large molecules, such as T-cells and antibodies, limiting antigen accessibility and impairing adaptive immunity. Exopolysaccharide-bound enzymes serve as a communal extracellular digestive system providing both nutrients for the cell and forming a defensive perimeter of antibiotic resistance enzymes. Negative charges of exopolysaccharides serve to control flow of polar molecules such as by chelating toxic heavy metals and repelling nucleic acids backbones. The outer and inner cell membranes of Gram-negative bacteria form a set of densely hydrophobic barriers requiring molecules to undergo selective transport to enter the periplasm and cytoplasm*

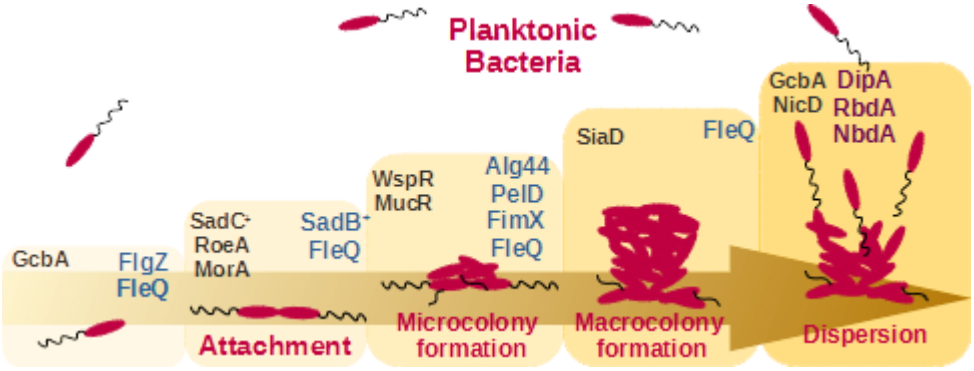

Figure 3: Illustration of the stages of biofilm formation. The process of biofilm formation is cyclical and begins when a planktonic bacterium encounters a surface to which they attach lengthways. The bacterium continues to move along the surface, secreting an extracellular matrix of polysaccharides which attract other bacteria to the surface. As more bacterium join together, they form a microcolony which then begins to expand in the vertical axis to form a microcolony. The final stage of the cycle is the dispersal of the biofilm, with bacteria exiting into the planktonic state. This process is mediated by diguanylate cyclases (black) and phosphodiesterases (purple) which control the availability of cyclic-di-GMP, increasing and decreasing respectively. Cellular effectors (blue) respond to these signals to drive the cycle. Dominant diguanylate cyclases, phosphodiesterases and effectors are labelled for each stage of development.

Biofilm formation is a multi-step process that begins when a planktonic bacterium comes into contact with a solid surface (see Figure 3). The bacterium attaches to the surface along its long axis and begins to the process of forming a monolayer by cell division and adhering to neighbouring cells via a secreted extracellular matrix of exopolysaccharides, lipids, proteins and scavenged extracellular DNA. As the biofilm matures it grows in distinctive environmentally determined shapes that maximise nutrient and oxygen distribution throughout the colony. Once a critical mass is reached, or in response to environmental stimuli, such as shear stress of carbohydrate starvation, part of the biofilm will break down its surrounding extracellular matrix and disperse as planktonic bacteria to find new sites to colonise.

1.1.5 Biofilm regulation in *Pseudomonas aeruginosa*

Biofilm formation is primarily under the control of the secondary messenger cyclic-di-GMP (see Figure 4), with higher levels of the compound encouraging biofilm formation and growth while lower level encourage the adopting of a planktonic phenotype[75]. In *P. aeruginosa*, planktonic cells are estimated to contain less than 30 pmol of c-di-GMP per mg of total cell extract whereas this level is elevated to 75-110 pmol.mg^{-1} in biofilms[76]. The molecule is synthesised by diguanylate cyclases (DGCs) from two GTP molecules and degraded by phosphodiesterases (PDEs) which contain GGDEF and EAL/HD-GYP domains, respectively[77,78]. At least six DGCs have been identified as controlling the transition from planktonic to surface-associated phenotypes: MucR, WspR, SadC, RoeA, SiaD and YfiN/TpbB[75]. Conversely, two DGCs, GcbA and NicD, and three PDEs, DipA, RbdA and NbdA, have been associated with biofilm dispersal[75]. While global intracellular c-di-GMP levels are generally those measured, there is evidence that effectors respond instead to local c-di-GMP concentration[79], such as been observed for the requirement of the SadC enzyme for Pel

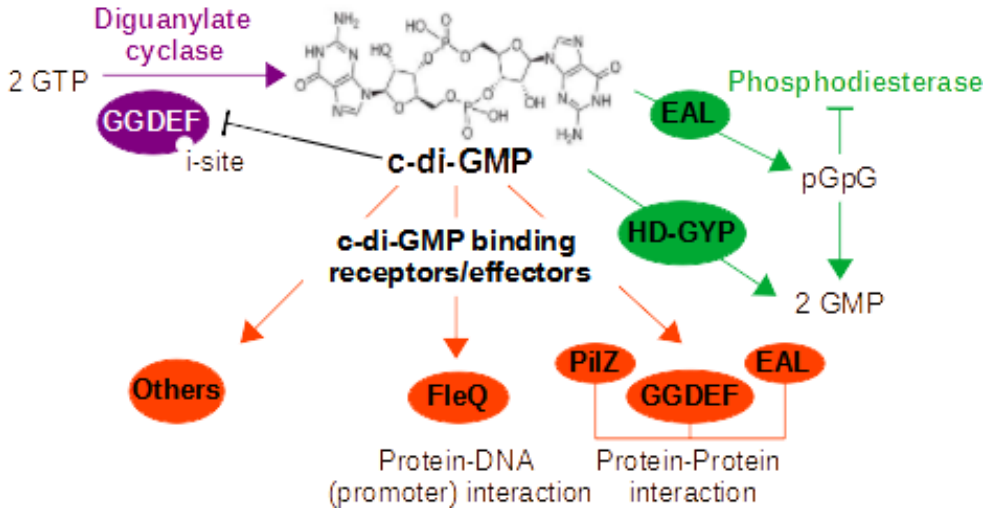

Figure 4 : Cyclic-di-GMP mediation in regulation of biofilm formation**.** Cyclic-di-GMP (c-di-GMP) is synthesised by diguanylate cyclases, containing a GGDEF motif and c-di-GMP recognising allosteric I-site, from two GTP molecules. Phosphodiesterases, baring EAL or HD-GYP domains, hydrolyse c-di-GMP into pGpG or two molecules of GMP, respectively. Diguanylate cyclases and EAL-domain containing phosphodiesterases are inhibited by their products. Increased levels of c-di-GMP enhance gene expression via the promoter FleQ and stimulate cellular effectors via proteins containing PilZ or degenerate GGDEF or EAL domain containing proteins.

production to occur even at comparable levels of global c-di-GMP[80] and the requirement of the GGDEF and EAL domain containing MucR for alginate synthesis[81,82].

Exopolysaccharide synthesis is under control of the control of c-di-GMP, both directly and indirectly, and a variety of two-component regulators. The Pel operon is activated by the IM protein PelD, which contains a degenerated GGDEF domain and I-site, whose activity occurs in response to c-di-GMP binding[83]. Likewise, alginate synthesis is dependent upon the Alg44 receptor which transduces signal from c-di-GMP binding via a PilZ domain[84]. Alternatively, FleQ is a c-di-GMP-responsive enhancer of gene expression which improves expression of flagellum-related genes and represses genes involved in adhesion and exopolysaccharide synthesis, such as *pel* and *psl*, at low levels of c-di-GMP but has the inverse effect in high c-di-GMP conditions following a conformational change[85].

Two-component system (TCS) regulators also affect gene transcription as both *pel* and *psl* are indirectly activated by the GacS/GacA system[86] (see Figure 5). Activation of GacA leads to the transcription of two small regulatory RNAs, RsmY and RsmZ, which inhibits RsmA which would then typically prevent ribosomal translation of the exopolysaccharide genes[86]. Each step of this process is under further regulatory cross-talk as GacS is inhibited by RetS while RsmYZ is enhanced by HptB and the TCS BfiS/BfiR and GacS is activated by LadS[86]. The ligands that activate these systems have yet to be identified[87], although a crystal structure of RetS identified a

potential carbohydrate binding module[88] which suggests the potential integration of carbohydrate catabolism/metabolism into regulation.

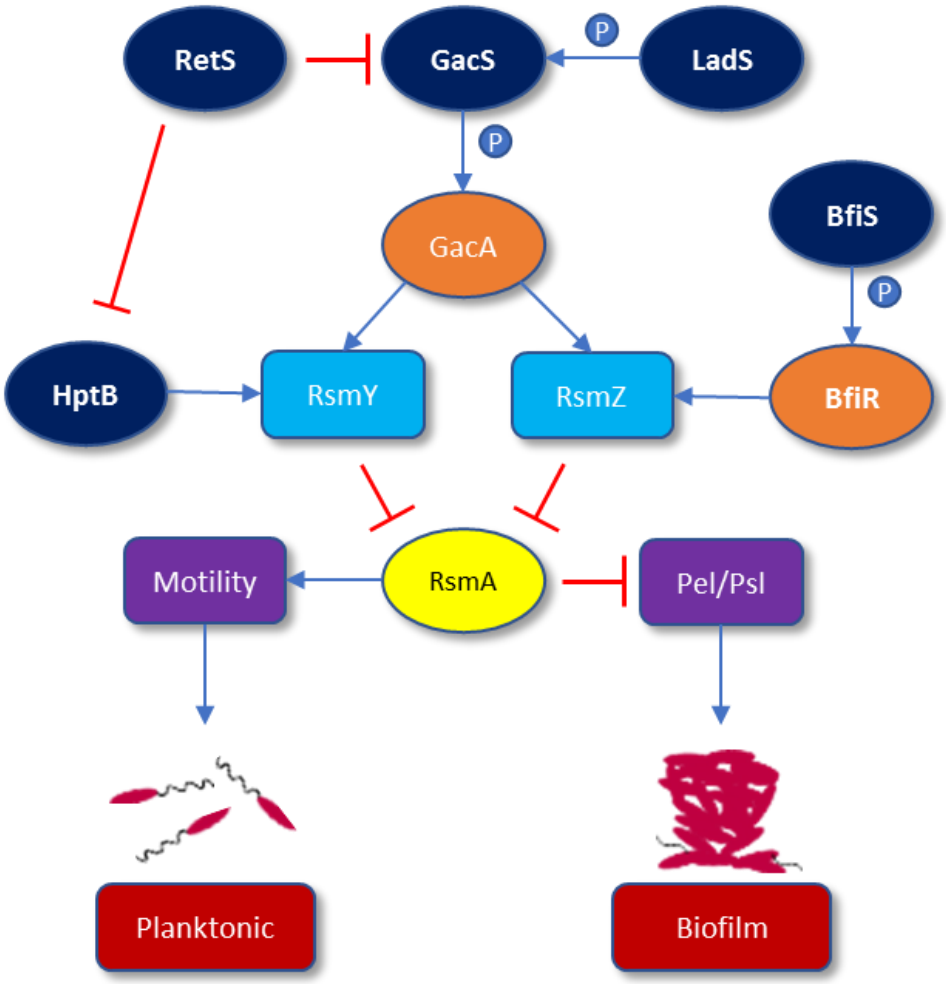

Figure 5: Signalling interplay in biofilm formation. GacS is the sensor histidine kinase component of the GacS/GacA two component system. Activation of the regulatory component GacA is caused by phosphorylation by GacS, itself regulated by sensor signalling kinases RetS and LadS which inhibit or enhance activity through dephosphorylation or phosphorylation, respectively[87]. Active GacA leads to increased transcription of the small regulatory RNAs RsmY and RsmZ. RsmYZ are also enhanced by signals from the DGC HptB[89] and by repression of RNase G transcription by the BfiS/R two component regulator[90]. RsmYZ in turn inhibits RsmA by sequestering it. RsmA is a protein which inhibits biofilm formation by post-transcriptional repression of pel and psl genes; activation of RsmY/RsmZ therefore drives biofilm formation. Arrows represent activation, "T" represents inhibition and "P" shows phosphorylation.

The attachment process is mediated in part by flagellum-based surface sensing. When a bacterium comes into contact with a solid surface this increases the flagellar load and induces an interaction between the GTPase FlhF and pole organiser FimV[91]. This signal leads to the activation of CyaA

and CyaB which in turn synthesise the secondary messenger cyclic-AMP[91]. The virulence transcription factor Vfr becomes active when bound to cyclic-AMP and represses the flagellum synthesis regulator FleQ[92] and activates the expression of Type IV Pilli[93–95]. Type IV Pilli act as a mechanosensor to confirm the presence of a solid surface which activates the Chp system to stimulate CyaB and further elevate cellular cyclic-AMP[96,97]. Vfr-cyclic-AMP also activates biofilm-associated virulence factors[93,95,98]. Vfr can also bind c-di-GMP and is inactive when it does so[94], suggesting that c-di-GMP may act as a competitive inhibitor and providing a mechanism for the shift from attachment to maturation[99].

Transition from the attachment phase to the microcolony phase of biofilm formation is dependent upon exopolysaccharide production. There are two major clinically relevant strains of *Pseudomonas aeruginosa* used in research: PA01 and PA14. PA01 contains both the *pel* and *psl* operons and it has been shown that Δ*psl* is unable to progress beyond the formation of a monolayer[100]. PA14 contains *pel* but not *psl* and Δ*pel* PA14 shows a similar inability to form a three-dimensional colony[100]. It has been shown that attached PA01 leaves a trail of Psl behind it as it migrates which then serves as a scaffold and attractant for other PA01 bacterium[101] and a similar mechanism may apply for Pel. Extracellular DNA is also important at this stage of development as treatment with DNase I has been shown to arrest colony formation but is ineffective at disrupting a mature biofilm[102]. It has been shown that eDNA contributes towards cell-cell adhesion[103] and that Pel is able to mediate its coordination[10,104]. Psl has a further impact upon maturation as the presence of Psl stimulates c-di-GMP production via SadC and SiaD[105], producing a positive feedback loop that raises and maintains levels at the higher concentrations associated with a mature biofilm.

Biofilm dispersal is also a c-di-GMP dependent process, although it focuses upon activation of PDEs to lower levels rather than DGCs to enhance them. Nitrous oxide and carbon substrates have been identified as signals the precipitate dispersal in a PDE-dependent manner[106,107]. Increased availability of carbon sources or nitrous oxide has been linked to increased DGC activity of DipA and RbdA[106,108], implying a central pathway in dispersal regulation. The protein BdlA has been identified as a downstream regulator necessary for dispersal[109]. BdlA contain two putative PAS domains between which the protein undergoes proteolysis by ClpP following a c-di-GMP dependent phosphorylation step, therefore a lowering of c-di-GMP by DGCs such as DipA promotes dispersal by preventing proteolysis of BdlA[109,110].

1.1.6 Biofilm exopolysaccharides in *Pseudomonas aeruginosa*

In *P. aeruginosa*, three operons are responsible for the production of the polysaccharide chains that hold the ECM together: Alg, Psl and Pel (see Figure 6). All three operons are under the control of the secondary messenger bis-(3′–5′)-cyclic dimeric guanosine monophosphate (c-di-GMP), Alg and Pel at the post-translational level[111–113] and Psl at the transcriptional[114]. The Alg operon is involved in the synthesis and secretion

of alginate which has been shown to be the determining factor in forming the persistent mucoid phenotype associated with cystic fibrosis; this has led to intensive study and characterisation of this pathway. Alginate is believed to be of little importance for the bacteria outside of this specialised environment. Psl and Pel act as biofilm determinants with similar, if not redundant, roles both in general infectious and non-infectious environments by controlling cell-cell interactions; while both operons are conserved they are rarely expressed simultaneously. The Psl operon was first identified using a reverse genetics approach whereas the Pel operon was determined as the source of pellicle type biofilms that form at the air-liquid surface in standing cultures[115]; the morphology of this colony type is wrinkled which affords greater access to limiting factors such as oxygen, water and nutrients by providing a larger surface area to volume ratio[116]. Of these operons, *pel* will be the focus of the research described here, however, a brief description of how the other, better characterised, operons function is necessary to place the work in the appropriate biological context.

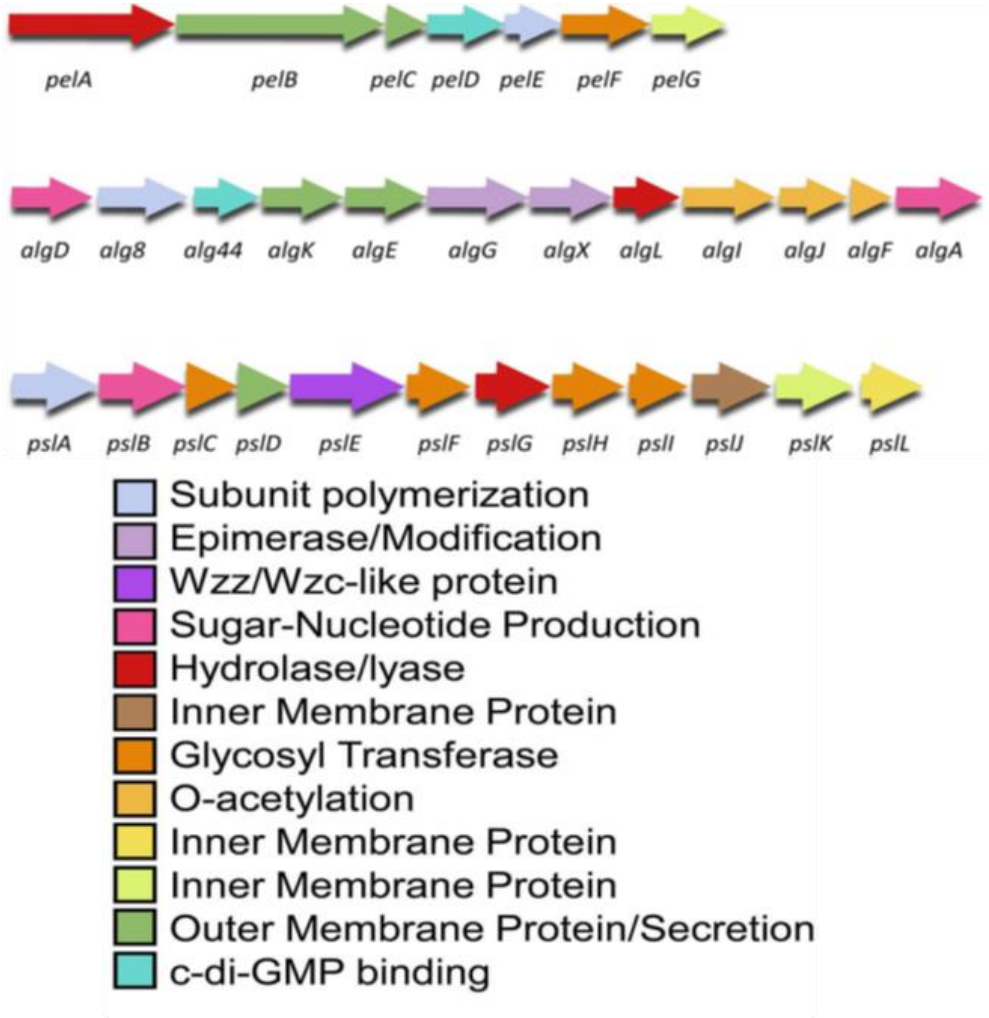

Figure 6 : Organisation and content of the Pel operon of Pseudomonas aeruginosa. ***A: Biofilm exopolysaccharide operons*** *with functional attributions based upon bioinformatics modelling (pel, psl) and structural and functional studies (alg). Figure adapted from Franklin, M.J. et al, 2011.*

Alg synthesis involves 12 proteins[117]: three cytoplasmic biosynthetic enzymes (AlgACD) which generate a linear homopolymer of D-mannuronic acid, two inner membrane proteins (AlgJI) and a periplasmic protein (AlgF) that perform selective O-acetylation of the polymer following translocation across the inner membrane (Alg44). Biosynthesis and transport are regulated by Alg44 in response to the secondary messenger c-di-GMP through a cytoplasmic PilZ domain in combination with Alg8[82]. A multiprotein scaffold is formed by a periplasmic lipoprotein (AlgK) through protein interfacing tetratripeptide-like (TPR) domains which disables the polysaccharide degrading activity of AlgL and assists polysaccharide epimerisation (AlgG), acetylation (AlgX[118]) and finally outer membrane transport through the β-barrel of AlgE. The resulting negatively charged exopolysaccharide has a five sugar repeating unit of 1-4 linked D-mannuronic acid with subunits 2 and 3 being selectively acetylated while subunit 4 is epimerised to L-guluronic acid; alginate is secreted from, and not covalently bonded to, the cell.

There are also twelve components in the Psl operon[117], which also relies upon the glucose-1→6 phosphate isomerase activity of AlgC to make two of its substrates. Unlike Alg, and the proposed mechanism for Pel, Psl is believed to make use of an isoprenoid-lipid based biosynthesis and transport due to the lack of a TPR scaffold and potential structural similarity of PslA, PslD and PslE to several proteins from the *E.coli* extracellular and capsular polysaccharides (EPS/CPS), WbpA, Wzc and Wza respectively, which also use this mechanism. A combination of PslB, AlgC, GalU and RmLC are responsible for generating the nucleotide activated sugar precursors of the Psl pathway. PslB is not essential to this pathway as WbpB shares a redundant role in GDP-mannose production. Three proteins, PslF, PslH and PslI, are predicted to belong to glycosyltransferase (GT) family 4 while a fourth protein, PslC, is believed to belong to GT family 2. The principle difference between these GT families is that they adopt different folds that cause GT 2 enzymes to use an inverted mechanism while GT 4 enzymes retain[119]. These four GT enzymes are predicted to be involved in cytoplasmic biosynthesis of the Psl repeating unit. PslA, PslE, PslJ, PslK and PslL contain inner membrane domains and are proposed to form a polymerisation complex for assembly onto an isoprenoid lipid which is then flipped across the membrane by PslK. PslD, potentially an octameric lipoprotein forming multiple rings, is involved in trans-periplasmic transport and outer membrane export. PslG contains a glycoside hydrolase domain which may be involved in polysaccharide processing but is suggested to break down aberrant polymer in a manner similar to AlgL. Indirect evidence for this is based upon the mutants of homologues Wza/Wzc, which produce short oligosaccharides rather than the expected polysaccharide. The Psl polysaccharide is a repeating pentamer formed from D-mannose, L-rhamnose and D-glucose that forms a helical coating around the cell surface.

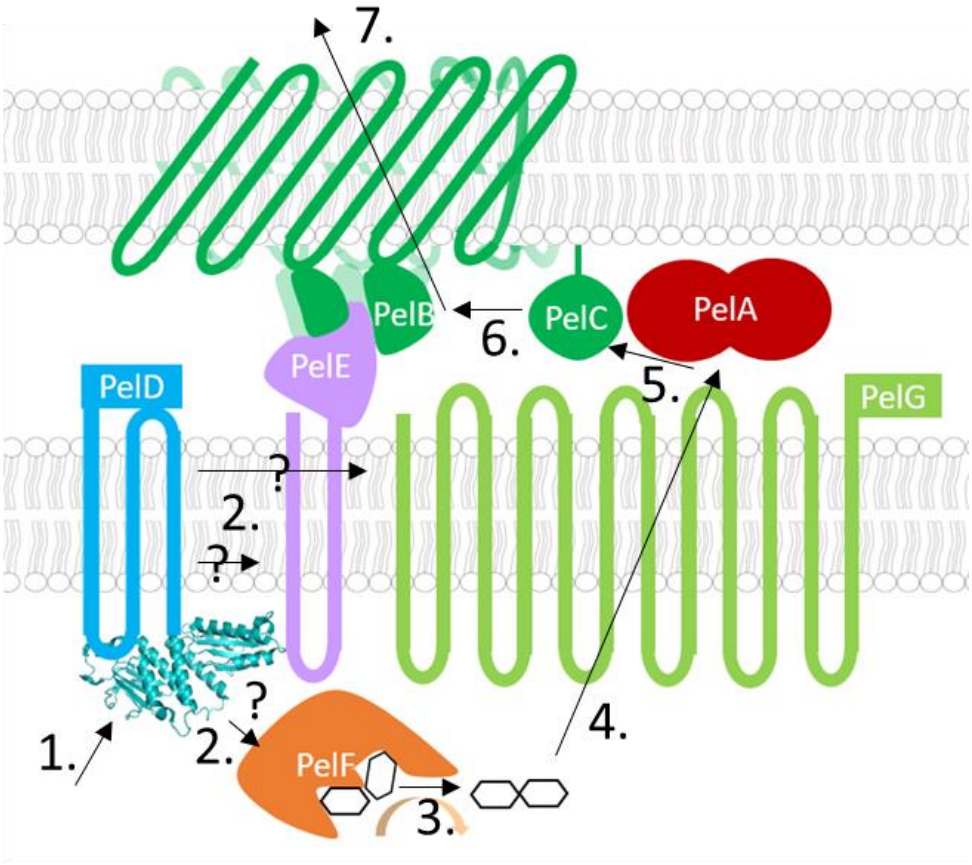

Figure 7 : Working model of functionality of the Pel operon. **1.** PelD responds to c-di-GMP binding to the I-site of its periplasmic domain. **2.** PelD activates the Pel system in response to c-di-GMP through a GAF domain. The target for this response could be PelE, PelF and/or PelG with PelF being the most likely. **3.** PelF catalyses the condensation of UDP-glucose and UDP-N-acetylgalactosamine sugars through a retaining glycosyltransferase reaction to form a polysaccharide. **4.** PelG transports condensed sugars across the inner membrane using electrochemical gradients. **5.** PelA is involved in catalytic processing of the condensed sugars and possibly regulation via degradation; it has an unusual combination of a membrane associated glycoside hydrolase 114 family TIM barrel predicted to catalyse hydrolysis of the polysaccharide and a possible C-terminal β/α barrel involved in (de)acetylation of the Pel polysaccharide at N-acetyl units. **6.** PelC is an outer membrane lipoprotein involved as an accessory to outer membrane transport. **7.** PelE is a scaffold protein supporting protein-protein interactions with a solvent accessibility-related armadillo-like helical domain and tetratripeptide-like (TPR) protein interaction domain. PelE is likely responsible for the transperiplasmic localisation of PelG, PelA and PelB. PelB is a large outer membrane β-barrel transporter with sixteen TPR domains. The end product is a partially acetylated poly(1-4)galactosamine-glucosamine exopolysaccharide.

In contrast, the Pel operon comprises only seven proteins, PelA-G, all of which are essential for functional Pel to be produced (see Figure 7). The

genes of this operon are under control of the FleQ repressor/promoter and located on the antisense strand of the chromosome. The AlgC enzyme catalysing the conversion of glucose-6-phosphate to glucose-1-phosphate has also been shown to be essential to Pel production[120] and is believed to be an important upstream determinant which integrates the three biofilm polysaccharide pathways. While functional studies of the *pel* operon are lacking, bioinformatics analysis suggests that Pel production is regulated in response to c-di-GMP by PelD[83], initial polymer/repeat unit biosynthesis occurs in the cytoplasm by PelF, transport across the inner membrane is first performed by PelG and then across a transperiplasmic complex by PelABCE. Further catalytic activity by PelA occurs before translocation across the outer membrane by PelBC. The precise structure of the Pel polysaccharide remains undetermined but appears to be composed of 1-4 linked partially acetylated galactosamine and glucosamine; the exact nature of the acetylation remains unknown[10]. This gives the polysaccharide an unusual positive charge which would allow it to interact with anionic polymers such as extracellular DNA (eDNA), with which it co-localises at the biofilm stalk, and the mammalian extracellular matrix component hyaluronan[10]. *P. aeruginosa* has also been shown to secrete DNases which were primarily related to the cannibalisation of lysed cells but which may also have a role in dynamic biofilm remodelling[121] and eDNA has been shown to be an important virulence factor in biofilms of uropathogenic *E. coli*[122].

1.1.7 Proteins of the *pel* operon

The biochemical and functional characterisation of the *pel* operon proteins will be discussed here, alongside predictions based upon the sequences of the proteins (see Figure 8). The proteins will be discussed sequentially in the predicted order of activity within the biosynthetic pathway of PelDFGEABC.

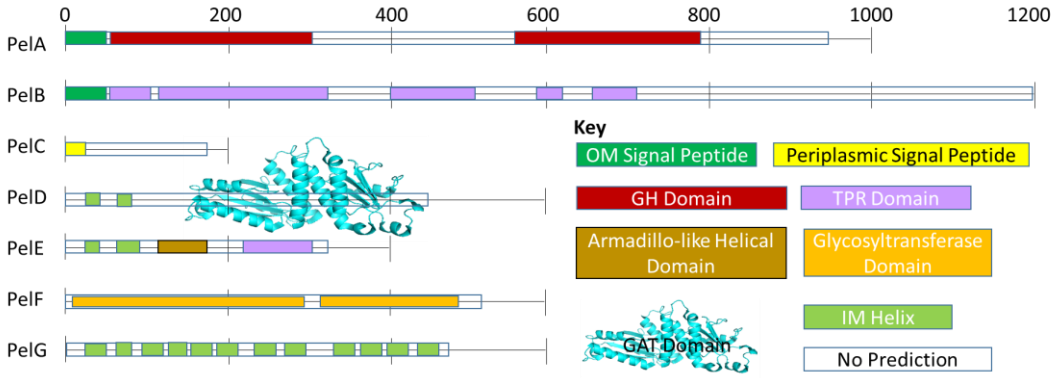

Figure 8: Domain prediction of Pel proteins based upon amino acid sequence analysis. The scale represents number of amino acids into the sequence with a major unit of 200 residues. Coloured bars show predicted domain boundaries within the proteins, with the pre-determined cytoplasmic domain of PelD represented as a cyan ribbon diagram. White bars show regions of the protein for which the amino acid sequence lack sufficient similarity to predictively assign a domain. Outer membrane signal peptides

are involved in transport recognition to the outer membrane, periplasmic signal peptides likewise localise the protein to the periplasm. Glycoside hydrolase (GH) domains are involved in the hydrolysis or modification of glycosidic bonds, such as those linking individual sugar subunits of the pel polysaccharide. Tetratricopeptide repeat (TPR) and armadillo-like helical domains are broad-surface solvent exposed domains involved in mediating protein: protein interactions. Glycosyltransferase domains are involved in the recognition of acceptor and donor sugars and work in concert to mediate condensation of the two, in this case polymerising glycosidic bonds. The GAF/GGDEF domain of PelD is a c-di-GMP responsive domain involved in activation of the pel synthetic system. Inner membrane (IM) helices are a readily identifiable hydrophobic membrane-spanning region of the protein that are found in IM proteins but not those in the outer membrane.

1.1.7.1 The Secondary Messenger Sensor: PelD

The first protein in the Pel pathway is PelD, which acts as a sensor for c-di-GMP[83]. This is an inner membrane (IM) protein with four predicted

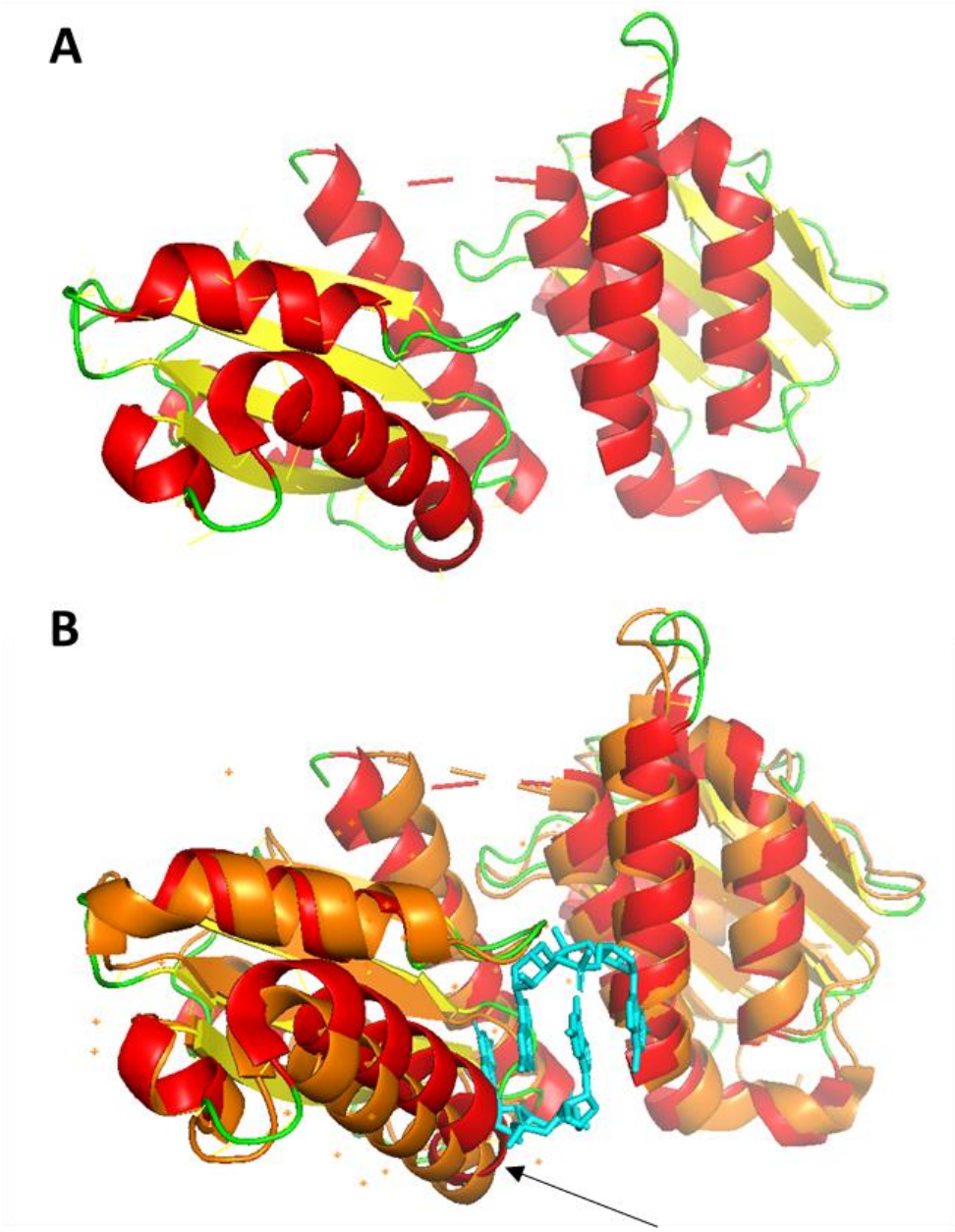

Figure 9: The structure of the cytoplasmic tail of PelD. ***A: The apo structure of PelD 156-C terminus.*** *The N-terminal GAF domain (right) is a common feature of cyclic nucleotide receptors. The C-terminal GGDEF domain is degenerate, lacking the ability to act as a diguanylate cyclase, but maintains a c-di-GMP-binding I-site. Cartoon representation showing alpha helices (red), β-sheets (yellow) and loops (green). PDB: 4DMZ*[113]. ***B: The structure of PelD 156-C-terminus in complex with c-di-GMP superimposed on the apo structure.*** *PelD binds two molecules of c-di-GMP at it's I-site but there is very little conformational change in response (arrow indicates largest shift), suggesting c-di-GMP acts as bridging unit to a second protein. Cartoon representation of complexed PelD (orange) and stick representation of c-di-GMP (cyan). PDB: 4DMZ and 4DN0*[113]

transmembrane helices as well as a structurally determined cytoplasmic tail containing a GAF domain and GGDEF domain[112,113]. The GAF domain is a common feature of cyclic nucleotide receptors while the GGDEF domain is typical of c-di-GMP synthesising diguanylate cyclases (see Figure 9: A). The active site of the GGDEF domain is degenerate (RNDEG) but the allosteric I-site, typically involved in inhibition by c-di-GMP and containing an RXXD motif, is conserved. This indicates that this portion of the protein has been

repurposed from its evolutionary roots to allow it to bind and transduce signals from the secondary messenger c-di-GMP. The overall function of the protein is to increase the synthesis of Pel in response to higher levels of c-di-GMP through another cytosolic member of the Pel operon, either PelF or PelG. In the Alg operon the function of secondary messenger sensing is by the IM transporter Alg44, which implies that the equivalent protein PelG may be the target of PelD, but also involves the glycosyltransferase Alg8[82], which might implicate PelF. C-di-GMP binding to PelD causes no significant structural rearrangements (see Figure 9: B), suggesting that the c-di-GMP molecule itself is involved in either forming of breaking contacts between PelD and its partner[112].

1.1.7.2 The Pel Polymerase: PelF

The second protein in the Pel pathway is PelF. This is a cytosolic glycosyltransferase that was suggested to use UDP-glucose (UDP-Glc) as a substrate[123]. Disruption of a UDP-Glc precursor pathway does not affect Pel production[10], suggesting that this is a false positive. The end product of the pel pathway was also found to be composed of subunits of N-acetylglucosamine and N-acetylgalactosamine (GalNAc)[10]. A possible explanation is that PelF may show multiple substrate specificities such as has been observed with the *S. thermophilus* exopolysaccharide system[124], where galactose (Gal) was incorporated in the absence of the preferred GalNAc.

Sequence analysis of PelF using PHYRE2[125] suggests the nearest structural homologue to be sucrose synthase 1 (SS1, PDB: 3S29) from *Arabadopsis thaliana*[126]. This enzyme is involved in the reversible cleavage and formation of glycosidic bonds between glucose and fructose, using UDP-Glc as a donor in the formation of sucrose. While this may further support UDP-Glc as the most viable substrate for PelF, this particular evidence is flimsy at best as SS1 belongs to the GT-4 family which are notorious for sequence and structural similarity between homologous proteins using divergent donor-sugars[127]. It can, however, be inferred that PelF is highly likely to adopt the canonical GT-B family fold and that its catalytic activity is more likely to retain stereochemistry than invert it[119,128]. SS1 does diverge from PelF in that it contains a regulatory domain at its N-terminus that does not overlay with the predicted fold of PelF, this domain is involved in membrane-association[129] and its absence supports cytosolic localisation of PelF.

1.1.7.3 The Inner Membrane Transporter: PelG

The third protein in the Pel pathway is PelG. This is predicted to be an IM protein with 12 transmembrane helices and is the most likely candidate within the pathway to act as a transporter across the IM to allow immature Pel polymer to reach the periplasm. The strongest matches using PHYRE2[125] are somewhat surprisingly not traditional sugar transporters but a combination of multidrug/oligosaccharidyl-lipid/polysaccharide (MOP) superfamily proteins, including flippase MurJ[130,131] (PDB: 5T77 & 6CC4), and

a selection proteins from the MATE family of multi-drug efflux transporters[132–134] (PDB: 5XJJ, 5Y50 & 6IDR). While the sequence identity of these proteins is very low (11-13%), this would not be unusual in integral membrane proteins as the hydrophobic environment of the membrane requires the presence of similarly hydrophobic amino acids throughout the membrane interface but specific hydrophobic amino acids are not required and therefore sequence conservation in these areas tends to be exceptionally low.

MurJ, of the oligosaccharidyl-lipid flippase (OLF) family, is involved in peptidoglycan synthesis and is responsible for flipping peptidoglycan-bound lipid II from the cytoplasmic to the periplasmic side of the inner membrane. While it is possible that a similar mechanism might be used for the transfer of the Pel polysaccharide across the membrane, this seems unlikely as catalysis would be required for both the addition and removal of Pel from the lipid which would require additional functionality from multiple proteins within the operon which has not been computationally predicted and diverges from the other exopolysaccharide export pathways used by *P. aeruginosa*.

Of the MATE transporters identified, two are from plants[132,133] and one is prokaryotic[134]. While there does appear to be overlap between these phylogenies[132], they are more typically attributed to one eukaryotic and two prokaryotic (NorM, DinF) subfamilies[135]. These proteins are involved in the export of, typically noxious, substances coupled to a cationic gradient, with broad substrate specificity[133]. This seems an unlikely fit for an exporter of a polysaccharide that has of yet to be related to any other functionality.

The MOP superfamily does, however, contain a family that relates directly to the proposed function of PelG; the polysaccharide transporter (PST) family. This family includes both flippases and proteins involved in the transport of exopolysaccharides and capsular polysaccharides in bacteria[136]. A structural characteristic of the MOP superfamily also corresponds to that proposed for PelG; a set of 12 transmembrane helices arranged in two bundles of 6. At 456 residues long, PelG also fits neatly into the typical 400-500 residue size of this family[136]. No structure has been determined from a protein of the PST family to date, explaining the divergence during homology modelling, but homology with MATE transporters suggests that a similar antiporter mechanism may be utilised in substrate transport.

1.1.7.4 The Periplasmic Scaffold: PelE

The remaining four proteins of the *pel* operon are believed to be involved in a trans-periplasmic complex of unknown stoichiometry formed around PelE at the IM and PelB at the OM. PelE is an IM protein with two N-terminal transmembrane helices and a large soluble domain located within the periplasm[117]. This was initially predicted to contain an armadillo (ARM)-like helical domain preceding at a protein-protein interactive tetratricopeptide repeat (TPR) domain containing at least three repeats[137].

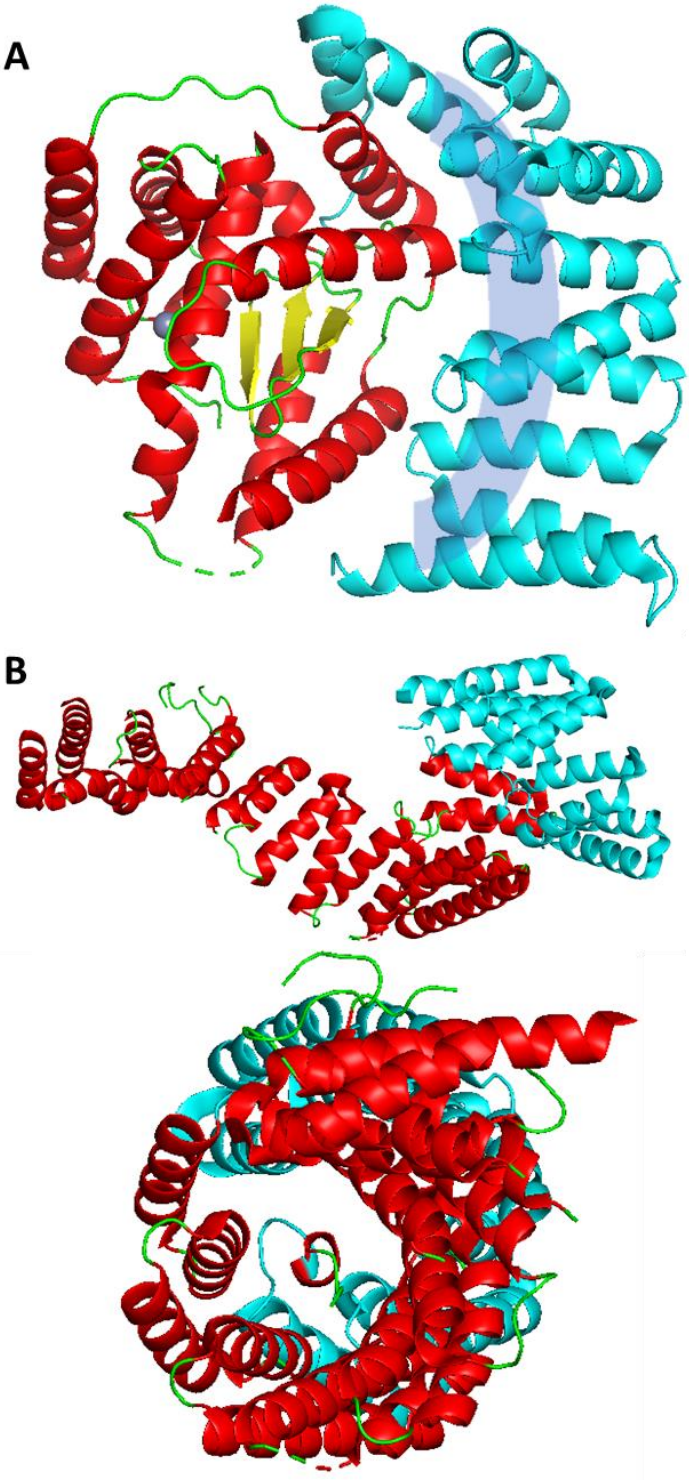

Figure 10: Tetratricopeptide repeat (TPR) domain containing proteins with homology to PelE. ***A: Cartoon representation of periplasmic chaperone BepA.*** *The protein has an N-terminal metalloprotease domain (helices in red, sheets in yellow, loops in green, metal ion in grey) adjoining a C-terminal TPR domain (cyan). At the interface between the two the TPR domain forms a large concave α-helical binding pocket (shaded) supporting the activity of the N-terminal domain for misfolded proteins. PDB: 6AIT*[138]*.* ***B; Cartoon representation of protein: protein interaction scaffold TTC7B in side and end-on views.*** *TTC7B is an elongated super-helical TPR protein capable of binding three proteins simultaneously to draw them into close proximity and enhance their activity. Non-competitive binding indicates three independent binding sites. The predicted homology region for PelE (cyan) most likely represents one binding site. PDB: 5DSE*[139]*.*

From this, the proposed function of the protein is to act as a scaffold at the IM for the assembly of proteins to transport Pel across the periplasm. TPR domains (see Figure 10) are typically degenerate repeats around 34 residues long that adopt a helix-turn-helix fold in 3-16 tandem arrays that together form an anti-parallel superhelix with a concave/convex surface for ligand binding (typically at the concave surface)[140]. Ligand binding by these proteins is typically highly specific as the large surface area binding site allows very precise and extensive amino acid placement to generate ideal hydrophobicity, charge and electrostatics for the interaction. The ligands themselves typically conform to either an extended coil, α-helix or combination of the two, making other TPR domains a viable interacting partner that can lead to oligomerisation. While the interacting partner of a TPR domain can be predicted by opposing surface charges, this requires structural information for both proteins.

The ARM domain was named for the arthropod from which proteins containing the domain were first identified[141]. They are repeats of around 42 residues in length that fold into three α-helices. Tandem ARM domains fold into a right handed superhelix that generates a concave protein binding surface held in place by a hydrophobic convex core, giving it very similar structural and functional properties to TPR domains[142,143]. The preferred binding target of an ARM domain is a disordered region of a protein which then adopts an ordered conformation based on the interaction[142]. The number of tandem repeats is variable but typically involves 3-12 which are generally contiguous, but can also be separated by small linker regions, and have previously been reported to be present on the same protein as TPR domains[143].

The strongest matches on PHYRE2[125] are BepA[138] (PDB: 6AIT) from *E.coli*, TTC7[139] (PDB: 5DSE) from *Homo sapiens*, Naa15[144,145] (PDB: 5NNR, 4KVM, 6C95, 4HNX) from *Chaetomonium thermophilum*, *Schizosaccharomyces pombe*, *Homo sapiens* and *Saccharomyces cerevisiae* (respectively), *O*-

GlcNAc transferase (OGT)[146] from *Homo sapiens* and Cut9[147] from *Schizosaccharomyces pombe*. A second search with just the C-terminal soluble domain provided near identical results. The sequence identity of these results is low (~15%) which suggests that these matches are unreliable for a protein with a large soluble domain, however, it is worth noting that the sequence conservation between TPR domains is relatively low as the consensus sequence for each repeat involves only positions 4, 7, 8, 11, 20, 24, 27 and 32, of which only 3 positions are conserved for a single amino acid, with the remaining sequence being tuned towards individual ligand recognition and therefore highly variable[140,148]. This means that I could expect a sequence identity for the consensus sequence of such domains to vary from 9-25% for an individual repeat.

BepA is a periplasmic chaperone protein involved in the correct folding of OM proteins and the proteolysis of misfolded OM proteins (see Figure 9: A). The protein comprises two domains, an N-terminal Zn metalloprotease domain (45-288) and a C-terminal TPR domain (305-482). The predicted

homology covers the region of 103-445 so includes incomplete contents of both domains. The canonical HEXXH motif of the metalloprotease domain is not found within the sequence of PelE, suggesting this would represent a degenerate protein-binding domain or a false positive. The TPR domain consists of 10 antiparallel α-helixes which form an extended sheet which interacts with the catalytic domain over a large surface area as a hydrophilic and negatively charged concave binding surface and cavity.

TTC7B is reported to be capable of binding three protein ligands non-competitively, therefore most likely through three separate binding sites. It appears to act as a direct enhancer of phosphorylation by phosphatidylinositol 4-kinase IIIα complex as well as recruiting further enhancers. The entire protein in an α-helical TPR superhelix with the predicted alignment covering 580-835 of the protein's C-terminal region where phosphatidylinositol 4-kinase IIIα is predicted to bind.

Naa15 is the auxiliary subunit of the NatA complex which supports the N-acetylation of proteins by the catalytic Naa10 subunit[144]. Naa15 is composed of 13 TPR motifs that adopts a ring-like topology in which Naa10 sits[145]. Complex formation with Naa15 induces a conformational change in Naa10 that enhances its stability. Additionally, Naa10 can bind HypK in its C-terminal region which induces a conformational change in Naa10 that negatively regulates Naa15[144]. PelE corresponds to the N-terminal 7-153 region of the protein which is not involved in the binding of either of the identified partners.

OGT is involved is a glycosyltransferase involved in the addition of GlcNAc to serines and threonines to act as a signal of the metabolic state of the cell in glucose regulation. The protein is subdivided into a flexible "hinge-like" N-terminal 13 TPR motif domain thought to be involved in stabilising protein interactions and a C-terminal glycosyltransferase domain. Unlike the other homologues found, OGT displays activity against a wide range of targets, suggesting that its N-terminus might represent a "broad spectrum" TPR domain. The predicted homology of PelE corresponds to 314-461 on this structure, the first 8.5 helices of the TPR motif.

Cut9 is a component of a hetero-oligomeric complex of 13 different proteins that acts as an E3 ubiquitin ligase for targeting proteins for degradation. The protein itself is a 14 TRP motif homodimeric superhelix that binds and sequesters N-acetyl-methionine of Hcn1 to protect it from ubitiquin-dependent degradation. The homologous region is 342-591 which has not been identified as a site important for binding heterologous molecules but, alongside the rest of the molecule, contributes towards the concave surface that interlocks with the subunit of its dimer.

These results underlie the diverse nature of TRP domain containing proteins and support the existence of such a domain within PelE. Conversely, none of the proteins identified at high confidence (>95%) were found to contain an identified ARM domain yet all the proteins discussed

aligned over both the predicted TPR domain and the predicted ARM domain, suggesting that one may function as an extension of the other.

1.1.7.5 The Proof-Reading and Modification Enzyme: PelA

PelA is a large, multi-domain protein with diverse functions. The protein is localised within the periplasm but has also shown to associate with the membrane fraction[149], indicating potential activity in the periplasm and at both or either of the IM and OM. Further work[150] has confirmed that PelA co-localises with the OM protein PelB and suggests that PelA is active both within the periplasm and at the OM. Thus far, two functions have been attributed to PelA; it acts as a glycoside hydrolase[71,72,150] and a de-N-acetylase[149].

The glycoside hydrolase domain of PelA is located at the N-terminus of the protein and has been shown to break down Pel-type biofilms *in vivo*[71,72,150], which has garnered interest in using it and similar enzymes as treatment options for biofilm associated infections. This domain of PelA is constituently active, however, association of PelA with PelB inhibits this activity[150] (see Figure 11); a function of the domain can therefore be suggested. In the absence of PelB, PelA causes immature Pel polysaccharide to be broken down, this prevents a potentially dangerous build-up of the insoluble polymer where it is not able to be exported from the cell. This might be important in situations where assembly of PelB is incomplete or defective or if transport across the periplasm proves to be defective. This is reminiscent of PslG and the interaction between AlgK and AlgL. Interestingly, the inhibition of this domain of PelA by PelB must be the result of a co-operative conformational change as the glycoside hydrolase domain itself does not interact with PelB[150].

The second identified domain of PelA is the de-N-acetylase domain. Mutational studies have suggested that this adopts a carbohydrate esterase family 4 fold as mutation of sites predicted to be involved in divalent metal co-factor binding successfully reduced the ability of the enzyme to deacetylate p-nitrophenol acetate *in vitro* and abolished the ability of *P. aeruginosa* to produce a Pel-type biofilm *in vivo*[149]. Interestingly, the cell contents of the *in vivo* mutants did not contain detectable Pel, which was taken to mean that Pel had been either broken down or had not been synthesised. This would imply that the glycoside hydrolase domain of PelA may reactivate in the presence of unprocessed Pel. Alternatively, it could also be that the anti-Pel antibodies used in the assay require this deacetylation to recognise the molecule.

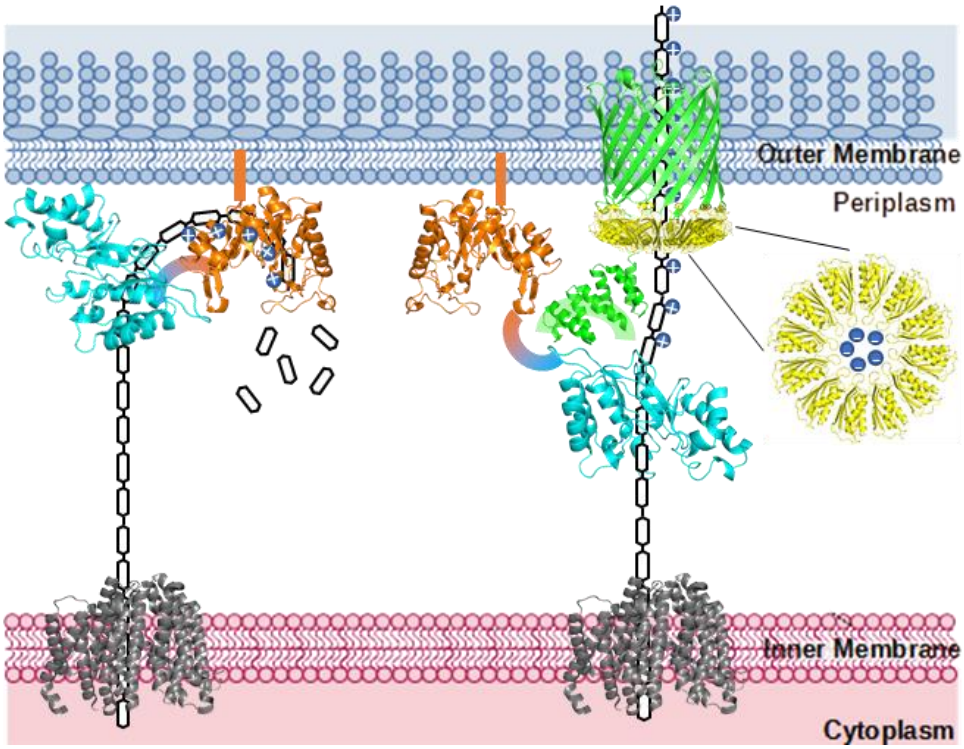

Figure 11: Model of pel transport and regulation across the periplasmic space**.** Pel polysaccharide is fed into the periplasmic space by PelD (grey). If correct assembly of the pel machinery is incomplete (left) then the polysaccharide encounters PelA and is degraded by its N-terminal glycoside hydrolase domain (orange). Correct assembly (right) directs pel to the de-N-acetylase domain of PelA which matures the polymer and gives a positive charge. The periplasmic domain of PelB (green) interacts with PelA, C-terminal to the glycoside hydrolase domain, to prevent degradation of the polymer. PelC (yellow) forms a dodecameric ring at the periplasmic face of PelB's β-barrel which attracts the polysaccharide into its negatively charged lumen. Passing through the PelC complex guides pel into the PelB's transmembrane channel and allows secretion from the cell. PelA structures are based on PHYRE2 homology models[125]. PelB uses PgaA[151] (PDB: 4Y25) for it's β-barrel and a resolved structure of its TPR domains[150] for its periplasmic domain (PDB: 5WTF). PelC uses the structure of PelC from Paraburkholderia phytofirmans[152] (PDB:5T11). PelG uses the structure of MurJ[131] (PDB: 5T77).

1.1.7.6 The Outer Membrane Scaffold and Transporter: PelB

PelB is an extremely large (135 kDa) OM protein. Following the OM signal peptide (1-46), the protein forms a large periplasmic domain containing 16 TPR repeats (47-880) with the remaining protein predicted to fold into an OM spanning β-barrel porin[150] for transport of the mature Pel polymer across the OM. TPRs 9-15 have been implicated in binding and inhibition of PelA[150] while the purpose of TPRs 1-8 and 16 remain to be determined. Candidates from the Pel system which could be involved in an interaction with the undesignated TPRs include the periplasmic domains of PelE and the lipoprotein PelC. While PelC is proposed to interact with PelB, there has

been a failure to confirm this interaction[152]. It is also possible that an unrelated but abundant periplasmic protein may be recruited for an as-yet unidentified purpose.

PHYRE2[125] homology modelling of the periplasmic domain of PelB returns a series of TPR rich proteins: Superkiller Protein 3[153] (PDB: 4BUJ), Pre-mRNA-processing factor 6[154] (PDB:5O9Z), Ctr9[155] (PDB: 6AF0) and Flagellar associated protein IFT70 (PDB: 4UZY) as well as TTC7B[139] (PDB: 5DSE), Nat15[144] (PDB: 5NNR, C695, 4KVM) and Cut9[147] (PDB: 2XPI), previously discussed for the homology of PelE. Sequence identity for those matches remains within the 10-16% region, as might be expected from modelling of a TPR-rich protein. For brevity, TPR containing proteins from this list shall only be discussed where they provide a novel interaction or activity.

Superkiller protein 3 (Ski3) is involved in the formation of a heterotetrameric Ski complex which acts as a subunit of the exosome RNA degradation complex. Ski3 forms two helical arms made up of 33 TPR motifs. The N-terminal arm causes a 5-fold reduction in the activity of the catalytic helicase subunit Ski2 but increases the RNA binding affinity of the enzyme while the C-terminal arm incorporates the Ski8 dimer into the complex. All three subunits are bound within the concave face of the protein with an extension of the N-terminal arm of Ski3 point out into solution and containing a potential binding site for a fifth partner, Ski7, and forms a RNA extrusion pathway between the helicase and Ski7 leading towards the exosome. This last property of Ski3 is interesting as this illustrates the possibility for TPR domains to guide sugar rich molecules towards a second subunit for further processing or transport.

A PHYRE2[125] search using the C-terminal domain of PelB produced relatively few results at a confidence of greater than 90%: PgaA[151] (PDB: 4Y25), FadL[156,157] (PDB: 1T16, 3DWO), TodX[158] (PDB: 3BRZ) and Ail[159] (PDB: 3QRA). As expected of a membrane protein component, sequence identity from the matches is low, corresponding to 10-16%. Only the PgaA structure provides a solid match across the majority of the sequence whereas FadL and TodX structures correspond only to 900-~1088 and Ail to 1015-1188 of PelB.

The PgaA structure corresponds to the C-terminal domain of the protein which is an OM β-barrel porin involved in the translocation of de-*N*-acetylated poly-β-1,6-*N*-acetyl-d-glucosamine (dPNAG) from the periplasm to the extracellular space as a component of biofilm formation. The protein additionally has N-terminal periplasmic domain roughly half the size of that found on PelB containing an 8-TPR motif domain that binds PgaB, a de-*N*-acetylase potentially equivalent to PelA, at a 1:1 ratio. The protein's 16 strand β-barrel contains a negatively charged periplasmic face and inner pathway across one side of the barrel and to attract and guide positively charged dPNAG. The periplasmic face of the protein is open but the extracellular face contains numerous small loops which appear to obstruct the opening but do not impact biofilm formation when deleted.

FadL is an OM transporter of long-chain fatty acids. The section corresponding to PelB is a series of β-sheets forming the C-terminal half of the β-barrel that does not contain any loops within the barrel's lumen or external to the barrel. This section of the protein is not involved in the proposed fatty acid binding groove but does contribute towards the high affinity hydrophobic binding site proposed to draw fatty acids down from the binding groove by diffusion towards the lateral groove generated by the N-terminal half of the barrel that would allow entry of the hydrophobic molecule into the intermembrane space. It is unlikely that a lateral diffusion model would function for PelB as the necessary structural region is absent and the direction of transport would not support exit from the cell. It is, however, worth noting that pel is insoluble and as such the predicted hydrophobic β-barrel may be important to maintain the flow of the molecule. TodX is a homologue of FadL with little structural divergence within the region of homology to PelB.

Ail is an 8-stranded OM β-barrel with 4 extracellular loops involved in adhesion to host extracellular matrix proteins. The lumen of this barrel is closed and not involved in transport across the membrane. The homology alignment with PelB covers the entire Ail molecule. While the shared role of mediating attachment implies that extracellular loops on a PelB β-barrel may be functional, the divergence in known roles of these proteins suggest adhesin activity is unlikely.

While the PgaA protein appears to be an excellent model for the activity of PelB, some contrasts should also be noted. Pel is an insoluble 1, 4-linked heteropolymer primarily composed of N-acetyl-galactosamine at a stoichiometric ratio of 5:1 with N-acetyl-glucosamine[10] while dPNAG is a soluble 1, 6-linked homopolymer of N-acetyl-glucosamine[151]. Pel is a linear polymer whereas dPNAG is helical. De-*N*-acetylation of both polymers occurs in the periplasm and the extent of this may be important in determining whether the polymer is secreted[160] or remains cell-associated[161], however, the PelA homologue PgaB is an outer membrane bound lipoprotein that does not exhibit the same proof-reading activity as PelA[161]. The *pgaABCD* operon is also much smaller as PgaC is proposed to be an IM GT-2 family glycosyltransferase that incorporates IM translocation in co-ordination with PgaD.

1.1.7.7 The Outer Membrane Accessory Lipoprotein: PelC

PelC is a small periplasmic/OM lipoprotein[162] that is proposed to form a complex with PelB. While *P. aeruginosa* PelC has been found to not be suited to crystallisation, structures have been solved from two other bacteria[152], showing that the protein forms into a 12 subunit ring with a periplasmic facing negative charge, a membrane facing aromatic ring and a 30Å pore at its centre. It is proposed to assemble around a section of the periplasmic domain of PelB that connects the TPR domains to the OM β-barrel and over the periplasmic entrance of that prion. The electronegativity of the periplasmic surface would attract the pel polymer once a positive charge had been applied to it through deacetylation by

PelA. The pel polymer would then feed directly into the porin through PelC's pore for transport across the OM. While an elegant interpretation of the available data, it is important to note that this proposal is based upon circumstantial rather than direct evidence.

1.1.8 The benefits of characterising pel synthesis

The pel polysaccharide makes contributions towards ABR through a variety of means. Disruption of a pel biofilm has additionally been shown to reduce ABR and increase vulnerability of *P. aeruginosa* to host defences[71,72]. This raises the possibility that better understanding of pel synthesis will provide alternative means of disrupting a biofilm to impair the ABR of a clinical infection through development of a drug targeting a specific process within that synthesis. The strongest approach to support this goal is through structural biology, as determining the structure of components of the pel operon will provide a wealth of information on the functionality of the protein as well as spatial information allowing rational drug design.

While all components of the pel system are required for successful synthesis of pel, some are more attractive targets than others. *P. aeruginosa* is a Gram-negative bacterium and therefore has an impermeable outer membrane which limits the scope of molecules that can gain access to the cell to ones that are less than 600 Da in size or that match the intake specificity of a transporter[163]. PelB, as an outer membrane protein, is therefore an attractive target as the outer surface of the protein's β-barrel is relatively exposed and therefore accessible. Given that pel is insoluble, there is a possibility that block export via PelB might have a lytic effect upon the cell.

Periplasmic components of the system are the next most druggable elements as only a single membrane needs to be bypassed for a drug to reach its target. PelA has shown that it is capable of hydrolytic activity that breaks down pel and that this activity is suppressed by interactions with PelB[150]. Interfering with this interaction would prevent secretion of pel. Likewise, supressing the de-*N*-acetlyase activity of PelA would likely prevent maturation of the polysaccharide and its recognition by the PelBC complex. Blocking the lumen of the PelC complex or preventing its oligomerisation would also prevent correct export. The role of PelE is more opaque but identifying the binding partner(s) of its periplasmic domain and inhibiting that interaction would likely prevent formation of the periplasmic machinery required for pel production.

Perhaps one of the most interesting targets would be the periplasmic face of the inner membrane protein PelG, as blocking export by this transporter would prevent the polysaccharide from reaching the periplasm and access to the proof-reading enzyme PelA. Unfortunately, ΔpelG is not constitutively lethal, suggesting that there might be a cytoplasmic pathway capable of breaking down the immature polysaccharide, however, this may not be the case if activation of PelF is dependent upon the presence of PelG. Should a cytoplasmic pathway for pel degradation not exist then blocking PelG might

be very damaging to the organism, especially is the immature polymer is an insoluble as the mature polysaccharide.

The cytoplasmic components of the pel system are the least attractive targets for drugs, as their entry into this compartment would require bypassing both membranes and small molecules are routinely removed from the cytoplasm by multi-drug efflux transporters. Nevertheless, both PelF and PelD are essential to the operation of the pel system and preventing either the glycosyltransferase reaction or signal recognition, respectively, would prevent pel production. The cytoplasmic domain of PelD has already been structurally determined but the PelF protein has not and may provide an ideal target for structural biology, given its likely high solubility.

1.2.1 Outer membranes in antibiotic resistance

As a Gram-negative organisms, *Pseudomonas* and *Salmonella*, both have an inner membrane (IM) and an outer membrane (OM), as opposed to the single IM of a Gram positive bacteria. Both membranes provide barriers to the diffusion of antibiotics into the cell and so Gram negative bacteria are typically more difficult to treat as the additional membrane prevents the entry of many drugs, necessitating the use of broad spectrum antibiotics for treatment as they are able to bypass the OM. Finding methods of disrupting the OM would allow a greater pool of antibiotics to be used in treatment, as well as potentially having an antimicrobial effect in their own right, and so proteins involved in the synthesis of the OM are of great interest as drug targets.

Lipopolysaccharides (LPS) are a key component of the outer leaflet of the asymmetric OM, taking the place of phospholipids that populate the majority of other bacterial membranes. LPS have a number of functions but were first identified as an important molecule due to their ability to act as an endotoxin; LPS is released from dividing or dying bacteria which can act as an antigen in small quantities but also cause septic shock, kidney and liver failure when released at higher concentrations, such as caused by antibiotic clearance of an infection[164]. In *E.coli* and *Salmonella*, impaired LPS biosynthesis and transport is associated with impaired growth and hyper-susceptibility to antibiotics[165,166]. LPS co-interact through divalent cations and hydrophobic interactions of lipid A to form a densely packed permeability barrier which excludes small hydrophobic compounds such as detergents, antibiotics and bile salts to protect the bacteria from harsh environmental conditions[167,168]. Furthermore, LPS is the major amphoteric molecule composing the outer leaflet of the OM; its depletion therefore impacts a variety of factors including folding and function of OM proteins (OMPs) and synthesis of peptidoglycan and secondary cell wall polymers[168]. Inhibition of LPS transport has been shown to cause a toxic build-up of LPS precursors at the IM in *Salmonella*[169], resulting from sequestration of resources that inhibits essential cell wall biosynthesis, and has been of great interest as a potential drug target.

1.2.2 *Salmonella enterica* infections

Salmonella species are the world's leading cause of foodborne illnesses and are responsible for hundreds of thousands of deaths annually[170], making them a significant public health concern. They are rod-shaped Gram negative facultative anaerobes of the *Enterobacteriaceae* family. *Salmonella* is primarily associated with poultry and fish and causes fever, acute gastroenteritis and septicaemia in humans[171], with some strains also responsible for abortions in ruminant animals[171]. Two serotypes are of particular importance to human health; *S. typhi* is the etiological agent of enteric fever, which offers the most dangerous infection, and *S. typhimurium*, which is the most ubiquitous and the cause of most *Salmonella*-associated gastroenteritis.

Infections are typically treated with antibiotics, however, drug resistance is becoming increasingly a barrier to *Salmonella* treatment[69,172,173] as antibiotics are routinely overused in farming practices as growth promoters and many of *Salmonella*'s preferred hosts are heavily farmed animals such as chickens, pigs, salmon and cattle. The development of resistance underlies a need to produce new therapies and antibiotics for the treatment of such pathogens, as is supported by the World Health Organisation's identification of fluoroquinolone-resistant *Salmonella* as a high priority in its list of top 10 bacterial pathogen list[69].

1.2.3 Outer membrane biogenesis

While the IM and OM both contain transmembrane proteins, those found in the IM typically adopts one more more α-helices to span the membrane while outer membrane proteins (OMP) generally consist of anti-parallel β-strands that adopt a β-barrel conformation with a hydrophilic interior to allow transit through the membrane and a hydrophobic exterior to anchor it in place[174]. The periplasmic leaflet of the OM also contains periplasm facing lipoproteins anchored via an N-terminal N-acyl-diacylglycerylcysteine. While the IM is energised by a proton gradient, this is not the case for the OM and ATP availability is limited in the periplasm[175]. Nutrient diffusion through the OM typically occurs passively through porins[167]. Porins are OMPs that contain a hydrophilic channel through which small hydrophilic molecules of around 600 Da or less may pass.

Components of the OM are synthesised in the cytoplasm, or at the cytoplasm-IM interface, and then transported across the IM and periplasm to be inserted into the OM[175] (see Figure 12). OMPs in the cytoplasm are targeted to the OM with a signal peptide which acts as recognition sequence for the Sec protein complex for translocation from the cytoplasm to the periplasm[176]. Chaperones then guide the OMP across the periplasm for insertion into the OM by the B-barrel Assembly Machinery (BAM) A-E complex[177–180]. Lipoproteins are also translocated to the outer leaflet of the IM by Sec machinery, but the processes diverge at this point. Lipoproteins are matured at the IM-periplasm interface by Lgt, LspA and Lnt[181–183]. OM lipoproteins are sorted by the LolCDE complex (see Figure 12) and transferred to LolA[184]. LolA transports the lipoprotein across the periplasm

and transfers it to LolB, which then inserts the protein into the inner leaflet of the OM[185,186].

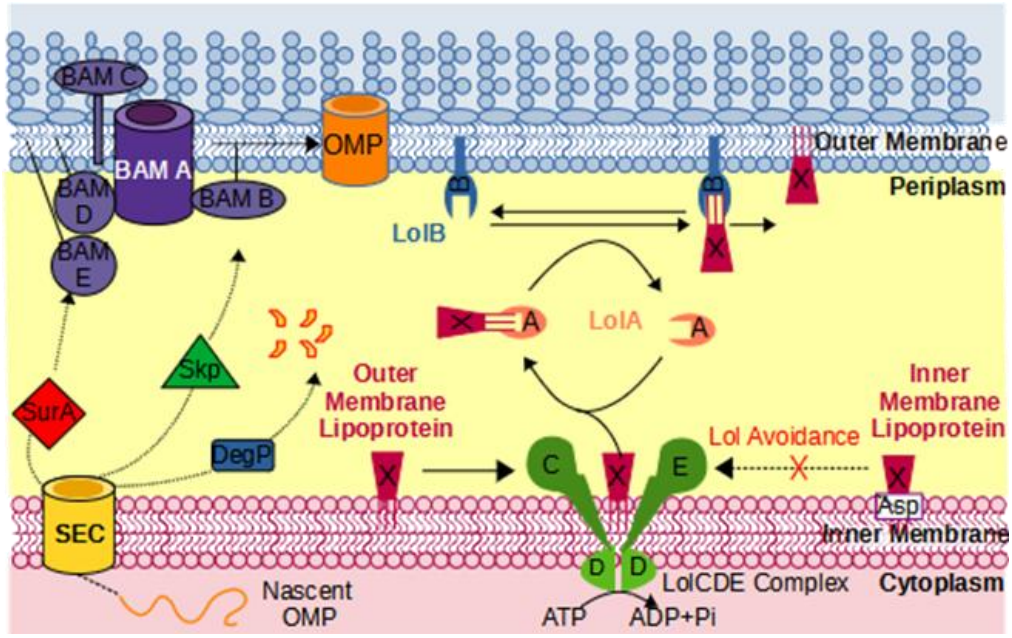

Figure 12: Model of outer membrane protein and lipoprotein insertion into the outer membrane. Nascent outer membrane proteins (OMPs) are recognised based upon their signal peptides by the Sec machinery and transported across the inner membrane where they are recognised by chaperones SurA or Skp, or degraded by DegP. The chaperones transfer the protein to the BAM complex which inserts them into the outer membrane. Lipoproteins are sorted by the LolCDE complex, which avoids signal peptides with an asparagine recognition sequence to maintain them in the inner membrane. Lipoproteins destined for the outer membrane are energetically extracted from the inner membrane by the LolCDE complex and passed to the periplasmic chaperone LolA, which buries their membrane anchor in a hydrophobic pocket. LolA exchanges the protein with LolB which then inserts the membrane anchor into the outer membrane.

Phospholipids are synthesized at the cytoplasmic face of the IM[187,188]. The majority of phospholipids used in the inner leaflet of the OM are phosphatidylethanolamine, phosphatidylglycerol or cardiolipin, comprising around 70%, 20% and 10% of total lipid by mass respectively[189]. To reach the OM, phospholipids are flipped from the inner to the outer leaflet of the IM. A dedicated flippase for this process has yet to be identified and it has been suggested that the α-helical sections of general integral membrane proteins may facilitate flipping[190]. The process of migration from the IM to the OM is poorly understood as, unlike other components of the OM, a dedicated group of proteins effecting transport has not been identified. PbgA has been proposed to be involved in cardiolipin transport, it's increased expression correlates with increased level of cardiolipin at both membranes and its IM-periplasmic positioning and OM binding by its periplasmic domain suggests a role in transport[191]. Retrograde phospholipid transport (OM →IM) has been identified as effected by Tol-Pal trans-

envelope complex and by the OmpC-Mla pathway but as phospholipid transport across the periplasm is bidirectional it has proven difficult to establish the exact roles of these complexes[192–194].

In contrast, the transport of LPS is well understood (see Figure 13). LPS is synthesized at the cytoplasm-IM interface and then flipped to the periplasmic leaflet of the IM by the flippase MsbA[194]. The molecule is maturated at this location and then undergoes transport via the Lpt system. The $LptB_2CFG$ complex at the IM is an ABC transporter which transfers LPS to periplasmic LptA[195]. LptA is a chaperone which encases the hydrophobic component of LPS and forms an oligomeric bridge across the periplasm[196]. At the OM LPS is collected by the lipoprotein LptE and fed into the OMP LptD, which forms a β-barrel with a lateral gate which is opened on LPS binding and deposits LPS directly into the outer leaflet of the OM[165].

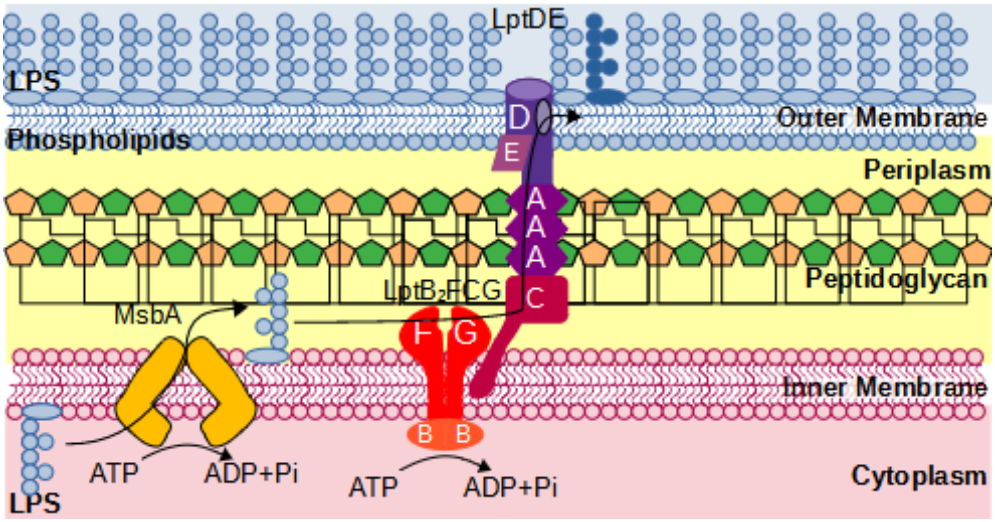

Figure 13: Model of outer membrane lipopolysaccharide biogenesis. LPS is synthesised at the cytoplasmic side of the inner membrane and energetically transferred to the periplasmic leaflet by the flippase MsbA. LPS is extracted from the periplasmic leaflet by the $LptB_2FCG$ complex and transferred via a series of periplasmic LptA chaperones to the LptDE outer membrane complex. LptE feeds LPS into the lumen of LptD which then pushes LPS directly into the outer leaflet of the outer membrane through a lateral gate.

1.2.4 Lipopolysaccharide synthesis in *S. enterica*.

LPS is a saccharolipid composed of three parts (see Figure 14A); lipid A, which serves to anchor LPS to the membrane, the core oligosaccharide, which forms a scaffold connecting the other two components, and the o-antigen, which is a variable repeating polysaccharide of 40 to 200 units that contributes towards surface adhesion and resistance to host defence mechanisms and antibiotics. The core oligosaccharide can be further subdivided into inner and outer core regions. Variation in LPS structure between different species and even strains is quite significant, with some bacteria producing only rough LPS, containing no O-antigen component,

while other, frequently pathogenic, bacteria generate the tripartite smooth LPS or semi-rough intermediate.

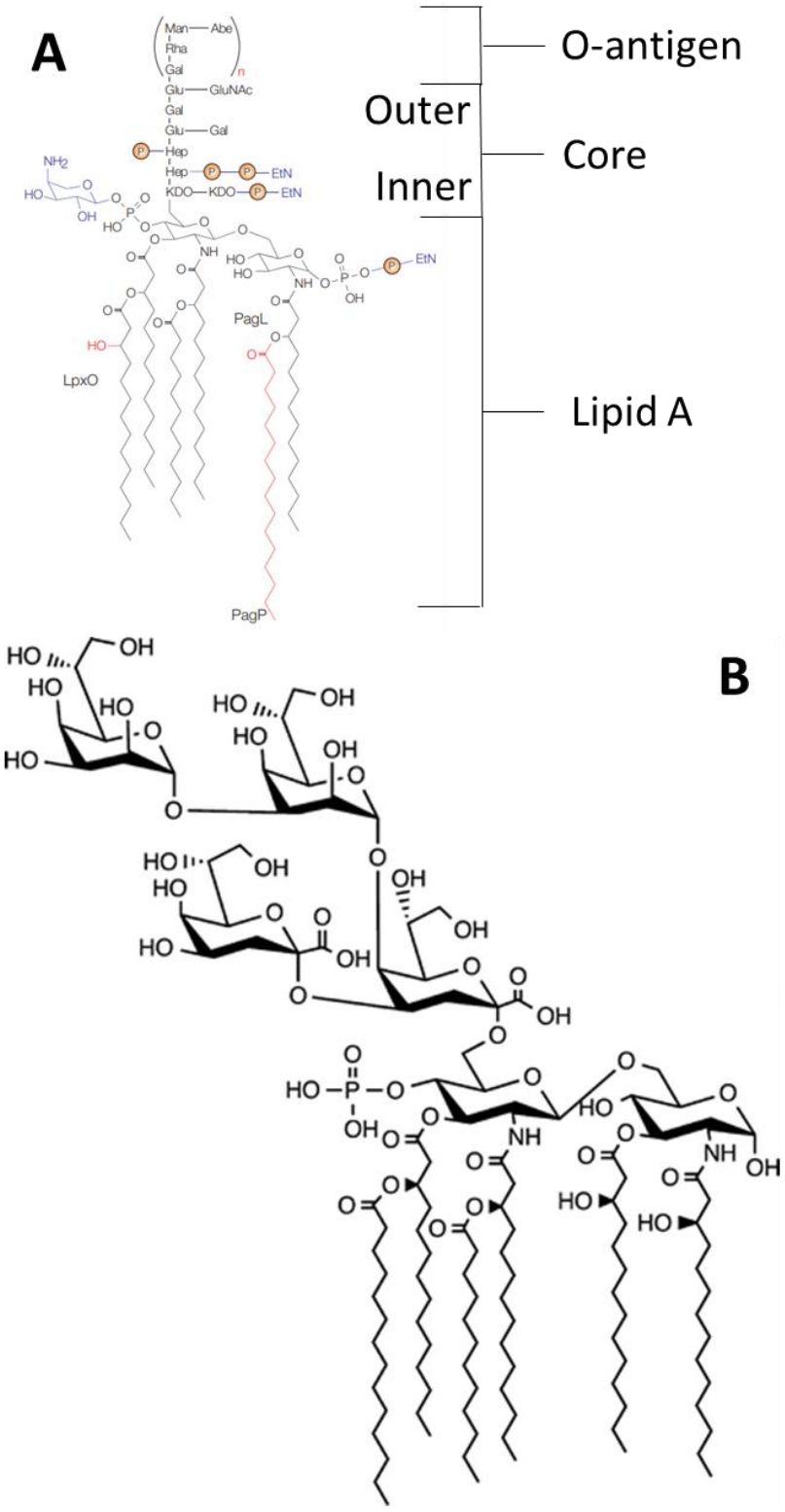

Figure 14: Lipopolysaccharide structure and synthesis. ***A: The structure of mature LPS from Salmonella enterica serotype Typhimurium (figure from Miller, S.I., et al, 2005)****. This is an essential component of the bacterial outer membrane. Subdivisions of the molecule are noted to the right. Lipid A anchors the molecule into the membrane. The core region acts as an invariant scaffold linking Lipid A and O-antigen components. O-antigen is a variable region that provides many of the specific functions of the molecule.* ***B: The structure of Hep_2-Kdo_2-lipid A (Figure from Qian, J., et al, 2014)****.*

This molecule represents the substrate modified for the beginning of construction of the outer core and therefore the substrate that the enzyme of interest WaaB will act upon.

A typical lipid A is a 3′, 3′ fatty acid substituted bisphosphorylated β-(1→6) linked glucosamine disaccharide with 2′, 2′ amide linkages[197]. The lipid A fatty acids vary in number (4-6), acetylated (4-7) and length (C10-20) between species and under different regulation conditions but consistently link to positions 2 and 3 of the glucosamines, with secondary chains sometimes attached to primary fatty acids. The 6′ position of glucosamine^{II} forms a glyosidic bond with the 2′ position of conserved α-3-deoxy-d-manno-oct-2-ulosonic acid (Kdo) of the core. The core consists of 8-12 sugars branching from the 5′ position of the conserved α-Kdo and with up to two additional Kdo subunits branching from the 4′ position of α-Kdo. The first three 1→5 linked sugars are L-glycero-d-manno-heptose residues (Hep) with frequent substitutions of heptose residues for phosphate, pyrophosphate, phophorylethanolamine or a different sugar. This forms the inner core. The outer core consists of an oligosaccharide of up to 6 units 1→3 linked to Hep^{II} and frequently branched with combinations of glucose (Glu), galatose (Gal) or their derivatives. The final component of smooth LPS is a highly diverse repeated oligosaccharide; each repeat unit may contain 2-5 subunits which may also branch from the main chain.

The lipidA biosynthesis pathway[198], also known as the Raetz pathway (see Figure 15A), is well characterised, highly conserved amongst Gram-negative bacteria and begins at the cytosol-IM surface. LpxA adds the first 3′ acyl chain to UDP-GlcNAc and LpxC removes the acetyl group covering the 2′ position to be replaced by the second acetyl chain by LpxD. The resulting UDP-di-acyl-GlcN either undergoes hydrolysis by membrane bound LpxH at the pyrophosphate bond to form lipid X, or undergoes condensation with existing lipid X to form the characteristic tetra-acyl-disaccharide of lipid A by the GT-B type glyosyltransferase LxpB, removing the final UDP in the process. Integral membrane proteins then take over as the LpxK expends ATP to add phosphate to the 4′ position of GlcN^{II} and addition of the inner core begins as WaaA, a second GT-B enzyme, incorporates two Kdo residues donated from cytidine monophosphate (CMP)-Kdo from the 6′. Two additional acyl chains are then added sequentially to the existing 2′ and 3′ acyl chains of GlcN^{II} by LpxL and LpxM respectively; at low temperature growth a homologous acetyl transferase, LpxP, supplants LpxL to add a longer acyl chain. The Kdo_2-lipidA is then flipped from the cytoplasmic leaflet of the IM to the periplasmic by the ABC transporter/flippase MsbA which shows high specificity for its substrate, particularly the phosphorylated regions of the saccharolipid. Once exposed to the periplasm, the LPS precursor is able to undergo numerous modifications dependent upon environmental conditions in the lipid A region and is then transported to the OM by the LptA-E system[165]. The O-antigen is synthesised and transported separately then ligated to rough LPS by WaaL[197].

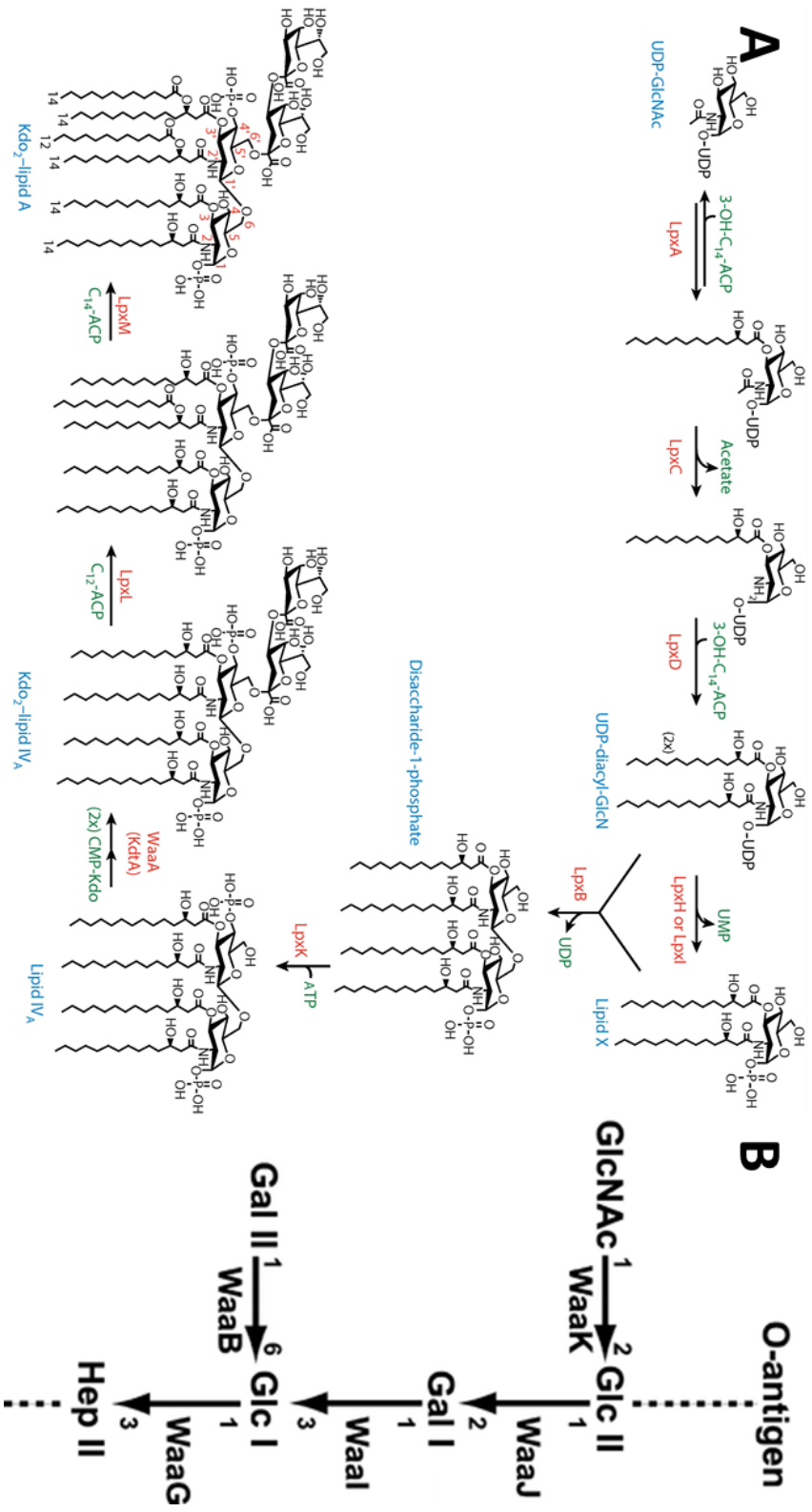

Figure 15: Lipopolysaccharide synthesis. ***A: Outline of the biosynthesis of Kdo$_2$-lipidA (the Raetz pathway, figure from Whitfield, C., Trent, S.M. 2014).*** *Intermediates are named in blue, enzymes catalysing reactions in red and substrates/co-products in green.* ***B: Conserved pathway elements of LPS outer core biosynthesis in E.coli and Salmonella enterica serotype Typhimurium (Figure from Qian, J., et al, 2014)****. The Hep II residue corresponds to the terminal sugar in* ***A****. The protein of interest in this work in Salmonella's WaaB, catalysing 1→6 condensation of Gal II to Glc I.*

The remainder of the biosynthesis of the inner and outer core is less well understood. The inner core contains an additional two L-glycero-D-manno-heptose sugar residues, 1→4 linked to Kdo^{II} and 1→3 linked to Hep^{II} respectively (see Figure 14B), and is well conserved amongst Gram-negative bacteria[199]. The outer core is more variable with five known structures in *E.coli* and two in *S. typhimurium*[200,201], however, an invariant structural feature in this region is a three linked hexose chain, beginning with glucose, which branch at sugars 1 and 3 (see Figure 15B). The addition of glucose to Hep^{II} is achieved by WaaG, a membrane bound cytosolic glycosyltransferase[202], prior to membrane flipping by MsbA. A second enzyme, WaaB, has been identified as adding 1→6 linked Gal^{II} to Glc^{I} to form the first outer core branch while a third enzyme WaaI add a 1→3 hexose sugar to the same Glc^{I} [199].

1.2.5 Structural and Functional Investigation of WaaB

Changes to the LPS core regions have been found to have potentially profound impacts upon the ability of *S. enterica* to invade epithelial cells[203], an important determinant in the ability of the bacteria to cause infection. Of these enzymes, WaaB was found to cause a reduction in invasion capacity, but to not prevent it entirely, and to lower the permeability of the OM[203]. Undermining the activity of this enzyme may therefore render the oft-antibiotic resistant organism more susceptible to treatment.

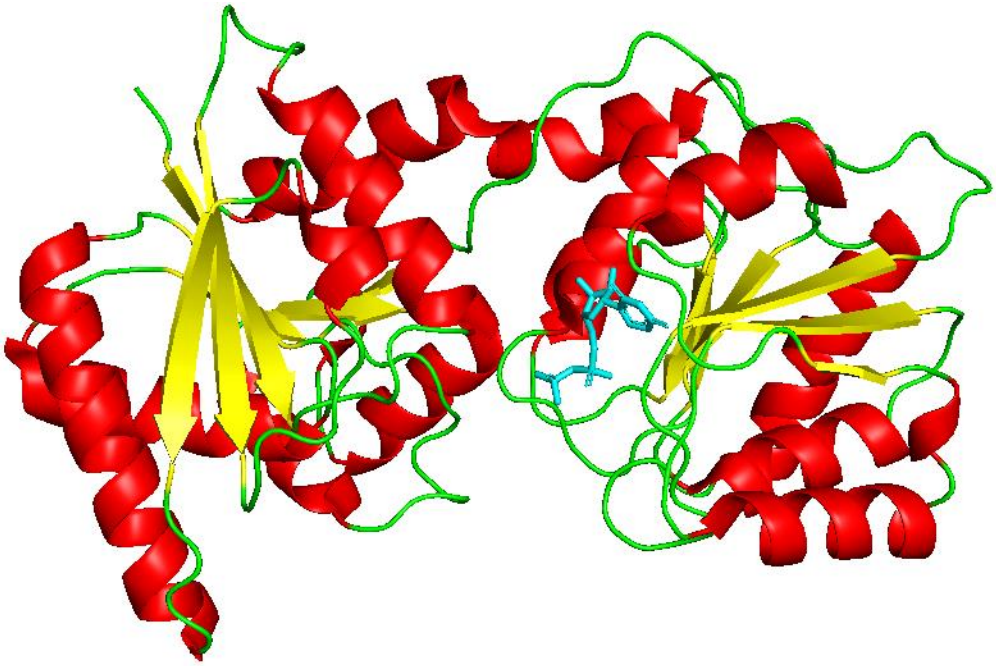

Figure 16: **Crystal structure of WaaB in complex with UDP.** Cartoon representation of PDB 5N80 showing α-helices (red), β-sheets (yellow), loops (green) and UDP (sticks, cyan). WaaB adopts a canonical GT-B fold with two facing Rossman-like domains. The C-terminal domain (right) is involved in UDP binding so is the donor-substrate binding domain while the N-terminal domain (left) is therefore a acceptor substrate binding domain. The active site cleft is found between the two domains and catalyses the 1→6 addition of UDP-galactose to the glucose of Glc-Kdo_2-lipidA. This

structure forms the basis for identification of residues important in substrate binding and catalysis for later investigation in this thesis.

The structure of this glycosyltransferase (GT) has recently been solved by the Dong group (awaiting publication) in the *apo* and UDP-bound forms (see Figure 16). These structures form the basis for the identification of residues important to UDP binding and catalysis, both processes essential to enzyme activity and potential drug targets, but such predictions need to be validated before further resources are dedicated to such projects. One objective of this thesis will be to identify the residues implicated in these processes and validate their importance to the enzymatic activity of the protein to provide evidence to guide rational drug design for this protein.

1.3.1 Antibiotic resistance in gut infections

The human gut lumen is a space that is extensively colonised by bacteria from a young age. The bacteria that occupy this space are diverse and there is a complex interplay between species which limits or negates the impact of potentially pathogenic species such as *P. aeruginosa*, *S. enterica* and *C. difficile* that causes them to act as commensal species[204] or preventing their colonisation altogether, termed colonisation resistance[205]. An unfortunate side effect of antibiotic treatment is that gut bacteria lacking ABR are often eliminated during treatment, allowing excessive colonisation of the gut by more resistant bacteria[204,205]. The previously mentioned, often Gram-negative, bacteria all possess extensive manifestations of ABR that allow this to occur[206] and their establishment within the gut is associated with intense, acute pathogenesis, including inflammation, diarrhoea and gastroenteritis[207]. Diarrhoeal diseases are one of the five leading causes of death in children under five and leads to the over 1.6 million mortalities per annum[208].

The natural and acquired resistance of these pathogens has led to novel methods of treatment that avoids use of antibiotics and instead focuses on re-establishing a health gut microbiome. Faecal transplants are gaining traction as a method of treating ABR colonisation of the gastrointestinal and urinary tracts[209,210], however, such methods do pose a real risk of introducing new resistant organisms to the patient with the potential to exacerbate their current circumstances[211]. A more considered approach would be to engineer probiotic cocktails that can be guaranteed to be free of contamination by pathological bacteria. Such an approach would greatly benefit from an improved understanding of the means by which colonisation is achieved by such probiotic bacteria.

1.3.2 *Lactobacillus reuteri* in health and disease

Lactobacillus reuteri is a well-studied Gram-positive bacteria that commonly colonises the gut in numerous vertebrate species[212,213]. Whereas other gut microbes are commonly associated with disease states, *L. reuteri* is considered a symbiotic, pro-biotic bacteria that fundamentally improves the health of its host[214,215]. This is achieved through the production of antimicrobial molecules that inhibit colonisation by pathogenic microbes[216–

[218], modulation of the host immune system to promote T cell development while limiting inflammation by reducing the production of cytokines[216,219–221] and by strengthening the intestinal barrier to prevent microbial penetration of the epithelium that can lead to infections[222,223] such as those caused by *S. enterica*. The reported abundance of *L. reuteri* in humans has been steadily decreasing and has been associated with increased incidence of inflammatory diseases such as irritable bowel syndrome[224]. *L. reuteri* has therefore become of great interest as a measure of maintaining human health and as a potential treatment target for infectious and inflammatory diseases[217,218,220,223].

These benefits are reliant upon successful colonisation by *Lactobacillus* species. Biofilm formation has been found to contribute towards resistance to low pH and bile salts within the gastrointestinal (GI) tract[225]. Additionally, biofilm-related exopolysaccharides of *L. reuteri* have been shown to interfere with the adhesion of enterotoxigenic *E.coli* in the porcine gut[221]. The GI tract contains a mucosal and epithelial layer, both of which bacteria can attach to. The attachment process is necessary for biofilm formation and is mediated by adhesins[226–228].

1.3.3 Protein glycosylation in biofilm formation

While *L. reuteri* is able to colonise a variety of different hosts, this is achieved in a lineage-specific manner with different hosts providing different ecological niches to which the host strains have become adapted. This is shown through the ability of rodent-host strains successfully colonising mice while strains from humans, pigs and chickens were unable to do so[229,230]. In particular, the serine-rich repeat protein (SRRP) surface adhesin of *L. reuteri*$_{100\text{-}23}$ from rats was found to be essential to the colonisation of mice by this strain[230]. SRRP$_{100\text{-}23}$ is secreted through an accessory SecA2-SecY2 pathway[230], homologues of which have been identified in the pig isolate *L. reuteri*$_{ATCC\ 53608}$[228] and the crystal structures of the binding regions of SRRP$_{100\text{-}23}$ and SRRP$_{ATCC\ 53608}$ have been determined[231].

These SRRP proteins undergo post-translational *O*-glycosylation in a strain-specific manner, with SRRP$_{100\text{-}23}$ receiving Hexose-Glc-GlcNAc and SRRP$_{ATCC\ 53608}$ GlcNAc-GlcNAc respectively[232], with the corresponding enzymes responsible arising from their individual Sec2-SecY2 accessory secretion systems. The second glycosylation step was found to be performed by the enzyme GtfC in both cases, with an impressive 97% sequence identity found between the two despite their catalysis of glycosylation using UDP-Glc and UDP-GlcNAc for *L. reuteri*$_{100\text{-}23}$ and *L. reuteri*$_{ATCC\ 53608}$ respectively. This high level of similarity has the potential to be revealing about the binding of donor-substrates in this class of enzyme so is of interest to structural studies. Similar studies on other glycosyltransferases have shown that substrate specificity can be altered through mutation of as few as a single residue[127,233,234].

1.3.4 Structural Features of Glycosyltransferases

All structures of glycosyltransferases (GT) solved to date adopt one of three folds[235], either GT-A, GT-B or GT-C. The GT-A comprises two adjacent βαβ Rossman-like domains and contains a conserved DXD motif within the active site that coordinates a divalent metal ion. The GT-B fold likewise contains two βαβ Rossman-like domains but they instead face each other with an active site formed between the two with no need for a metal ion. The GT-C fold consist of integral membrane proteins that use lipid phosphate-like sugar donors. The GtfC enzymes are predicted to belong to GT family 3, which adopt a GT-B fold[119].

The GT-B fold contains additional domain assignments (see Figure 16). The N-terminal domain is involved in binding the acceptor molecule, in this case the SRRP, and are noted for showing large degrees of variability reflecting diversity amongst the acceptors. The C-terminal domain is involved in the binding of a UDP-sugar donor and is relatively conserved in comparison to the N-terminal domain. Binding of the donor causes a shift from the open to closed conformation of the protein that involves a 10-25^{o} rotation of the N-terminal domain towards the C-terminal domain[236,237]. Metal ions are sometimes required for optimal activity, this has been linked to product release[238,239], which appears to be the case for GtfC as Mg^{2+} ions are implicated in efficient product turnover[232]. The loop preceding the Cα4 helix in this protein fold is associated with donor-substrate binding[236].

A major objective in this thesis is to determine the structures of GtfC from *L. reuteri* strains 100-23 and ATCC 53608. Determining these structures would provide an unparalleled view of sugar-donor substrate specificity in glycosyltransferases, which would be of great interest in the optimisation of probiotics for effective colonisation of the gastrointestinal tract as well as generalists involved in enzyme engineering for industrial applications. The methods used to determine such information are detailed in the following section.

1.4.1 Structural Biology

Structural biology is the study of the 3D arrangement of macromolecular complexes and components such as those formed by DNA or proteins. Of specific interest is the way in which the individual amino acids of a polypeptide chain will arrange themselves relative to one another based upon their position within the chain, the properties of the amino acid relative to those within its local microenvironment, and the effects of other specialised interacting molecules. This arrangement is referred to as the "fold" of a protein and is responsible for the formation of different domains within the protein which are the cause of its functionality. If we can determine the 3D structure of a protein we can therefore compare the fold of that protein to previously characterised proteins to identify domains and thereby formulate a hypothesis as to not only the function of the protein itself but the functions of the different sections that make it up.

Structural models of protein are essential for rational drug design and rational enzyme engineering as they form the basis for the associated

computational docking and molecular dynamic studies. Additionally, a vital step in the drug development pipeline is identification of the mode of action and binding of the drug to the target macromolecule which is often required for optimisation of a compound and as evidence required for by regulators before the compound can be used clinically. Determining the structure of a protein therefore makes it a more attractive potential target for drug discovery provided that it is involved in a relevant or essential biological process, such as antibiotic resistance or cell membrane biogenesis.

X-ray crystallography is one of the three dominant techniques used to study the structure of macromolecular biological molecules as part of structural biology, the other two methods being cyro-electron microscopy (cyro-EM) and nuclear magnetic resonance (NMR) spectroscopy. X-ray crystallography has been the most abundantly successful of these techniques, representing the source of more than 85% of the structures within the protein data bank (PDB)[240]. There are multiple advantages to using X-ray crystallography over other techniques as it can resolve small proteins (<20 kDa) as well as larger macromolecular complexes (>100 kDa), has a proven history of producing high resolution data and is applicable to both soluble and membrane proteins. This makes the technique broadly applicable for the study of proteins with diverse functions, such as those found within a synthesis and transport operon.

In contrast, NMR spectroscopy is not broadly applicable across proteins as the technique becomes increasingly convoluted with increases in the size of the protein, proteins studied must be natively soluble and the process of converting experimental information into a structure included limited automation and if therefore time intensive. The use of NMR spectroscopy in structural biology is therefore typically limited to proteins that are small, soluble and that pose issues for X-ray crystallography either through an inability to grow crystals and/or an inability to understand a structure due to native disorder of the protein.

Cryo-EM is an advanced from of electron microscopy that involves cryogenic cooling of samples to reduce vibrations. Typical instruments of this type can achieve resolutions of 5 Å for large proteins or complexes (> 100kDa) such as ribosomes or whole virus particles. These resolution and size limitations make cryo-EM a specialised technique for gaining more general information about the organisation of proteins within a complex than to help determine the function of a protein contained therein. Recent advances in electron detector usage and image deconvolution algorithms[241] have improved the potential resolution of this method to be competitive with those achieved by X-ray crystallography but access to these more advanced microscopes is limited and remains expensive.

Given the advantages and limitations of these techniques, and the available expertise and equipment at the laboratory of study, X-ray crystallography was chosen as the technique with which to investigate the proteins of interest. NMR spectroscopy was eliminated as the majority of the proteins

studied are too large for structural deconvolution using this technique. Cryo-EM was likewise eliminated due to limited access to the instruments that would be necessary to provide resolution competitive with X-ray crystallography.

1.4.2 X-ray Crystallography

X-ray crystallography is a technique that allows the modelling of the three-dimensional structure of a relatively homogenous compound, composite or macromolecule such as nucleic acids or proteins. Information of the positions of atoms relative to one another within the lattice of a crystal are reconstructed from the diffraction patterns generated by exciting that crystal with monochromatic radiation at X-ray wavelengths. The key stages within this discipline are growth of the crystal, diffraction experiments, data processing and computational reconstruction and refinement.

1.4.3 Growing protein crystals

Protein crystals are highly organised heterogenous aggregates containing protein, solvent and molecules bound to the protein, such as co-factors, binding partners or drugs. Solvent content within a crystal can vary between 26-90%[242]. Crystals are highly variable in size, although protein crystals larger than 200 µm in length are typically considered the most ideal for handling. Unlike inorganic crystals, which frequently grow to cms in scale and are formed through strong ionic bonds, protein crystals are held together by weaker forces, such as hydrogen bonds, salt bridges and Van der Waal's interactions, and are thus extremely fragile.

The two most popular methods used to grow protein crystals are the hanging and the sitting drop vapour diffusion techniques. These methods place a droplet of concentrated protein on a surface and mix it with a droplet of buffered precipitant; the combined droplet is then incubated in a closed system connected to a larger well of the aqueous precipitant. Diffusion of vapour within this environment (evaporation and condensation) causes increasing incremental dehydration of the protein-precipitant droplet towards the precipitant concentration of the well solution. In an ideal crystallisation experiment, the concentration of precipitant will be just below what is needed to cause the protein to precipitate therefore the loss of water will cause [precipitate] and [protein] to increase to a point of gradual precipitation where nucleation of a crystal can occur (see Figure 17) The generation of a crystal will serve to lower [protein] which causes a transition from nucleation to growth and causes elongation of the crystal until [protein] falls to a point where it is no longer saturating the droplet.

Many variables exist which influence the formation of protein crystals. Temperature, pH, precipitant, [protein], additives, ionic strength, time and protein purity are important considerations, to name a few. With so many variables a broad screening approach is needed to assess as many of these factors as possible simultaneously. This is typically done using a specialised liquid handling robot to enhance the reproducibility of the results in combination with commercial sparse matrix screens which contain a larger assortment of buffers, precipitants and salts in predetermined conditions

which have previously producing protein crystals. These crystallisation experiments are sealed and incubated in duplicate at different temperatures and checked daily under a light microscope to identify potential crystal growth. Combined with multiple preparations of soluble protein to enhance protein purity and the requirement of mgs of protein per set of screens this approach rapidly becomes labour intensive.

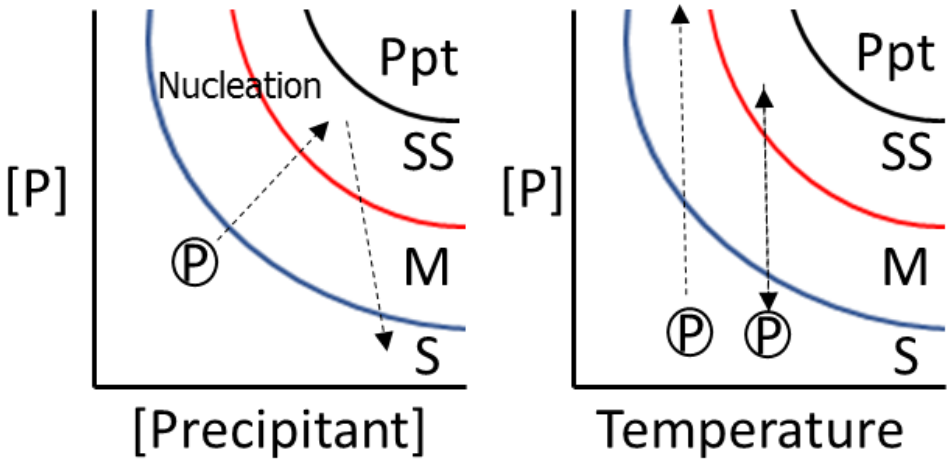

Figure 17: Crystallisation phase diagrams. [P] represents protein concentration, (P) represents the start point of a crystallisation experiment, S, M, SS and Ppt stand for soluble, metastable, supersaturation and precipitate, respectively. Limits of soluble, metastable and supersaturation are indicated by coloured curves. **Left:** At the start of a crystallisation experiment the protein is preferably at a concentration where it is just soluble. As the experiment progresses, diffusion from the droplet steadily increases concentrations of protein and precipitant towards a state of supersaturation. Nucleation of crystals occurs in the metastable and supersaturation states, which decreases the concentration of free protein in the droplet. Nucleation is most common in the supersaturated state but can occur in the metastable. The decrease from supersaturation reduces the incidence of nucleation and instead favours elongation until a soluble concentration is reached. If concentrations of protein or precipitant become too elevated, then the protein will form an aggregate precipitate rather than a crystal. An ideal crystal will reach the nucleate at the metastable/supersaturation boundary as this provides the most room for elongation and reduce nucleation to few crystals so that they can grow larger. **Right:** Temperature has a similar effect upon crystal growth to precipitant concentration but it typically static during a crystallisation experiment. Temperature has an important effect upon when, and if, a droplet will reach a state of supersaturation and nucleate so experiments at multiple temperatures are encouraged.

Once a successful crystallisation condition has been identified that condition can be optimised to generate crystals of a higher standard. This is done by manually producing an array of conditions with slightly altered pH and [precipitant] to produce an optimal crystal for further processing.

Unfortunately, it is not possible to reliable tell through visual inspection how amenable a crystal will be to further experimentation, although larger, fragile crystals with prismatic qualities and smooth or sharp faces are considered a good sign whereas small crystals may be difficult to work with and hard crystals are likely to be salt originating from either the protein's or screen's buffer solution. It is also possible to determine whether a crystal is salt or protein using UV light, as the protein will absorb while the salt will not, or through specialised protein-binding dyes.

1.4.4 Harvesting Protein Crystals

X-ray diffraction experiments typically require the crystals of interest to be removed from the crystallisation experiment and transported to a synchatron light source to be able to produce a high level of diffraction. This can be somewhat problematic as protein crystals are both fragile and unstable. To minimise crystal deterioration by the environment it is preferred to flash-freeze the crystals in liquid nitrogen. This process of rapid freezing has the potential to both damage the crystal and to form ice crystals which can interfere with the experiment; therefore, crystals are typically bathed in a solution based upon the mother liquor including a cryoprotectant compound to limit the formation of ice crystals. Cryoprotectants include molecules such as glycerol, sugars or low molecular weight polyethylene glycol compounds. The choice of cryoprotectant is based upon known literature examples[243] of successfully diffracting crystals drawn from the same or similar mother liquors as well as taking into account whether an existing component of the mother liquor would act as a cryoprotectant.

Once a suitable cryoprotectant mixture is selected, the crystal is removed from the crystallisation plate using a light microscope and µm scale loop and soaked in the cryoprotectant solution for up to 1 minute. Ligands and related compounds may also be added to the cryo solution to hopefully be incorporated within the crystal. Quite often multiple solutions need to be tested as a cryoprotectant will be found to cause the crystal to dissolve or become damaged. Due to the small volumes involved, this process is performed below room temperature to minimise evaporation but above 4°C to prevent the condensation of excess ambient moisture around the liquid nitrogen used in the next step. The cryoprotected crystal is then lifted in a suspension of buffer using a loop and plunged into liquid nitrogen to flash-freeze.

1.4.5 X-Ray Diffraction Experiments

The solution of an atomic structure requires excitation of the crystal with a wavelength of light that is small enough to distinguish between the individual atoms of the excited molecule. For this reason, X-rays are preferentially used here as they fit within the desired resolution with a wavelength of around 1 Å but are not as energetic as smaller waveforms, such as electrons, which damage the sample more rapidly. It is possible to produce x-rays at in-house facilities, however, the intensity and deviation of such sources is low in comparison to specialised synchrotron facilities

which produce beams orders of magnitude more intense and stable and therefore produce a higher quality of spots (smaller and shaper) on the diffraction images. The higher beam intensity also reduces the minimum size of crystal required to around 20µm and reduces the required exposure time to collect a complete dataset. Synchrotron facilities additionally are able to perform diffraction experiments at cryogenic temperatures (100 K) which presents several advantages; the cooling reduces the thermal motion of the sample, which improves the consistency and quality of the data gathered, and decreases the degree to which free radicals are generated during the experiment, thereby limiting damage to the sample which would lower the quality of the data set collected.

In an X-ray diffraction experiment, a crystal-containing loop is extracted from a cryogenically cooled puck. The X-ray beam is manually aligned to the crystal through use of multi-axis goniometer alongside a camera feed before excitation. The sample is then excited with an X-ray beam and rotated, during which time multiple diffraction pattern images are taken. These images are automatically indexed by the ISPyB image suite.

When radiation hits an electron associated with an atom within the crystal the energy of that radiation is absorbed and then reemitted at a different angle than from which it was struck. This phenomenon is known as diffraction. The angle of the reemitted energy can be predicted using Bragg's law:

$$2\mathrm{d}_{\mathrm{hkl}} * \sin\theta = \mathrm{n} * \lambda$$

where d is distance between reticular planes (h, k and l are Miller Indices describing the reticular planes within the crystal), θ is the angle of incident radiation, n is a positive integer value and λ is the wavelength of the radiation (See Figure 18). As θ is measured during the experiment and λ is a controlled variable this allows us to determine the organisation of the crystal lattice, or space group, and the interatomic distances between constituent atoms by solving this equation.

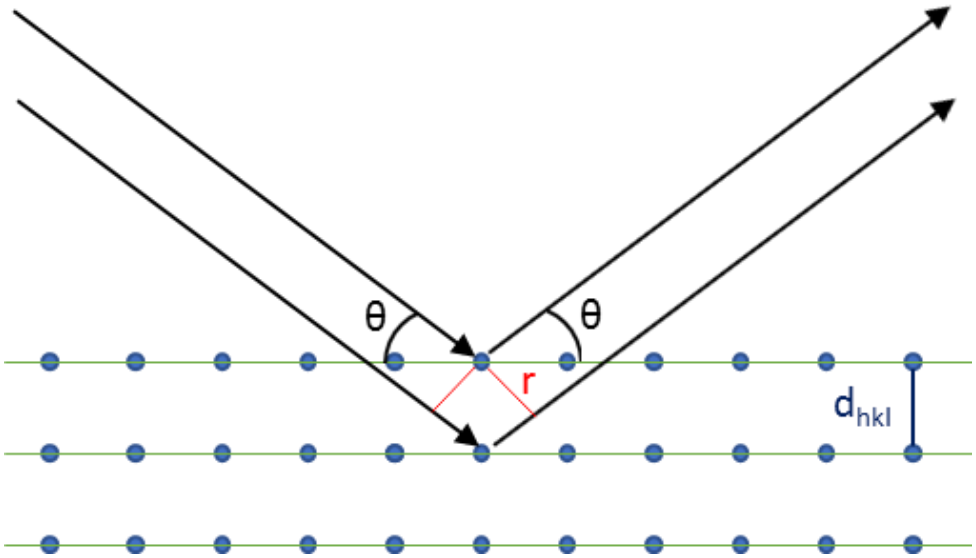

Figure 18: Graphic description of Bragg's law**.** Reticular planes are represented by green lines passing through blue atoms. Distance between reticular planes (d_{hkl}) is shown by a blue line. Incident radiation is indicated by black arrows and its angle shown by θ. 'r' is represented by red lines and is equal to d_{hkl} * sin θ. The extra path of the lower beam, whose length is 2r, corresponds to a multiple of the wavelength λ.

The scattering of an atom is determined by its number of electrons. Its scattering power is described by the atomic scattering factor fj, resulting from the ratio between the amplitudes of the wave scattered by the atom and of the wave scattered by a single electron.

$$f_j(\theta, B) = f_j^0 e^{\left[\frac{-B_j(\sin^2\theta)}{\lambda^2}\right]}$$

The atomic scattering factor f0j decreases with increasing θ because contributions between different regions of the electron cloud tend to cancel as θ increases. It also depends on the temperature factor, Bj, which describes the standard deviation of the atomic position due to temperature-induced vibrations.

The spots on a diffraction image represent the radiation emitted by a family of reticular planes with Miller indices h, k, and l that are in the condition of positive reflection according to Bragg's law. Their intensity is described by the following:

$$I = k\lambda|\mathbf{F}_{hkl}|^2 \frac{V_{cryst}}{V_{cell}}$$

where k is a proportionality coefficient, λ is the wavelength of the incident radiation, Fhkl is the structure factor, Vcryst is the crystal volume and Vcell is the unit cell volume. The structure factor can also be represented in the

complex plane as a vector with magnitude |Fhkl| and phase αhkl of the wave reflected by the family of reticular planes of indices (h, k, l).

$$\mathbf{F}_{hkl} = |\mathbf{F}_{hkl}|e^{i\alpha_{hkl}}$$

The structure factor for each (h, k, l) is also the result of the scattering contribution of all atoms in the unit cell and it depends on the atom types and on their locations in the unit cell. Each single spot therefore contains information on all the atoms in the unit cell:

$$\mathbf{F}_{hkl} = \sum N_j = f_j e^{2\pi i(hx_j+ky_j+lz_j)}$$

the equation above sums the contribution to the structure factor of each jth atom in the unit cell at positions (xj, yj, zj).

A structure factor is the Fourier transform of the unit cell's electron density; to determine the protein's electron density, ρ (x, y, z), the inverse transform of the structure factor in each point of the unit cell needs to be calculated:

$$\rho(x,y,z) = \frac{1}{V}\sum_h\sum_k\sum_l \mathbf{F}_{hkl}\, e^{-2\pi i(hx+ky+lz)}$$

$$= \frac{1}{V}\sum_{hkl} |\mathbf{F}_{hkl}|\, e^{i\alpha_{hkl}} e^{-2\pi i(hx+ky+lz)}$$

where V is the volume of the unit cell. A diffraction experiment can measure the absolute values of the structure factors $\mathbf{F}_{hkl}$, but not the phases α_{hkl}. However, they can be estimated by applying several methods, for example SIRAS/MIRAS, SAD/MAD or molecular replacement. The latter is the method used in this thesis.

1.4.6 X-Ray Diffraction Data Processing

X-ray images are automatically integrated and scaled by DLS servers immediately following collection using a variety of pipelines, including FAST DP[244], DIALS[245], Xia2 3d/3dii[246], multi-Xia2 and autoPROC[247]. These processes vary depending upon the hardware used at the site, as this can change the data format, and are run redundantly in parallel such that the strongest processed image can be selected from the results based upon the crystallographer's validation preferences. Due to the wide variety of methods used, only a general outline for this process is provided below.

The first step of data processing is known as indexing. Here multiple sample images taken at distinct angles are used to identify the spots from diffraction. This information then undergoes a 2D-3D vector transformation based upon Ewald sphere construction. The cell dimensions are identified and then reduced to match to the best fit Bravais lattice. Data is further reduced by detector and crystal parameter refinement and integration of spot intensities via discrimination from background noise and correction. This constricts the raw data into a more manageable form.

The data then undergoes a symmetry check to determine the Laue class (or shape of the unit cell) and the space group. Scaling is applied to reduce experimental error such as originating from instruments or crystal defects. Partial and symmetry-related reflections are then merged to provide a complete reflection dataset. Finally, intensity distributions are analysed to identify potential twinning, B-factors and structure factors.

1.4.7 Model Building

Following data process, the strongest resulting image file is selected for model building. The best results should have higher resolutions (closer to 0), multiplicity and completeness (>90%); the meaning of these statistics will be discussed in the next section.

An issue exists in crystallography commonly referred to as "the phase problem". X-ray detectors measure the intensity and position of an impact but are unable to determine the phase of the wavelength, meaning that this information is lost during the experiment and needs to be determined using another method. Several techniques are able to do this, such as anomalous scattering (MAD/SAD), isomorphous replacement or molecular replacement. The latter method is used in this work.

Isomorphous replacement involves the incorporation of "heavy" atoms within the crystal either through soaking or co-crystallising with iodine or a heavy metal or by expressing the protein in a selenomethionine medium such that the it is incorporated into the protein in place of methionine. Diffraction from native and heavy crystals are compared using the Patterson function within which peak height is proportional to the number of electrons and therefore easily identifies the positions of the heavy atoms, allowing phase and amplitude to be determined geometrically. The downsides of this method are that it requires either additional protein preparation and crystallisation using an expensive and toxic medium or the incorporation of a heavy atom via soaking, which may damage the crystal, and requires multiple well-diffracting crystals for the comparison to take place.

Molecular replacement is a computational technique where a protein with significant predicted homology to the protein of interest, and with an existing structural model, is identified and the phases from one model are used to estimate the phases of the other. The obvious advantage of this method is that it can be done rapidly without additional reagent consumption. The disadvantage is that the determined phases are estimates rather than experimentally determined and will be biased towards similarity with the existing structure. It is of course also necessary for a suitable similar structure to have already been determined, a novel protein may be unsuited to this technique, and any existing errors within the homology model regarding phasing will be carried across.

Before molecular replacement can be applied, the model used needs to be optimised for to correctly integrate with the experimental data. The first consideration is finding the model that has the greatest homology to the

experimental protein. A simple method of doing this would be to identify the protein with the greatest sequence identity to the experimental protein through a BLAST search; the uniprot server[248] is able to do this while screening for proteins that have an already-solved structure. This is a valid strategy if the results returned have extremely high sequence identity (>80%) but we can't rely solely upon the model with the highest sequence similarity for unrelated proteins as non-sequence factors influence protein structures, such as levels of disorder, conformational flexibility or the presence or absence of co-factors. For proteins with no such obvious homologous structure it is necessary to determine a model using other methods, the most applicable of which is the use of hidden Markov modelling alongside existing know protein sequence and folding libraries to predict the secondary and tertiary structure of the protein of interest. One of the foremost options at this juncture is PHYRE2[125], identifying regions within the protein of interest that conform to the parameters of existing domain architecture and scoring the degrees of similarity and confidence of the prediction.

The quality of the structure chosen should be as high as possible (high resolution, low R-free, etc.) to minimise the possibility that the selected model will transfer across errors that will mislead later model building. A model that is predicted by ensembles to match the conformation within the crystals is also preferable and a good determinant when examining potential models with low sequence similarity[249]. The RMSD between ensemble and model can also be improved by removing flexible regions such as loops from the model, however, doing so decreases the coverage of the model and will therefore impact negatively upon signal: noise ratio. When removing potentially large sections of a model in this manner it is important to make sure to preserve sections involved in crystal contacts as these segments are critical is preventing clashes during the molecular replacement and rebuilding processes. It may also be appropriate to prune away sections of the model that significantly deviate from the ensemble prediction or to remove sidechains that lack sequence identity to prioritise the correct assembly of the protein backbone[250]. This can be achieved either manually or using programs such as CCP4 Chainsaw[251] or Phenix Sculptor[252].

Once the parameters of the model are completed it is also necessary to estimate the number of monomers that are present within the experimental data so that the correct number of protein molecules will be inserted. This is done using the Matthew's equation[253]:

$$V_M = \frac{V}{n\ x\ M}$$

where V is the volume of the unit cell ($Å^3$), n is the number of monomers in the asymmetric unit and M is the molecular weight of the monomer (Da). V_M is the Matthew's Coefficient, which should present values within a range of 1.68 – 3.53; the most likely values for n can therefore be estimated based upon the experimentally determined V and the known molecular

weight of the protein and then cross-correlated against the known distribution of Matthew's Coefficients. If multiple possible values for n are present then molecular replacement can be attempted for each value and the best resulting model taken for further work; molecular replacement attempted with an incorrect number of monomers will frequently fail due to steric clashes or lack the necessary crystal contacts to realistically represent a stable crystal.

With the described preparation completed molecular replacement can be attempted. The purpose of this process it to search for a solution that allows the correct positioning of the structure within the experimental unit cell. This depends upon six variables, namely three rotational angles and three translations that define the correct position within the unit cell and phase information, the latter of which is taken from the target model. Performing a six-dimensional search is a time consuming and computationally expensive process that frequently struggles with high symmetry groups, low resolution data, densely packed unit cells or elongated proteins[254], therefore two-step approaches, where angles are determined independently of translations and then recombined, dominate the field.

Two popular programs for molecular replacement are MOLREP[255] and Phaser[256], both making use of two-step approaches. As structure factors cannot be calculated if orientations are not fully defined, they instead rely on approximations using Patterson methods or likelihood methods. Such methods can be described as Fast Fourier Transform methods as they generate all values simultaneously in reciprocal space[257,258] rather than using a "brute force" method that evaluates each transformation against a defined sampling grid in real space. Phaser is the program used for molecular replacement in this thesis.

Patterson methods, like isomorphous replacement, rely upon the Patterson function:

$$p^P(u,v,w) = \frac{1}{V}\sum_{h,k,l} |\, F_{hkl}\,|^2 \cos 2\pi(hu + kv + lw)$$

This method of processing intensity data removes the requirement for phase information from the equation and produces a series of vectors describing interatomic distances, with weight relative to their electron density, equal to the N(N-1), where N is the number of atoms. Patterson vectors are classed as sub-vectors, representing intra-molecular distances, or cross-vectors, representing inter-molecular distances.

In molecular replacement, Patterson functions are calculated from the experimental data and from the model then undergoes a RMSD minimisation to determine molecular orientation with best fit to the model. The Patterson transform for the search model is first generated and then placed in a simple triclinic P1 cell with size optimised to maintain sub-vectors and exclude cross-vectors. The Patterson transform of the experimental data is then generated and then reduced to an area around

the origin to reduce the complexity of the data and exclude as many cross-vectors as possible. The model's function in then rotated against the data's function and correlation co-efficients determined at every angle to determine a best fit. The structure factor amplitudes for several translations for model are then calculated in the orientation of best fit and then compared to the crystal to estimate phase information. Where the asymmetric unit contains multiple monomers, this process is repeated for each monomer and they are positioned in series to prevent overlap. The disadvantage on this method is that it does not provide an estimation of error[258,259].

Maximum likelihood measures are evaluated based upon likelihood scores of observed structure factor amplitudes, the probabilities of which are determined based upon a model's orientation and position. These probabilities can then be compared, and the most likely solution selected to determine the orientation of the crystal's structure. The total structure factor for each reflection is calculated:

$$F(h) = \sum_{j} \sum_{l} f_i \exp\left[2\pi i h \,.\, \left(c_j x_j + d_j\right)\right] = \sum_{j} \exp\left(2\pi i h \,.\, d_j\right) F(h, j)$$

where h is an individual reflection, j is the crystal's symmetry operator, c and d are its rotational and translational parts, x values are fractional co-ordinates and F(h, j) is the molecular transform of the molecule corresponding to the crystal's symmetry operator.

For a given orientation the model is placed sequentially at grid points positioned at translationally unique volumes of the unit cell. For the model, the phase and amplitude of structure factors is known which allows calculation of the total structure factor and determination of the probability of a given orientation and position based upon two-dimensional Gaussian modelling of atomic errors. As the phase in our crystal is unknown, it is integrated out as a nuisance variable to produce a Rice distribution. The model is then rotated sequentially on a grid and the angle with the highest likelihood selected. While exact structure factor sums cannot be determined, due to lack of phase information, the relative magnitude of components can be estimated and expressed statistically as a random walk of components to produce a two-dimensional Gaussian distribution and, though integrating out phase information, a rotational Rice distribution. Rice functions for each reflection are then combined and the highest likelihood across the data is selected.

1.4.8 Refinement

The process of crystallographic refinement is performed using a combination of manual and automated methods to modify the co-ordinates of the model generated from phase-solution of the experimental data. The goal of the refinement is to improve the fit of the model to the detected electron density within the physical restraints provided by factors such as bond length, Ramachandran angles, preferred side-chain orientations and steric hindrance. This is performed manually using the program WinCoot[260]

and in an automated manner using RefMacs (CCP4[249]) and Refine (Phenix[261]). Where gaps appear in the model these can be filled in manually using WinCoot. The primary guiding statistic used during the refinement process is the reduction of R-free, which is discussed in the validation section.

1.4.9 Validation - Crystallography Statistics

There are a wide variety of statistics used in structural biology and crystallography to assess the accuracy and validity of the data and models produced. This section of the thesis will provide explanations and definitions for how these statistics are interpreted and, where appropriate, calculated.

Resolution: Resolution is the distance at which two independent entities can be determined to be distinct. For a protein structure, this distance is measured in Å and is a good indicator for the potential quality of a processed image or a completed model, with a smaller number indicating a greater ability to distinguish between atoms and molecules and therefore presenting a higher resolution and quality. Resolution is determined experimentally directly from the diffraction pattern during image processing. While resolution < 1.0 Å can be achieved, this is relatively rare and typically reflects small proteins or fragments of larger proteins; the median resolution for a structure is around 2.0 Å (see Figure 19). A resolution of 5.0 Å allows the identification of the protein backbone but is insufficient to clearly define side chains; this is frequently referred to as "blobology" due to the unreliability of the fine details of such models. A resolution of 3.5 Å allows identification of bulky sidechains such as those containing aromatic groups; this represents the limit at which a structure can be considered informative. A resolution of 2.5 Å allows identification of all side chains, although there may be some ambiguity; this represents the resolution needed to have confidence in a structural model. A resolution of 1.5 Å or better allows the resolution of individual atoms, as a carbon – carbon bond is this length, and easy identification of water molecules. The better the resolution, the clearer the structure present, the easier it is for a model to be built and therefore the more confident someone can be in their interpretation of that model.

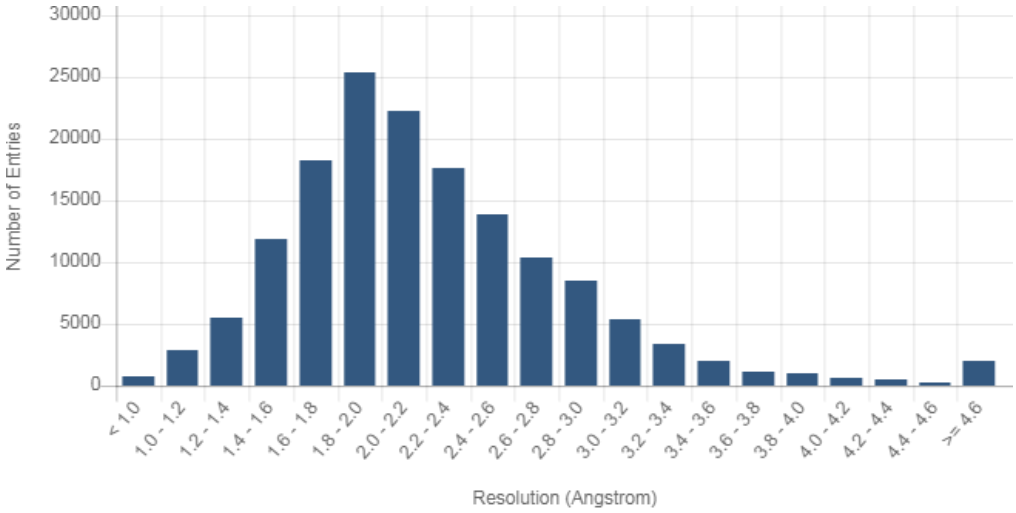

Figure 19: Graphical indication of number of structures in the protein data bank as a product of resolution. Structures are from include all experimental sources. Figures was copied from the protein data bank[240] (01/08/20).

Space Group: This describes the designation of the symmetry of the asymmetric group found within the crystal.

Unit Cell: The unit cell gives the dimensions of the asymmetric units that comprise the crystal.

Total Reflections: The total number of reflections observed within the dataset. The number of reflections will increase with the size of the unit cell and resolution and decrease if the symmetry of the space group is high. A larger number of total reflections is considered beneficial as a sign of higher resolution.

Unique Reflections: The total number of reflections following symmetry averaging. This is a subset of the subset of reciprocal lattice points in the asymmetric unit.

Multiplicity: Dividing the total reflections by the number of unique reflections produces the multiplicity, which reveals the overall redundancy of the dataset.

Completeness: The percentage of the reflections from the asymmetric unit that were successfully collected. A higher completeness indicates that the data covers a larger portion of the asymmetric unit. Low completeness data, even at high resolution, is not representative of the contents of the unit cell. This is calculated as number of unique reflections divided by number of possible unique reflections at this resolution.

Mean I/sigma (I): This is an indication of the signal-to-noise ratio within the dataset. A higher number here indicates less signal interference. A common cut-off for data is 2.0[262] but lower values can still be used. This is calculated as mean intensity divided by mean intensity error.

Wilson B-factor: This is a measure of the degree of order within the crystal. A factor of greater than 50 $Å^2$ indicates significant disorder within

the crystal which implies that large regions of the protein may not be detected or that there is a high solvent content.

R-merge: This measures the agreement between observed and calculated data making use of a subset of redundant reflections that were removed from the model during symmetry averaging. Historically, this was used to determine the cut-off for resolution in a model but increases with multiplicity, which is independent of model quality, therefore is somewhat unreliable[263]. A value below 0.5 is considered ideal. The value is calculated as follows:

$$R_{merge} = \frac{\sum_{\mathrm{hkl}} \sum_{\mathrm{i}} | I_i(\mathrm{hkl}) - \bar{\mathrm{I}}(\mathrm{hkl})|}{\sum_{\mathrm{hkl}} \sum_{\mathrm{i}} I_i(\mathrm{hkl})}$$

where I_i(hkl) represents independent measurements of the intensity of a reflection and $\bar{I}$ (hkl) is their average.

R-meas: This is a modification of R-merge that considers the issue of multiplicity[263] by adjusting by a factor of $\sqrt{\frac{n}{n-1}}$.

R-pim: This is a further modification of R_{merge}[264] which describes the precision of averaged measurements ($\bar{I}$ (hkl)) as a standard error of the mean and is lower than $R_{measure}$ by a factor of $\frac{1}{\sqrt{n}}$.

R-work: This measures the agreement between the observed and computed structure factor amplitudes. It is calculated using the following:

$$R_{work} = \frac{\sum | F_{Obs} - F_{Calc} |}{\sum F_{Obs}}$$

where F_{Obs} are the structure factors measured from the X-ray experiment and F_{Calc} are the structure factors calculated from the model.

R-free: This measures the agreement between the observed and computed structure factor amplitudes using a set of reflections omitted from modelling and refinement which allows the measure to be independent of introduced biases[265]. This value is generally considered to be the most accurate measure of the validity of a structural model[240] and is required to be below resolution/10 to be considered for publication.

CC1/2: Is used to determine whether the resolution limit applied to the data is appropriate. A Pearson's correlation coefficient is performed on the data, which is randomly split in half for the analysis[266]:

$$CC = \frac{\sum(x - (x))(y - (y))}{[\sum(x - (x))^2 \sum(y - (y)^2]^{\frac{1}{2}}}$$

A Student's T-test is then used to determine where statistical significance ends; typically at values of 0.15 or less.

CC*: This is an estimate of the true signal quality and allows a direct comparison of model and data quality by comparison to CC_{work} and CC_{free}[266].

A CC_{work} larger than CC* implies overfitting as it shows a stronger relationship between the experimental data and the model than the data and the signal. Likewise, a CC_{free} below CC* suggests underfitting of data. These statistics are useful in determining the point at which model building should terminate. CC* is calculated as:

$$CC^* = \sqrt{\frac{2CC_{1/2}}{1 + CC_{1/2}}}$$

CC_{work} and CC_{Free} are calculated in the same manner as CC* using the worked and an independent subset of the data similarly to R_{work} and R_{free}.

RMS (Bonds and Angles): Root mean square deviation measure of the bond length and bond angles within the model. The expected values are from measurements of the same parameters in high resolution small molecule structures. An ideal model deviates in bond lengths by less than 0.02 Å and in bond angles by less than 4°.

Ramachandran Plot: The Ramachandran plot describes the allowed Phi and Psi angles within the backbone of an amino acid chain based upon the secondary structure that chain exhibits. Each Phi and Psi angle is categories depending on where it presents on the Ramachandran plot as either favoured, allowed or an outlier. An ideal model will have all angles within the favoured or allowed categories and will have fewer than 2% classified as outliers.

Rotamer outliers: This gives the percentage of amino acid side chains within the model that do not conform to expected low-energy configurations based upon a library of high-confidence side chains drawn from the protein data bank[267]. An ideal model will have no outliers, however, the conformation adopted by side chains is affected by a significant number of variables beyond simple tetrahedral geometry, such as steric influences and atomic attractions, that may be relatively unique and therefore not occurring within the rotamer library, allowing some leeway here.

Clashscore: This indicates the number of too-close contacts between atoms within a model per 1000 atoms[268]. Lower values for this statistic are better, with a score of 30 or less being considered ideal.

Average B-factor: Also known as the temperature factor, this value gives the average uncertainty in the positions of atoms within the model based upon the density of the observed electrons at those positions. This is calculated as below:

$$B_i = 8\pi^2 U_i^2$$

where U_i^2 is the mean displacement of atom i. For a given atom, a value of less than 30 $Å^2$ indicates significant confidence in that position while values in excess of 60 $Å^2$ indicate disorder.

1.4.10 Glycosyltransferases

Glycans are ubiquitous throughout life with numerous biological roles. Relevant examples include adhesin modification for colony formation by GtfC, polysaccharide synthesis by PelF and LPS synthesis by WaaB. The addition and polymerisation of glycans is performed by enzymes named glycosyltransferases (GTs), of which there are currently 111 families[119]. Classification of these enzymes into families is based upon sequence homology, with members of the same family typically sharing fold, mechanism and stereochemistry.

These enzymes catalyse the formation of glycosidic bonds by transfer of a sugar moiety from an activated donor, containing a phosphate leaving group, to an acceptor. Substrate specificity varies wildly between members of the same family and thus needs to be determined experimentally. Donor substrates are typically nucleoside diphosphate sugars (e.g. UDP-glucose, GDP-galactose) but nucleoside monophosphate sugars, lipid phosphate-linked sugars and unsubstituted phosphate are also used by some enzymes[269]. Acceptor substrates vary significantly and include other sugars, proteins, lipids and small molecules, with oxygen, nitrogen, sulfur or even carbon acting as the target's nucleophile[269].

Despite the breadth of enzymes in the family, only three folds are canonically described (see Figure 20): types A, B and C[235,269]. Both the GT-A and GT-B folds consist of two β/α/β Rossman-like domains but differ in that the GT-A fold arranges the domains abutting one another, and generally requires a divalent metal ion coordinated by a semi-conserved DXD sequence within the active site, whereas the GT-B fold has domains facing one another to form a catalytic site in the cleft between the two[128,235,269]. The GT-C fold differs in that it is adopted only by integral membrane proteins and so contains a large span of hydrophobic helices and uses lipid phosphate-linked sugar donors[235].

The two Rossman-like domains found in GTs of the A or B fold are reminiscent of nucleotide binding domains. These N-terminal and C-terminal domains determine the specificity for the acceptor and donor substrates, respectfully. Given the relative variation in acceptor and donor substrates, it is unsurprising that C-terminal donor substrate binding domains show greater conservation than the N-terminal domains. In contrast, GT-C proteins are conserved in their transmembrane region, which is implicated in donor substrate binding, whereas their soluble domains are highly variable and do not conform to Rossman-like folds[235,269,270].

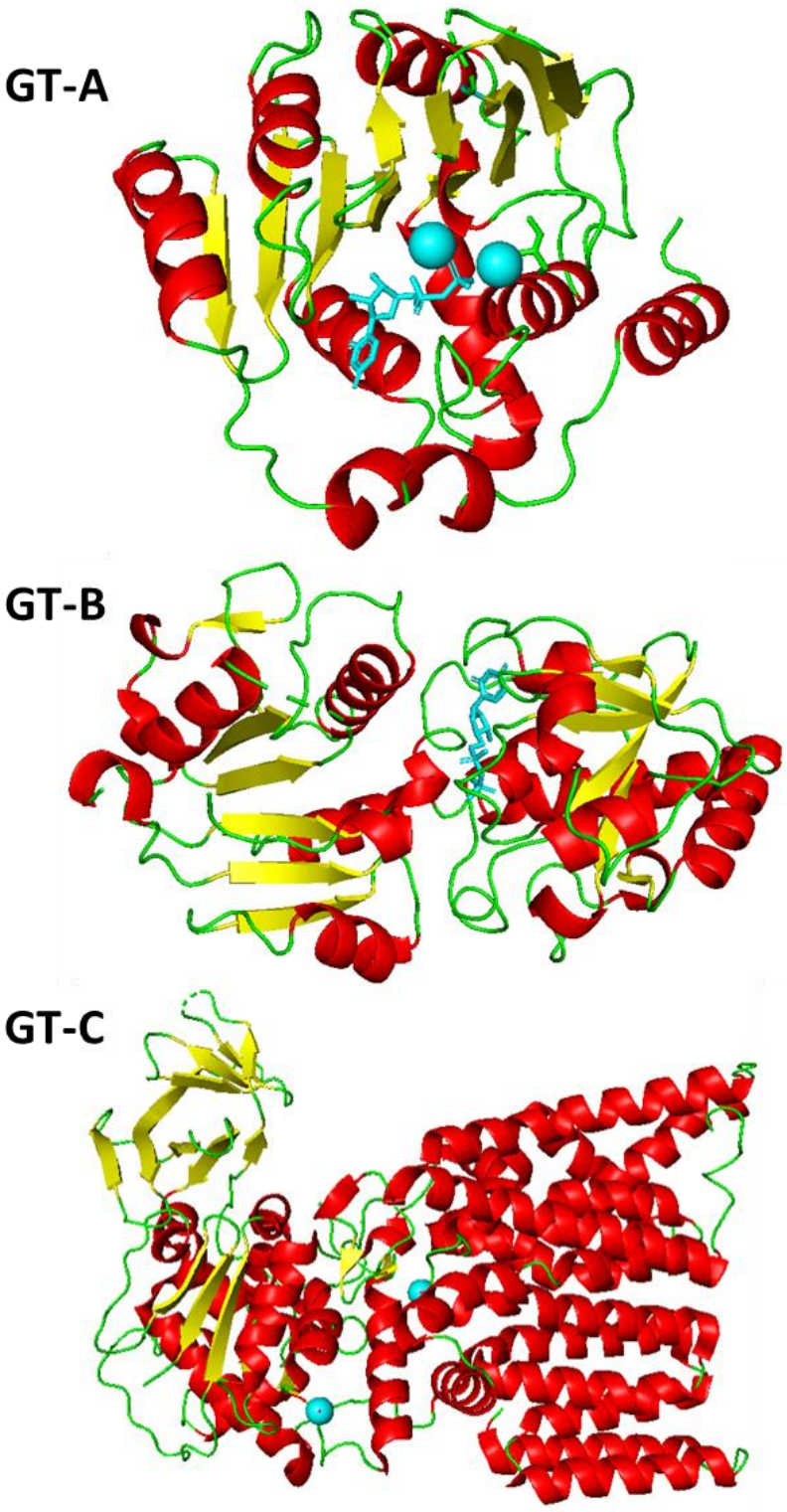

Figure 20: Cartoon representations of the three glycosyltransferase folds**.** Cartoon schematics showing secondary structure with α-helices (red), β-sheets (yellow), loops (green), metal co-factors (balls, cyan) and UDP (sticks, cyan). The GT-A fold consists of two abutting Rossman-like domains, giving the appearance of a single continuous domain, and rely upon divalent metal co-factors. GT-B fold enzymes instead have two facing Rossman-like domains forming an active cleft between them. The GT-C fold contains a

large transmembrane helical domain (right) with a less defined soluble C-terminal domain. Examples used are SpsA from Bacillus subtilis in complex with UDP and magnesium (PDB: 1QGS)[271], T4 phage b-glucosyltransferase in complex with UDP (PDB: 2BGU)[272] and the oligosaccharyltransferase from Campylobacter lari in complex with magnesium and peptide substrate (PDB: 3RCE)[270].

The formation of a glycosidic bond by a GT has two potential stereochemical outcomes with respect to the donor-sugar: inverted or retained. This property is determined by the mechanism used by the individual GT and therefore is conserved amongst members of the same GT family, allowing relatively reliable prediction based upon classification. This system of prediction is not fool-proof as new families are discovered experimentally and GTs reclassified based upon this new knowledge[273,274]. GT-A or B fold enzymes can either invert or retain stereochemistry, while GT-C enzymes have been found to exclusively invert[235,269].

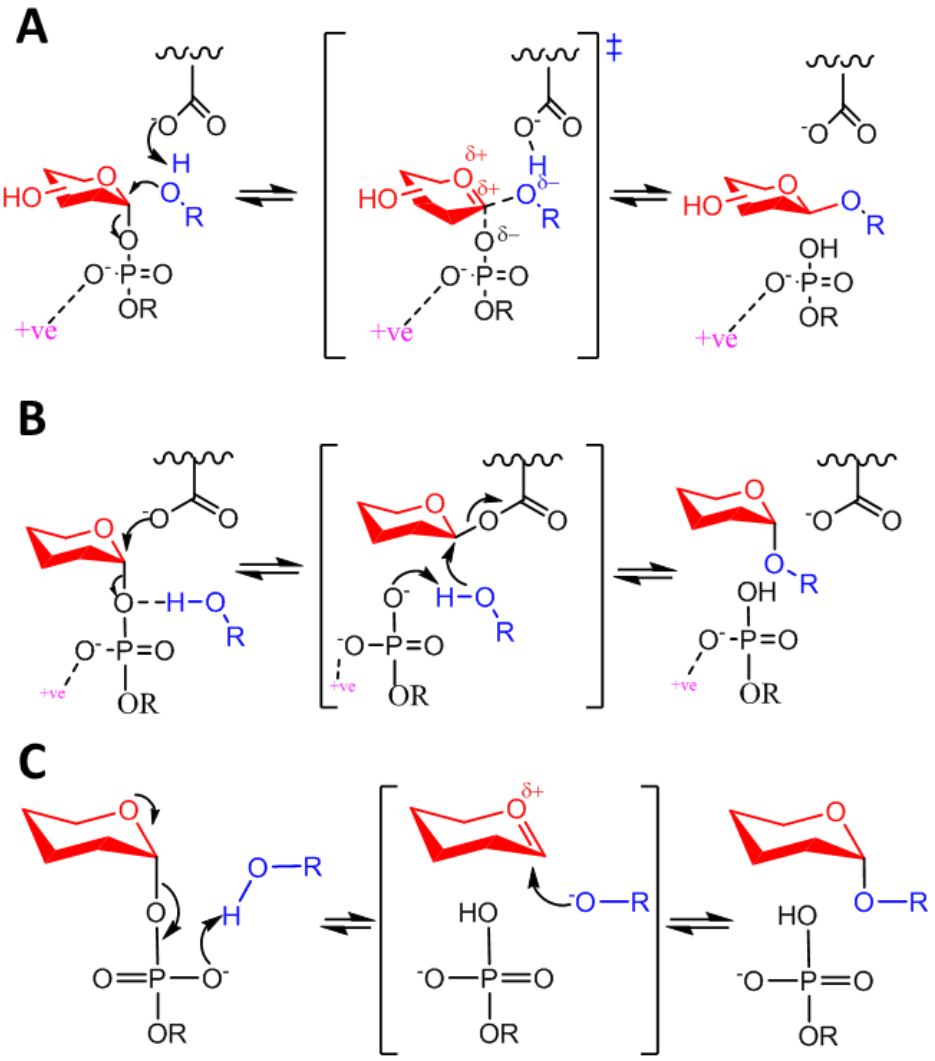

Figure 21: Mechanisms of glycosyltransferases. The stabilising positive charge for the leaving group is provided by a divalent metal ion in GT-A enzymes or a sidechain in GT-B enzymes. Simple hexose sugars (red) are shown for

illustrative purposes. 'R' represents residual extension of the molecule. Acceptor substrate shown in blue. Stereochemistry of the linkage of the anomeric carbon is indicated. **A: S_N2-like reaction of inverting glycosyltransferases. B: Historic double-displacement reaction of retaining glycosyltransferases**. Evidence supporting the existence of the glycosyl-enzyme intermediate has not been found. **C: Same-side S_Ni reaction of retaining glycosyltransferases.** This more recent model suggests no direct catalysis by the enzyme and that retaining glycosyltransferases achieve catalysis through subtle control of microenvironments[275].

Inverting glycosyltransferases employ a direct displacement S_N2-like reaction (see Figure 21A)[269]. A proton is extracted from the acceptor substrate by a general base residue to initiate nucleophilic attack of the donor-substrate's phosphate-binding carbon. The enzyme-substrate complex then enters an oxocarbenium ion-like transition state before dissociation into enzyme, organic product and phosphate leaving group. The leaving group is mediated by the co-ordinated metal ion in GT-A enzymes or a positively charged side chain in GT-B enzymes.

Retaining glycosyltransferases have been proposed to employ a double-displacement mechanism (see Figure 21B)[269]. The donor-sugar undergoes nucleophilic attack by a general base residue which displaces phosphate/pyrophosphate to form a glycosyl-enzyme intermediate. The displaced phosphate/pyrophosphate abstracts a proton from acceptor substrate to form a leaving group and initiate nucleophilic attack of the glycosidic bond of the intermediate at the glycosyl-carbon. Once more, the leaving group is mediated by the co-ordinated metal ion in GT-A enzymes or a positively charged side chain in GT-B enzymes. No evidence of the glycosyl-enzyme intermediate has been found, which has led to the proposal of a same-side S_Ni-like reaction mechanism (see Figure 22C) for retaining GTs. This mechanism proposes that a proton is abstracted from the acceptor-substrate by the phosphoryl leaving group which primes the acceptor for nucleophilic attack of the anomeric carbon while generating no covalent intermediates. This model suggests that microenvironment of the active site contributes towards the reaction without the need of general acids or bases and is better supported by the literature[275].

1.4.11 Glycoside hydrolases

A glycoside hydrolase (GH) can be considered the opposite of a glycosyltransferase: where one joins two substrates together with a glycosidic bond, the other breaks down a glycosidic bond to form two substrates. GH enzymes are therefore involved in almost all aspects of life. Enzymes of this superfamily have been organised into 168 sub families, based upon sequence homology, and 18 clans, based upon fold[119]. These clans include six variants of an $(\alpha/\alpha)_6$ domain, five of $(\beta/\alpha)_8$, two β-jellyrolls, two 5-fold β-propellers, one clan of 6-fold β-propellers and one clan adopting a β-helix[119] (see Figure 22).

Typical members of the GH superfamily adhere to a general acid/base catalytic mechanism with either stereoselection for inversion or retention[276]. The detailed catalytic mechanism is conserved in members of the same clan[277] and generally conserved within GH families. Exceptions do, however, exist as some families use both or even divergent mechanisms[276,277].

In the inverting mechanism, the catalytic base abstracts a proton from water, causing nucleophilic attack of glycosidic carbon and abstraction of a proton from the catalytic acid by the glycosidic oxygen, breaking the glycosidic bond. The retaining mechanism is a similar general acid/base catalysis with direct nucleophilic attack of the glycosidic carbon by the enzyme inducing proton abstraction from the general acid by the glycosidic oxygen, breaking the glycosidic bond and forming a glycosyl-enzyme intermediate. The general acid residue then regenerates by abstracting a proton from water, activating it to perform nucleophilic attack on the anomeric carbon and breaking the bond with the enzyme and regenerates the general base.

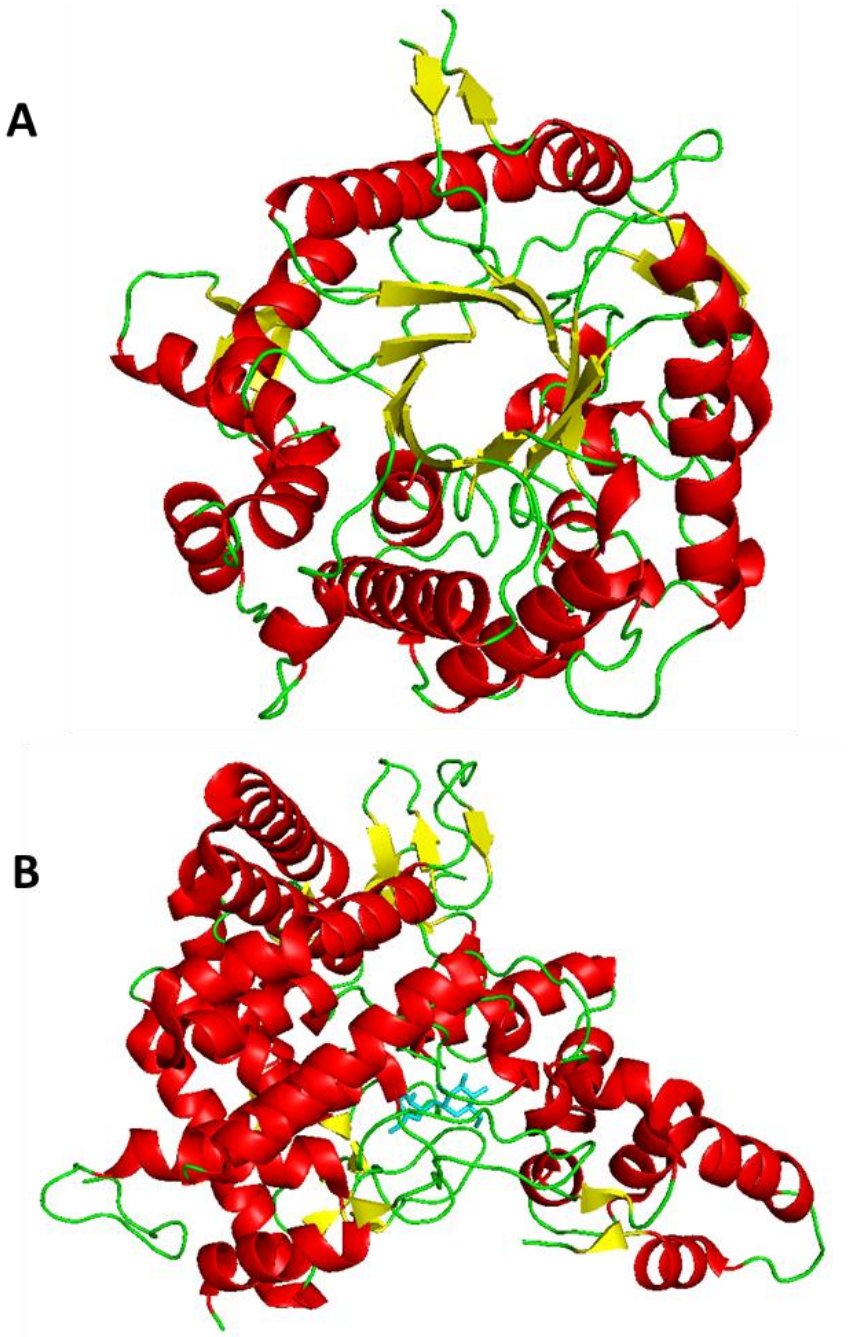

Figure 22: Cartoon representations of two major glycoside hydrolase folds. *Cartoon schematics showing secondary structure with α-helices (red), β-sheets (yellow), loops (green) and validoxylamine A (sticks, cyan).* ***A: The $(\beta/\alpha)_8$ TIM barrel fold.*** *This ubiquitous fold forms a core of 8 β- sheets surrounded and stabilised by 8 solvent-exposed α-helices. The catalytic site is formed at the more open face by the β-core with peripheral stabilising residues contributed by the surrounding helices.* ***B: The $(\alpha/\alpha)_6$ toroid barrel fold.*** *The barrel is formed by 6 sets of paired antiparallel helices staggered to form a hydrophobic core (left). The active site is present at the open face of the barrel (where the inhibitor is bound). Examples used are beta-glucosidase A from Bacillus polymyxa (GH family 1, PDB: 1BGA)*[278] *and trehalase from E. coli in complex with validoxylamine A (GH family 37, PDB: 2FJ4)*[279]*, respectively.*

Potential enzymatic nucleophiles are generally easy to identify from structural data as they will be located within the active cleft, are generally polar, hydrogen bond with a water molecule and are typically conserved[280]. Structures containing the enzyme's substrate are particularly helpful in

identifying residues important to both substrate binding and to catalysis based upon proximity; fluorinated substrate analogues can also provide snapshots of the start of the reaction and in cases where enzymatic activity would prevent the substrate from being present in the structural model[281]. In the case of structures not containing a catalytic ligand, docking experiments can be used to determine the identity of the catalytic nucleophile where the structure of the ligand is known[282,283].

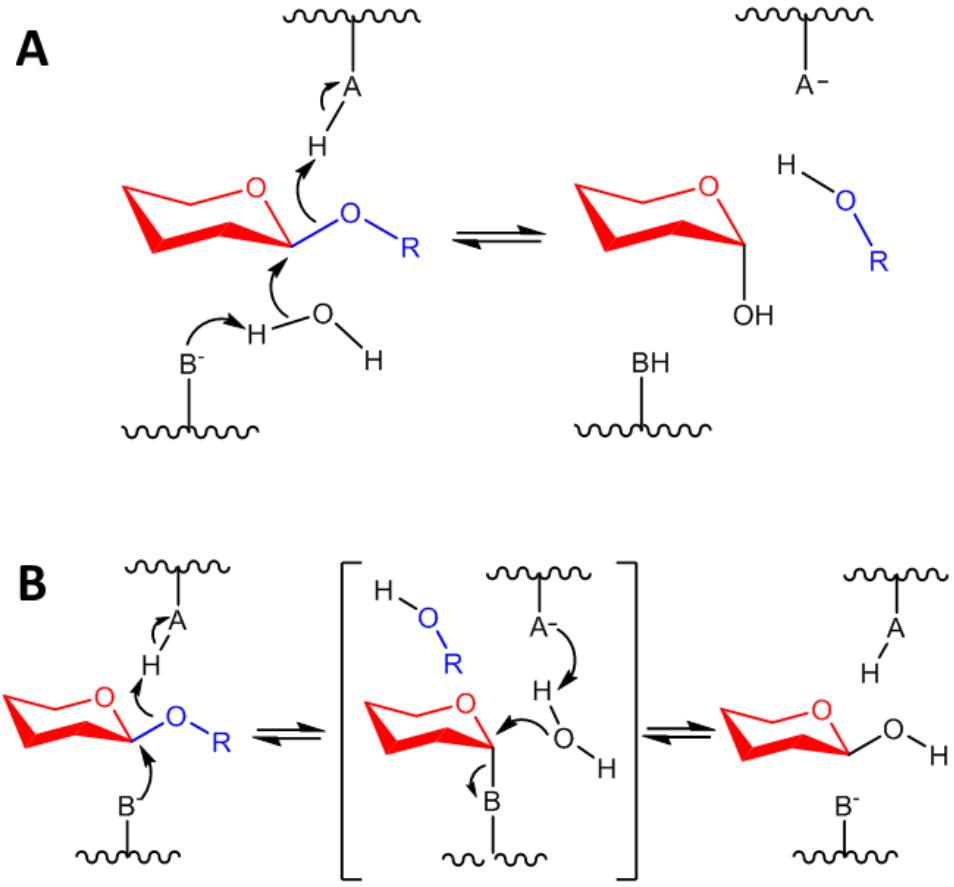

Figure 23: General acid-base reaction mechanisms of inverting (A) and retaining (B) glycoside hydrolases. Simple hexose sugars (red) are shown for illustrative purposes. 'R' represents residual extension of the molecule. Acceptor substrate shown in blue. Stereochemistry of the linkage of the anomeric carbon is indicated. A and B represent catalytic acid and base, respectively.

Despite these advantages, it remains necessary to confirm the proposed activity of a residue, usually via mutagenesis coupled with a biochemical assay of activity or binding. A major contributing factor to this is that enzymes frequently undergo significant conformational change upon binding which can result in residues previously distinct from the active site to be arranged into a catalytic formation[284–286].

1.5.1 Aims and Objectives

This thesis is focused upon three independent projects under the wider umbrella of combating antibiotic resistance via the generation of structural models of proteins either contributing towards antibiotic resistance, to

further our knowledge of their functions and provide models forming the basis for rational drug design, or that contribute towards a protective effect against antibiotic resistant organisms, to further our knowledge of their functions and form a basis for enzyme engineering to enhance their protective effects. The first project is the structural and functional characterisation of the proteins of the *pel* operon of *Pseudomonas aeruginosa* which generates the biofilm-forming exopolysaccharide pel and contributes towards antibiotic resistance. The second project is the structural and functional characterisation of the lipopolysaccharide synthesis glycosyltransferase WaaB from *Salmonella enterica* serotype Typhimurium, which performs a modification of lipopolysaccharide which improves cellular resistance to antibiotics. The third project is the structural and functional characterisation of the adhesin-modifying glycosyltransferase GtfC from *Lactobacillus reuteri* which allows species-specific colonisation of the gut where the bacterium acts as a potent probiotic providing colonisation resistance and as a means of displacing colonisation by antibiotic resistant microbes.

1.5.2 The Pel Project

The objectives for the pel project are:

1. To clone the proteins of the pel operon into plasmids and cells lines for protein expression.
2. To trial the expression of these protein to determine whether they can be expressed solubly
3. To determine optimal expression conditions and cell lines for highest yields of soluble protein.
4. To purify the proteins to sufficient quality and quantity for use in crystallisation experiments.
5. To optimise the protein and purification procedure to produce high quality diffracting crystals.
6. To use crystallised protein to determine the crystal structure of pel proteins.
7. To use the determined protein structure to identify likely functions for the protein and assess if these coincide with bioinformatic analysis of these proteins.
8. To confirm the hypothetical functions of the protein through functional assays.

Work towards aims 1-3 is outlined in Chapter 2 of this thesis while remaining work towards these aims is outlined in Chapter 3. Completion of these goals will provide powerful tools for pharmacological development of means to attenuate the antibiotic resistance provided by the pel polysaccharide of *P. aeruginosa*.

1.5.3 The WaaB Project

The objectives for the WaaB project are:

1. To use the crystal structure of WaaB to determine functional residues.

2. To clone mutations into the identified residues.
3. To express and purify WaaB mutants.
4. To perform glycosyltransferase assays with the WaaB mutants to assess their importance.
5. To optimise, express and purify mutants of particular interest to produce high quality diffracting WaaB crystals.
6. To use crystallised protein to determine the crystal structure of WaaB mutants of interest.
7. To determine the reason for the activity of interest.

The work contributing towards our understanding of WaaB is outlined in Chapter 4. Completion of these goals will provide powerful tools for pharmacological development of means to attenuate the antibiotic resistance provided by the outer membrane of *S. enterica* serotype *Typhimurium*.

1.5.4 The GtfC Project

The objectives for the GtfC project are:

10. To crystallise the provided samples of $GtfC_{100\text{-}23}$ and $GtfC_{ATCC\ 53608}$
11. To use crystals of GtfCs to determine their crystal structures.
12. To use the structure of the enzymes to determine functionally important sites within the enzyme.
13. To use the structures of the enzymes to identify the structural features leading to their different substrate specificities.
14. To present this data to our collaborators (Juge Group, Quadrum Institute, UK) so that they can perform functional assays to test our hypothesises.

The work contributing towards our understanding of the GtfC proteins is outlined in Chapter 5. Completion of these goals will provide a structural basis for sugar-donor substrate recognition by glycosyltransferases, which will be of great interest in the realm of synthetic biology, and improve our understanding of the species-specific adhesion process in probiotic bacteria that is vital to colonisation resistance and to treatment of antibiotic resistant infections of the gastrointestinal tract.

Chapter 2 – Methods

2.1 General Methods

This section will detail the typical methodology used to perform experiments; specific experiments will outline how their methodology deviated from the typical.

2.1.1 Cloning

The purpose of cloning during this research is to insert a gene of interest into a recombinant plasmid and to then transform said plasmid into a strain of *E. coli* for the purposes of either replicating the DNA or for expression of the gene of interest. This thesis uses a variety of cloning techniques and as such a standard protocol cannot be provided, however, some specific details of the cloning can be standardised and a summary of those is provided here:

2.1.1.1 Plasmids

For expression of a protein of interest to be successful, it needs to be cloned onto a suitable expression plasmid. The plasmids used in this work are summarised in Table 1.

Table 1: Plasmids

Name	Selection	Function
pBAD24	Ampicillin	L-arabinose controlled expression, optimised to minimise leaky expression, no tag.
pDEST17	Ampicillin	Gateway system IPTG-inducible expression plasmid, contains an N-terminal His_6 tag
pDONR207	Gentamycin	Gateway system transfer plasmid
pEHisTEV	Kanamycin	IPTG-inducible expression plasmid, N-terminal TEV protease cleavable His_6 tag
pET22b (+)	Ampicillin	IPTG-inducible expression plasmid, N-terminal PelB signal peptide sequence
pLou3	Ampicillin	IPTG-inducible expression plasmid, N-terminal TEV protease cleavable His_6-MalE tag

Plasmids selection was based upon the predicted solubility of the protein of interest. The pEHisTEV, pDEST17 and pLou3 plasmids were used for the expression of proteins, or sections of proteins, that were predicted to be cytoplasmic or periplasmic. Each contains an N-terminal His_6 tag that was cleavable either through a plasmid-incorporated Tobacco Etch Virus protease site or through a PCR-incorporated 3C protease site. This allows selection of the protein during purification using Ni^{2+} affinity chromatography. The pLou3 plasmid also contains a cleavable N-terminal

maltose binding protein (MBP) tag, which has been shown to improve protein solubility[287] and allows for selection using amylose affinity chromatography. A disadvantage of this tag is that MBP will often dimerise with native MBP which often results in an MBP contaminant. The pBAD24 and pET22b (+) plasmids are used for the expression of membrane proteins; pBAD24 helps to minimize leaky expression, which can have toxic effects upon cells in the case of membrane proteins, while pET22 (+) contains an outer membrane protein signal peptide that has been shown to be highly functional in *E.coli*. Both of these plasmids lack a purification tag so a C-terminal His_{10} tag was incorporated into the proteins via PCR.

2.1.1.2 Cell Lines

All cloning and expression of proteins in this thesis took place using recombinant *E.coli* cell lines (see Tables 2 and 3).

Table 2: *E.coli* strains

Holding Strains

Strain	Genotype
DH5α	*F– endA1 glnV44 thi-1 recA1 relA1 gyrA96 deoR nupG purB20 φ80dlacZΔM15 Δ(lacZYA-argF)U169, hsdR17(rK–mK+), λ–*
TOP 10	*F- mcrA Δ(mrr-hsdRMS-mcrBC) Φ80lacZΔM15 Δ lacX74 recA1 araD139 Δ(araleu)7697 galU galK rpsL (StrR) endA1 nupG*
XL-10 Gold™	*endA1 glnV44 recA1 thi-1 gyrA96 relA1 lac Hte Δ(mcrA)183 Δ(mcrCB-hsdSMR-mrr)173 tetR F'[proAB lacIqZΔM15 Tn10(TetR Amy CmR)]*

Expression Strains

Strain	Genotype
BL21 (DE3)	*B F– ompT gal dcm lon hsdSB(rB–mB–) λ(DE3 [lacI lacUV5-T7p07 ind1 sam7 nin5]) [malB+]K-12(λS)*
BL21 (DE3) pLysS	*B F– ompT gal dcm lon hsdSB(rB–mB–) λ(DE3 [lacI lacUV5-T7p07 ind1 sam7 nin5]) [malB+]K-12(λS) pLysS[T7p20 orip15A](CmR)*
BL21 (DE3) Rosetta2™	*B F– ompT gal dcm lon hsdSB(rB–mB–) λ(DE3 [lacI lacUV5-T7p07 ind1 sam7 nin5]) [malB+]K-12(λS) pLysSRARE[T7p20 ileX argU thrU tyrU glyT thrT argW metT leuW proL orip15A](CmR)*
BL21 (DE3) Rosetta2-gami™	*Δ(ara-leu)7697 ΔlacX74 ΔphoA PvuII phoR araD139 ahpC galE galK rpsL (DE3) F'[lac+ lacIq pro] gor522::Tn10 trxB pLysSRARE (CamR, StrR, TetR)*

OverExpress™ C43 (DE3)	*F – ompT hsdSB (rB- mB-) gal dcm (DE3)*
HDB150	*MC4100 ompT::spc ΔaraBAD leuD::kan*
SoluBL21™ (DE3)	*F- ompT hsdSB (rB-mB-) gal dcm (DE3)†*

Holding strains represent strains which are optimised for competency and are used to amplify the amount of available plasmid through replication. DH5α and TOP 10 cells are general purpose holding strains while XL-10 Gold™ are hyper-competent cells reserved for procedures where transformation success rate may be lower, such as from site directed mutagenesis.

Expression strains include chromosomal additions that allow for specific overexpression of the gene of interest by the T7 RNA polymerase enzyme, as well as modifications to allow that strain to express more effectively under specific circumstances. BL21 (DE3) represents the baseline expression strain with limited modifications. The pLysS addition helps to restrict leaky expression which can aid in protein expression by limiting cellular toxicity of proteins of interest until the logarithmic phase of growth where protein expression is induced. Rosetta2™ cells contain an additional plasmid which encodes additional copies of tRNAs that are rare in *E.coli*; this allows efficient expression of proteins from heterologous organism that have different codon biases, such as *Pseudomonas aeruginosa*, as expression of genes containing rare codons would normally be limited by the availability of these tRNAs and may lead to premature translation termination. Rosetta2-gami™ combines the advantages of the Rosetta2™ cell line with a more reducing cytoplasmic environment to encourage correct disulphide bond formation; this is an important consideration for proteins with multiple cysteines that would generally be localised into the periplasmic compartment but are being expressed in the larger cytoplasmic compartment as the cytoplasm is normally an oxidising environment which can lead to incorrect disulphide bond formation, including intermolecular disulphide bond formation between separate copies of the protein of interest. Overexpress™ C43 cells and HDB150 cells are optimised to reduce the toxic effects of expressing integral membrane proteins. SoluBL21™ cells are optimised for expression of soluble proteins via inactivation of proteases that would interfere with expression.

Table 3: Genotype Nomenclature

Nomenclature	**Explanation**
F-	Does not contain the F plasmid
F'[]	Carries the F plasmid with [chromosomal genes]. Can mate with F- via conjugation

ahpC	Mutation to alkyl hydroperoxide reductase conferring disulphide reductase activity
araD139	mutation in L-ribulose-phosphate 4-epimerase blocks arabinose metabolism
CmR	Chloramphenicol resistant
dcm	cytosine methylation at second C of CCWGG sites exist
deoR	Regulatory gene allowing large plasmid uptake
endA1	Endocnuclease 1 removed; cleaner DNA preparations
galE	Mutation blocks UDP-galactose production causing incomplete LPS synthesis associated with high competency
galK	Cannot metabolise galactose and resistant to 2-deoxygalactose
galU	Cannot metabolise galactose
glnV44	Suppression of amber (UAG) STOP codon by insertion of glutamine; phage growth requirement
gor52	Mutation in glutathione reductase which enhances disulphide bond formation
gyrA96	DNA gyrase mutation providing nalidixic acid resistance
hsdSB(rB–mB–)	Allows cloning of DNA without cleavage by endogenous restriction endonucleases
hsdR17(rK–mK+)/ hsdRMS	EcoK1 removed. Does not digest non-methylated (heterogenous) DNA but still methylates DNA.
hte	Allows high transformation of large plasmids inserts
Δ lacX74	Deletion of the lac operon
lon	Lon protease mutant
mcrA	Mutation eliminating restriction of DNA methylated at the sequence $C^{m}CGG$
mcrBC	gene encoding an endonuclease which cleaves DNA containing methylcytosine on one or both strands. McrBC does not act upon unmethylated DNA
mmr	mutation in methyladenosine-specific restriction
	system which allows more efficient cloning of DNA containing methyladenines)
nupG	deoR equivalent

ompT	mutation in outer membrane protein protease VII, reducing proteolysis of expressed proteins
phoA	Abolished activity of alkaline phosphatase
pLysS	Plasmid carrying chloramphenicol resistance and phage T7 lysozyme, effective at attenuating activity of T7 RNA polymerase, for better inhibition of expression under non-induced conditions.
pLysSRARE	Plasmid contains tRNA genes for rare codons (AGG, AGA, AUA, CUA, CCC, and GGA)
purB20	adenylosuccinate lyase mutation resulting in growth rate limitation in M9 media
proAB	Proline auxotroph
pvuII	Abolishes activity of restriction enzyme
recA1	DNA repair deficiency to prevent recombination of inserted DNA.
relA	relaxed phenotype; permits RNA synthesis in absence of protein synthesis
rspL	Ribosomal mutation grants streptomycin resistance
tetR	Tetracycline resistant
thi-1	Thiamine auxotroph
Tm10	Transposon carrying resistance marker
trxB	Mutation is thioredoxin reductase which enhances cytoplasmic disulphide bond formation
φ80dlacZΔM15	φ80 prophage carrying defective lacZ required for blue/white screening
Δ(araleu)7697	Leucine auxotroph
Δ(lacZYA-argF)U169	Hydrogen peroxide resistance from chromosomal deletion.
λ(DE3 [...])	Lambda prophage carrying T7 RNA polymerase genes and lac promoter and repressor
λ−	Lambda lysogen deletion

2.1.1.3 Polymerase Chain Reaction (PCR)

PCR is used to amplify the DNA of the gene of interest for cloning purposes. Unless otherwise specified, PCR was performed using the Q5 ® Hot Start Polymerase (New England Biosciences) kit and their recommended protocol. Primers used can be found in Table A1 in the

appendix. Pel genes were amplified from *P. aeruginosa* PA01 genomic DNA, or a plasmid containing the gene where it had been cloned from genomic DNA already.

2.1.1.4 Restriction Digestion

Ligation dependent cloning requires that overhangs are introduced to the amplified gene and plasmid to produce "sticky ends" that can then be ligated together to insert the gene into the plasmid. Restriction sites that are not present in the gene of interest, and that complement appropriately positioned restriction sites in the target plasmid, are incorporated onto primers at either side of the gene. Following amplification, the PCR product is digested using the pair of complementary restriction enzymes to produce the desired "sticky ends". This was performed using Fast Digest™ restriction enzymes (New England Biosciences) as per the manufacturer's protocol.

2.1.1.5 Agarose Gel Electrophoresis (GE)

Agarose gel electrophoresis is a technique where DNA samples are loaded onto a gel and then are separated by size during then migration through the gel matrix due to an applied current. In ligation dependent cloning, this is used following restriction digestion to produce a DNA product with suitable 5' and 3' overhangs that can then be ligated into a plasmid without interference from leftover complementary DNA and template. In ligation independent cloning, this technique is used to assess the size of the PCR product to ensure that the gene of interest has been amplified correctly.

0.8 - 1% w/v agarose gels were made in 100 mL TAE buffer and microwaved to dissolve the agarose, partially cooled and then allowed to set with ethidium bromide added as a dye. 4 µl samples of reactions were combined with 1 µl 5x loading dye then loaded alongside 2µl GeneRuler® 1kBP plus DNA Ladder (Thermo Fisher) and run at 110 V for 1 hour. Gels were imaged under UV light using a Sygene™ G: Box® Gel Doc and GeneSys® software.

2.1.1.6 Ligation

The ligation procedure is an enzymatic reaction in which the complementary 5' and 3' overhangs of restriction digested PCR product and plasmid are joined to form a circular piece of DNA. Ligations were performed at 10 µl scale using the NEB Quick Ligation™ kit at a 3:1 molar ratio between PCR product and plasmid at room temperature for 15 minutes following the manufacturer's instructions.

2.1.1.7 Transformation and Plating

Heat shock transformation was used to insert the modified plasmid into the cells of interest. Cells (50µl) and DNA (5µl) were incubated together for 30 minutes on ice then underwent heat shock at 42 °C for 45 seconds become being returned to ice for 2 minutes. 450 µl sterile LB or SOC media was added to each tube and they were incubated at 37 °C for 45 minutes. Samples were centrifuged at 3000 x g for 5 minutes. 400 µl supernatant was removed and remaining 100 µl was plated on LB agar plates containing

appropriate antibiotics. LB agar plates were incubated overnight at 37°C then single colonies from each plate were inoculated to 10 mL LB media containing appropriate antibiotics and grown overnight at 37 °C, 200 rpm.

2.1.1.8 Plasmid Extraction

Plasmids were first transformed into holding strains of *E.coli* for plasmid amplification; these plasmids required extraction before they could be cloned into expression strains of *E. coli*. 10ml overnight cultures of transformed bacteria were centrifuged at 4000 x g for 5 minutes then extraction was performed as per the manufacturer's instructions for either the GeneJet® (ThermoFisher) or GenElute® (Sigma-Aldrich) Miniprep kits.

2.1.1.9 Cloning Validation

Extracted plasmids require validation to confirm that the cloning procedure has been successful. For ligation dependent cloning, 100ng of plasmid underwent digestion with the restriction enzyme pair used in cloning and was separated using 0.8% agarose gel electrophoresis to examine for the presence of inserts. Plasmids producing two bands, showing both insert and plasmid at correct molecular weights, were sent for commercial Sanger sequencing (Source Bioscience) to provide further confirmation of successful gene insertion and to assess the sequence for errors. For ligation independent cloning methods, the restriction double-digestion step was skipped as restrictions sites were not built into the primers.

2.1.2 Protein Production

The process of protein production includes the growth of recombinant *E.coli* and their induction to overexpress the protein of interest once they have reached the logarithmic phase of growth. Following this, the protein must be extracted via cell lysis and then purified from other proteins using chromatographic methods. The requirements for protein crystallisation are for protein to be produced in large quantities (mgs) and at high purity (>95%), requiring large volumes for cell culture and purification.

2.1.2.1 Cell Culture

Validated plasmids containing the gene of interest are transformed into expression strains of *E.coli*, plated and grown overnight on a plate and then inoculated to LB media and grown overnight, as described in transformation and plating. For standard protein overexpression, the volume of LB inoculated overnight is 500 ml. Appropriate antibiotic selection is applied at all stages (ampicillin, 100 µg/ml; chloramphenicol, 50µg/ml; kanamycin, 50 µg/ml).

Each 1 L of the day culture is inoculated with 35ml of the overnight culture, for a total of 420 ml of overnight culture into 12 L of LB media. The day culture is incubated at 37 °C, 200 rpm, until an optical density (600 nm) of 0.6 – 1.0 is achieved; typically taking 3-4 hours. Overexpression of the protein of interest is then induced by adding an inducing agent (0.1 mM IPTG, or 0.1% w/v L-arabinose for pBAD24 vectors). For membrane proteins, the culture is induced after reaching an optical density of 1.0 or higher due to the increased stress overexpression of these proteins causes

on the culture. Once induced, cell cultures continue to be incubated at 20 °C overnight to allow expression of the protein of interest.

Following the incubation, cells are pelleted using centrifugation at 3994 xg for 20 minutes at 4 °C. Cell pellets are transferred to a beaker, sealed with aluminium foil and then stored at -20 °C for use within the next six months.

2.1.2.2 Glycerol Stocks

Glycerol stocks are for long term storage of cells that are used for inoculation purposes so that fresh transformations are not required every time a protein needs to be expressed. Glycerol stocks are made at the same time as a day culture, if one has not already been made, from leftover overnight culture; 1mL of grown culture is flash frozen in 2 aliquots with 2mL 50% w/v glycerol in liquid nitrogen and stored at -80 °C for future use. Future inoculations are performed using a 200 µl pipette tip, which is scraped into the frozen cells and then brought into contact with the inoculating media.

2.1.2.3 Protein Expression Trial

When expressing a recombinant protein for the first time it is advisable to determine optimal conditions for the expression of that protein as suboptimal conditions can cause the protein yield to be low, for the protein to have toxic effects upon the cell or for the protein to be expressed in an insoluble state. The conditions that have the most impact, and can be easily controlled, are the strain, plasmid, incubation temperature, the length of time cells are overexpressing protein and the concentration of inducer added to cause overexpression.

Expression trials use the same protocol as general cell culture for *E. coli*, except scaled down to 100 mL per condition during the day culture and the variation of independent variables. Plasmids and cell strains are outlined in Tables 1 and 2, respectively. Incubation temperature refers to the temperature following induction of overexpression with the temperatures 37, 30, 20 and 16 °C most commonly used. Expression time is either four hours or overnight (16-22 hours). For IPTG, the concentrations tested are 1 µM to 1 mM tested in orders of magnitude.

2.1.2.4 Protein Purification

Recombinant proteins expressed in bacteria are present in the cell amongst the host of proteins that that bacteria normally produces. For structural studies, it is necessary to obtain the recombinant protein at as high a level of purity as possible, therefore it must be separated from these other proteins. The same is also true, to a lesser extent, for biochemical studies. The recombinant proteins purified in this thesis all make use of immobilised metal affinity chromatography (IMAC), which specifically recognises the recombinant protein through a histidine tag located at either the N or C terminus of the protein, as the first step of the purification. Additional steps are then used to remove contaminants that were able to bind to the IMAC column, namely the removal of the histidine tag using a sequence specific protease in combination with a second affinity chromatography step and a

final size exclusion chromatography step to isolate the protein of interest based upon its molecular weight. Outlined below is the general protocol used.

The cell pellet is resuspended in lysis buffer (50mM Tris pH 8.0, 300 mM NaCl, 30 mM imidazole, 10% w/v glycerol, 1 cOmplete™ EDTA-free protease inhibitor tablet, 10 ng/mL DNase) as the pellet defrosts over a period of two hours using a magnetic stirrer. The resuspended cells are lysed twice using a Cell Disruptor (Constant Systems) at 30 kPSI and then clarified at 38,434 xg using a JA25.50 rotor for 30 minutes at 4 °C.

The supernatant, containing the soluble fraction of the cell lysate, is applied to a pre-equilibrated 5 mL HisTRAP FF column (GE Healthcare) which is then washed with 10 column volumes (CVs) of Wash Buffer (20 mM Tris pH 8.0, 300 mM NaCl, 30 mM Imidazole, 10% w/v glycerol). The protein is eluted from the column with 3 CVs of Elution Buffer (20 mM Tris pH 8.0, 300 mM NaCl, 300 mM Imidazole, 10% w/v glycerol) then buffer-exchanged using a HiPrep 26/10 Desalting Column (GE Healthcare) intp SEC Buffer (2 0mM Tris pH 8.0, 150 mM NaCl, 10% w/v glycerol). The protein concentration of the eluate is estimated using a NanoDrop™ (Thermo Fisher) and then an appropriate protease (Tobacco Etch Virus protease for pEHisTEV/pLou3 or C3 protease for pDEST14) is added to the solution at a 1:100 mass ratio. This protein solution is left overnight at 4 °C with gentle rocking (~50 rpm) to allow the affinity tag to be removed.

The digested protein solution is centrifuged at 3000 xg for 10 minutes to pellet aggregate protein and then applied to a pre-equilibrated 5mL HisTRAP FF column (GE Healthcare), to which the tag and previous contaminants should bind. The flow-through (FT) is collected and concentrated using a VivaSpin 20 Centrifugal Concentrator (Sartorius) at 4000 rpm, 4 °C to a volume of 5 mL. The concentrated sample is then applied to an AKTA™ PURE system (GE Healthcare) using a 10 mL Superloop (GE Healthcare) at 4 °C and passed through a 16/600 HiPrep Superdex 200PG column (GE Healthcare) equilibrated with SEC buffer to separate the protein of interest, protease and any aggregate. Fractions containing pure protein of interest are pooled concentrated using a VivaSpin 20 Centrifugal Concentrator (Sartorius) at 4000 rpm, 4 °C to a desired concentration. Concentrated protein is either used immediately or flash frozen in liquid nitrogen and stored at -80 °C. Samples are mixed to a final concentration of 33% w/v glycerol before being flash frozen for the cryoprotectant properties of that chemical.

The purification is assessed co-currently by monitoring the UV of the sample flow using the AKTA™ instrument and at the end of the purification through sodium dodecyl sulphate polyacrylamide gel electrophoresis (SDS PAGE).

2.1.2.5 Sodium Dodecyl Sulphate Polyacrylamide Gel Electrophoresis

SDS PAGE is similar to agarose GE in that samples are loaded onto a gel matrix and then migrate through it based upon an applied charge to allow separation based upon molecular weight. SDS PAGE is used for protein samples rather than DNA and includes SDS as a negatively charged denaturant to limit the effects of that the native shape and charge of the protein would have upon its migration.

SDS PAGE in this thesis is performed using 6-12% Tris-Glycine Novex® SDS PAGE gels. 30 µl of samples are combined with 10 µl 4x SDS PAGE loading dye, boiled for 5 minutes then 20 µl is run at 165 v, 200 mA for 38 minutes. The gel is then removed and soaked in Coomassie Instant Blue™ dye for 15 minutes before being destained in water. An initial inspection of the gel may take place after 30 minutes, although this level of destaining may be insufficient to resolve less abundant contaminants. Following inspection, the gel is destained in fresh water for a further 24 hours before the gel is imaged under white light using a Sygene™ G: Box® Gel Doc and GeneSys® software.

2.1.3 Crystallography

2.1.3.1 Crystallisation Experiments

Crystallisation experiments make use of the sitting drop vapour diffusion technique. The concentrated protein is screened against a standard array of sparse matrix screens including PEG Ions 1 and 2, Crystal Screen Cryo, Index, PEG RX, Salt RX (Hampton Research), JCSG Plus, Structure Screen, Pact Premier, Morpheus, Midas Plus and LMB Screen (Molecular Dimensions). A Crystal Gryphon (Art Robbins Instruments) liquid handling robot is used to distribute sample and screen into Swissi polystyrene MRC 2-drop crystallisation plates (Molecular Dimensions). Plates contained 0.5µl droplets at a 1:1 ratio of protein:well solution and were incubated at 21°C and 16°C.

The incubated trays were checked for crystal growth daily during the first week of incubation using a light microscope. Conditions that were found to produce crystals underwent optimisation. An 5x5 array of conditions based upon the crystal-producing condition are produced covering a pH range of -0.4 to +0.4 of the original condition's pH along one axis and -5% to +5% of the condition's precipitant concentration along the other axis; the central condition is therefore a repeat of the original. Crystallisation experiments are then set up as previously using the new optimisation array and at the temperature of the original condition.

2.1.3.2 X-ray Diffraction Experiments

Grown crystals were harvested from crystallisation plates using mounted LithoLoops (Molecular Dimensions) appropriate to the size of the crystal. Crystals were harvested at the temperature of incubation and then either flash-frozen in liquid nitrogen or added to a 2 µL droplet containing the mother liquor and suitable cryoprotectant (20-30% w/v glycerol or 20-30% w/v PEG 400) and allowed to equilibrate for 30 seconds before being

recollected and flash frozen in liquid nitrogen. Cryoprotectant were selected by comparison with cryoprotectants used for existing crystals[243]. As multiple crystals would be harvested at once, multiple cryoprotectant conditions would be screened, as well as a crystal without a cryoprotectant.

The frozen crystals were transported to Diamond Light Source (UK) Synchatron beamline station i03 or i04. X-ray diffraction data was collected remotely from the beamline using a wavelength of 0.9795Å and exposure time of 0.05 seconds with 360° of rotation. The resulting diffraction pattern was automatically processed on the Diamond server pipeline.

2.1.3.3 Structural determination and model building

The crystal structures determined in this thesis determined phases using molecular replacement. A suitable model for molecular replacement for each protein was determined based on the highest similarity existing structure as identified using PHYRE2[125]. A Matthew's analysis (Phenix[261]) of the diffraction data was used to estimate the number of monomers within the asymmetric unit and then the identified structure was used as a model for molecular replacement (Phaser[261]).

Following molecular replacement, the resulting structure was visually assessed for major errors using Coot[260], and where found these were removed. Where appropriate, an expected ligand of the protein was added to the model. The model then underwent multiple iterations of refinement using Refine (Phenix[261]) until minimal improvements in R-free or divergence between R-free and R-work were observed. The model was then manually reassessed using Coot to rebuild problematic sections, as highlighted by aberrant crystallographic statistics, and then returned to further automated refinement. This process continued until either no issues could be identified with the structure, refinement failed to improve the structure or R-free and R-work began to diverge by more than 0.1. The resulting structural model was sent to Professors Andrew Hemmings and Changjiang Dong for review and, once no further issues were identified, waters and molecules from the mother liquor, buffer and cryoprotectant chemicals were added to appropriate unfilled electron densities.

2.2 Pel Operon Methods (Chapter 3)

The methods described in this section are based upon on the basic protocols described in 2.1 with specific details for experiments that took place in place in Chapter 3.

2.2.1 Cloning of the Pel Operon

Proteins from the Pel operon of *Pseudomonas aeruginosa* PA01 were analysed by a variety of bioinformatics prediction servers to determine sections of the proteins that would be most amenable to cloning and expression, including PSIPRED[288], JNET[289], Prof[290], Coils[291], HMMTOP[292], MEMSAT-SVM[293], PHOBIUS[294], PREDISI[295], DISOPRED2[296], IUPRED[297] and InterProScan[137]. Full length genes of PelA (1-948), PelB (1-1192), PelE (1-329), PelF (1-501) and PelG (1-456) were selected for cloning to test whether the full-length protein was amenable to expression. In addition,

PelA 288-946 and PelC 20-172 were also selected based upon the predicted signal peptide and N-terminal domain of PelA and the signal peptide and lipation site of PelC as signal peptide and lipation sites might cause problems with protein expression. The N-terminal domain was removed in the PelA 288-946 clone to test the solubility of the C-terminal domains.

Genes were amplified from *Pseudomonas aeruginosa* PA01 genomic DNA using the standard protocol for ligation dependent cloning. Predicted soluble proteins (PelA 288-948, PelC 20-172, PelF 1-506) were ligated into pEHisTEV and pLou3. Predicted membrane proteins (PelA1-948, PelB, PelD, PelE, PelG) were ligated pBAD24 plasmid with a C-terminal octahistidine tag added by the reverse primer during PCR. Ligations used TOP 10 cells.

2.2.2 Expression Trials of Pel operon proteins

The purpose of these trials was to determine whether the cloned Pel proteins would express and identify an optimal temperature for expression. Positively sequenced plasmids containing Pel proteins were transformed into *E.coli* C43 cells using antibiotic selection. Single colonies were collected and individually inoculated to 10mL LB media (100 µg/mL ampicillin) and incubated overnight 37 ^{o}C, 200 rpm. 1mL of grown culture was used to make glycerol stocks while the remaining 9 mL was used to inoculate a further 100mL antibiotic LB culture. Expression was induced using either L-arabinose for pBAD24 vectors or IPTG for pLou3 vectors and left for up to 22 hours at 20, 30 or 37 ^{o}C before harvesting through centrifugation at 3994 xg for 15 minutes at 4 ^{o}C. The cell pellets were then frozen at -20 ^{o}C.

A single step Ni^{2+}-IMAC purification was performed to provide samples for analysis, as all of the cloned proteins contained a corresponding affinity tag that should allow these proteins to be specifically extracted in detectable amounts. Frozen pellets were resuspended in 20 mL Lysis Buffer A (20 mM Tris pH 8.0, 0.150 mM NaCl) then lysed using sonication (4 cycles of 30 s). Soluble proteins were centrifuged at 38,434 xg for 30 minutes to clarify soluble and membrane fractions. Membrane proteins were instead centrifuged at 4000 rpm for 20 minutes to pellet cell debris and then the supernatant was ultracentrifuged at 40000 rpm for 30 minutes to clarify soluble and membrane fractions. The ultracentrifuge supernatant, containing the soluble fraction, was discarded then membrane pellet was resuspended in Lysis Buffer B (20 mM Tris pH 8.0, 0.15 mM NaCl, 1.5% w/v DDM) through brief sonication (1-2 seconds) and 30 minutes of rotary stimulation. The resuspended membrane fraction was then centrifuged at 13,300 rpm on a benchtop microcentrifuge.

1.5 mL of supernatant fractions of both soluble and membrane proteins were applied to 50µl Ni^{2+}-NTA Superflow beads in a microcentrifuge filter/collection tube, incubated for 5 minutes then centrifuged at 12,000 rpm. These columns were then spin-washed with 500 µl Wash Buffer A (20 mM Tris pH 8.0, 0.15 mM NaCl, 30 mM Imidazole) to remove impurities and then 100 µl Elution Buffer A (20 mM Tris pH 8.0, 0.15 mM NaCl, 500 mM Imidazole) for soluble proteins, and identical buffers containing 0.15%

w/v DDM for membrane proteins, to elute protein of interest from the column. The presence of protein was assessed by SDS PAGE.

The initial trial used the available Pel proteins PelA1-948, PelA288-948, PelE1-329 in B and PelC20-172 and a miscloned PelF1-501 in pLou3 as a control (expressing the tag only). PelB, PelD and PelG were not successfully cloned at this point and issues were encountered with growing pEHisTEV vectors. This trial induced recombinant expression at OD 1.2 using 0.02 % L-arabinose or 0.1 mM IPTG, as appropriate; cells were then left to grow at 20 °C for 22 hours. This trial was then repeated, including PelG1-456 in B, with cells grown at 37 °C for 20 hours to determine whether a higher temperature would increase recombinant protein yield. The two most promising candidates, PelA288-948 and PelE1-329 in B, were then expressed at 6 L scale (inoculating 15/100 mL to each 1 L). Further trials were performed with the potential positives PelA288-948 in B and PelC20-172 in L, the first investigating response to different concentrations of inducer (0.002%-0.02% L-arabinose and 0.01-0.1 mM IPTG) and temperature (20-37°C) while the second assessed expression over time (1-16hrs) in response to different temperatures at the 2 L scale.

2.2.3 Expression Trial of PelA (46-948) and PelF (1-506) pLou3

This trial was to determine whether pLou3 PelA 46-948 and PelF 1-506 constructs expressed protein. 10mL LB precultures containing appropriate antibiotics (100 µg/mL ampicillin and 50 µg/mL chloramphenicol) were inoculated with *E.coli* and grown O/N at 37 °C, 200 rpm. These included PelA 46-948 pLou3 and PelF 1-506 pLou3 in *E.coli* BL21 (DE3) cells of strains pLysS, Rosetta 2 and Rosetta 2-gami. A control of pLysS cells with no plasmid (or ampicillin) was also used. All cells were induced with 1 mM IPTG and left to grown for 16 hours at 18 °C, 200 rpm. 90 mL of each culture was pelleted in 50 mL falcon tubes at 4000 xg for 10 minutes and then stored at -20 °C.

Cells were resuspended in 20 mL lysis buffer (PelA: 50 mM Tris pH 8.0, 0.3 M NaCl, 10% v/v glycerol, 1 cOmplete™ EDTA-free protease inhibitor tablet, DNase. PelF: the same at pH 7.0) and lysed via sonication at 15% 3mm probe amplitude for 1 minute. The sonicator's probe was cooled in ice slurry for 2 minutes between runs. 100 µl Ni-NTA Superflow beads (QIAGEN) was added to 1 mL collection tubes and centrifuged at 12,000 rpm using a benchtop centrifuge for 1 minute to remove storage solution. 1.5 mL clarified supernatant was added to each tube in two 0.75 mL batches and centrifuged at 12,000 rpm using a benchtop centrifuge for 1 minute. 10 column volumes (500 µl) of wash buffer (lysis buffer containing 20 mM imidazole, no DNase) was applied to each column and centrifuged as previously to remove contaminants. 2 column volumes (100 µl) of elution buffer (wash buffer containing 250 mM imidazole) was then applied and centrifuged as previously to elute protein of interest. Membrane fraction pellets were resuspended in 1 mL lysis buffer then sonicated for 5 seconds to break the pellet. Samples of cell lysate, resuspended membrane pellets and column elution underwent SDS PAGE as previously described.

2.2.4 IPTG Expression Trial of PelA (46-948) pLou3

The purpose of this expression trial was to determine an optimal concentration of IPTG with which to induce overexpression of cells. This was performed as per the expression trial of PelA (46-948) except as follows. Culture size was scaled up to 1 L and inoculated with 5 mL preculture of PelA46-948 plou3 Rosetta 2. Cultures were induced with 1 µM, 50 µM, 100 µM, 500 µM or 1000 µM IPTG upon reaching the desired OD. Cell pellets were transferred immediately to 50 mL falcon tubes and resuspended in 15 mL lysis buffer. Following resuspension, cells were lysed using a Cell Disrupter (Constant Systems) at 30 kPSI.

2.2.5 IPTG Expression Trial of PelF (1-506)

The purpose of this expression trial was to determine an optimal concentration of IPTG with which to induce overexpression of cells, and whether treatment with 25% w/v trehalose would improve the solubility of PelF 1-506. 1.5mL of preculture was inoculated to seven flasks of 100 mL antibiotic LB and protein overexpression was then induced in duplicate cultures at 1, 10 and 100 µM IPTG. Half of the cultures were incubated at 37 °C, 200 rpm for 4 hours and then harvested. The remaining duplicates were incubated at 20 °C, 200 rpm for 20 hours and then harvested. The seventh flask was "induced" with 100 µl of filtered distilled water as a negative control and was incubated at 37 °C, 200 rpm for 4 hours and then harvested. Cells were harvested to form 2 pellets per culture using equal volumes by centrifugation at 4000 xg for 10 minutes into 50 mL falcon tubes.

Half the pellets were resuspended in 10 mL lysis buffer (50 mM Tris pH 7.0, 0.3 M NaCl, 10% v/v glycerol, 1 cOmplete™ EDTA-free protease inhibitor tablet, DNase) and the remaining half were resuspended in lysis buffer containing 25% w/v trehalose. Cells were then sonicated and treated as described for the previous PelF expression trial. Samples were taken of the soluble fractions and an elution fraction combined from all samples.

2.2.6 ArcticExpress and SoluBL21 Expression Trials of PelF (1-506) pLou3 and pEHisTEV

These trials were performed to determine whether the solubility of PelF 1-506 could be enhanced by use of different cell strains. PelF (1-506) pLou3 and pEHisTEV were transformed into ArcticExpress (DE3) RP Competent cells (Agilent) and SoluBL21™ Competent cells (AMSBIO) of *E.coli* using the previously described heat shock method. ArcticExpress cells have been optimised for growth at lower temperatures with additional chaperones to improve protein folding and reduce insolubility. SoluBL21™ have been optimised to improve protein solubility.

For SoluBL21™ cells, 1.5 mL of precultures were each inoculated to two flasks containing 50 mL LB containing respective antibiotics and grown at 37 °C, 200 rpm. Protein overexpression was induced using 0.1 mM IPTG. One flask was incubated at 37 °C, 200 rpm for four hours then harvested

and the other was incubated at 20 °C, 200 rpm for 20 hours before harvesting.

ArcticExpress (DE3) RP (AERP) cells containing PelF (1-506) pLou3 or pEHisTEV were grown at 37 °C, 200 rpm overnight in 10 mL LB containing either 100 µg/mL ampicillin or 50 µg/mL kanamycin, respectively, and 20 µg/mL gentamycin. 600 µl of cells for each plasmid were subcultured into eight 30 mL LB flasks containing no selection antibiotics and incubated at 30 °C, 200 rpm for 3 hours. Cultures were transferred to 10 °C and 13 °C at 200 rpm evenly and left for 10 minutes to equilibrate. Protein overexpression was induced using 0.1 and 1.0 mM IPTG in duplicate flasks for each plasmid. One duplicate was harvested after 4 hours and the remaining one after 20 hours.

Harvested SOL and AERP cells were resuspended in 1 mL (50 mM Tris pH 7.0, 0.3 M NaCl, 10% v/v glycerol, 1 cOmplete™ EDTA-free protease inhibitor tablet, DNase) and sonicated for 20 seconds, with the probe cooled in ice-slurry for 1 minute 30 s between runs the clarified at 15,000 rpm for 30 minutes at 4 °C using a benchtop centrifuge. Samples of soluble fractions were taken for SDS PAGE.

2.2.7 Cloning of PelBC

PelB and PelC were not successfully cloned using previous cloning methods; having the DNA synthesised bypassed this issue. Additionally, they were identified as forming a complex which might stabilise the proteins expressed. The sequence for PelBC was synthesised and cloned into the pET22b(+) commercially by GenScript®. This sequence involved removal of the PelB signal peptide sequence for compatibility with the PelB (different protein) signal peptide encoded by the plasmid. A ribosome binding site sequence was placed between PelB and PelC (1-172). A hexahistidine tag was added to the end of PelC. A 5' EcoR1 site and 3' HindIII site were added to PelB and C respectively to allow further sub cloning. PelBC pET22b(+) was transformed into TOP10, C43 and HDB150 competent *E.coli* cells.

2.2.8 Expression Trials of PelBC

The first trial was to determine whether the PelBC pET22b(+) construct successfully expressed the proteins. A larger volume expression trial was used to account for the smaller volume of the outer membrane space in comparison to the cytoplasmic space. 500mL LB containing 100 µg/mL ampicillin was inoculated from the glycerol stock of PelBC pET22b(+) C43 and incubated overnight at 37 °C, 200 rpm. 35 mL of preculture was inoculated to 12 flasks of 1 L LB containing 100 µg/mL ampicillin and incubated at 37 °C, 200rpm. Once an optical density of 1.0 was achieved, protein overexpression was induced with 0.2 mM IPTG and temperature was reduced to 20 °C. Culture was harvested after 20 hours at 5000 xg for 30 minutes then stored at -20 °C.

Cell pellet was resuspended in TBS buffer (50 mM Tris pH 8.0, 0.3 M NaCl, 5% w/v glycerol, 10 mM imidazole) with small amounts of DNase and lysozyme for 1 hour 30 minutes. Sample was lysed using a Cell Disruptor (Constant Systems®) at 30 kPSI then centrifuged at 15000 rpm for 20 minutes at 4 °C using a JA25.50 rotor. Supernatant was removed and ultracentrifuged at 45000 rpm for 1 hour. Pellet was collected and resuspended in TBS buffer containing 1% w/v N-dodecyl-β-D-maltopyranoside (D310) over two hours. Sample was ultracentrifuged at 45000 rpm for 1 hour and supernatant loaded to a 5mL HisTRAP FF column. Column was washed with 10 CV TBS buffer containing 0.04% w/v N-Dodecyl-Beta-D-Maltoside (DDM) then 3 CV of TBS buffer containing 0.04% w/v DDM and 250 mM imidazole. Eluted protein would then be concentrated to a maximum of 5 mL using a VivaSpin® 20 centrifugal concentrator (Sartorius) and injected through a HiPrep 16/600 Superdex 200PG column using SEC buffer (20 mM Tris pH 8.0, 150 mM NaCl, 0.04% w/v DDM, 5% w/v glycerol).

A second expression trial was necessary due to unavoidable disruption of work during the previous purification. Second trials were conducted using a 150 mL preculture of PelBC pET22b(+) HDB150 into 4 flasks of 1L LB antibiotic. Cultures were induced with 0.1 mM IPTG with one serving as a control with no IPTG. Cultures were grown at 37 °C for four hours, 20 °C for 20 hours or 16 °C for 20 hours. 1 mL samples of each were taken and pelleted at 5000 xg for 20 minutes then stored at -20 °C. Samples were resuspended in 80 µl TBS buffer with 20 µl loading dye and boiled for 5 minutes. Samples were sonicated for 2 seconds using a 3mm sonicator probe then centrifuged at 12000 rpm for 5 minutes to reduce the viscosity of the samples before loading 20 µl samples for SDS PAGE as previously described.

2.2.9 Purification of PelA 46-948

PelA 46-948 pLou3 was expressed in BL21 (DE3) Rosetta2 *E.coli* cells using the standard cell culture protocol, except that the culture was harvested and frozen as two equal portions. The cells were resuspended, lysed and purified as per the standard protocol, except that one pellet was resuspended in lysis buffer additionally containing 0.75 M trehalose. Samples were taken of the clarified soluble and membrane fractions of both pellets, to determine whether the additive improved relative soluble yield of PelA, before soluble fractions were pooled for the remainder of the purification.

Fractions deemed pure were pooled and divided into two lots then concentrated to 10 and 27.5 mg/mL. Concentrated protein was screened against the standard array of crystallisation screens. Remaining protein was flash frozen in liquid nitrogen using 30% w/v glycerol as a cryoprotectant.

Expression and purification of PelA 46-948 pEHisTEV Rosetta2 used the same protocol as for PelA 46-948 pLou3 Rosetta2 except that ampicillin was exchanged for 50 µg/mL kanamycin and samples were not split to trial

the effect of trehalose as an additive. Pooled fractions of D12-D4 (68-81.5 mL) and D3-E7 (81.5-95 mL) were each trialled for crystallisation as previously described.

A subsequent purification used the same protocol except that a gradient elution from a HisTRAP FF column (GE Healthcare) was performed over 5 CV on an AKTA Prime (GE Healthcare) instrument at 4 °C rather than being eluted isocratically with 3 CV. The next purification used a 30 CV elution gradient for a HisTRAP column on an AKTA Pure (GE Healthcare) instrument at 4 °C and replaced the HiPrep Superdex 200PG column (GE Healthcare) step with a 75 PG variant.

A final variant purification incorporated anion exchange to improve protein purity. This method was the same as the standard protocol except that IMAC elution was performed over a 50 CV elution gradient, digested protein was desalted a second time using a HiPrep 26/10 Desalting Column (GE Healthcare) into ANX Buffer A (20 mM Tris pH 8.0, 10% w/v glycerol) and then loaded onto a 5 mL HiTRAP ANX FF column (GE Healthcare), washed with 10 CV ANX BufferA and then protein was eluted over a 20 CV gradient with ANX Buffer B (20 mM Tris pH 8.0, 1 M NaCl, 10% w/v glycerol). Fractions containing PelA were then concentrated and passed through a 16/600 HiPrep Superdex 75PG column (GE Healthcare) equilibrated with SEC buffer (20 mM Tris pH 8.0, 150 mM NaCl).

2.2.10 Mass Spectrometry of PelA 46-948 contaminant

SDS PAGE gel slices were excised with a scalpel and destained with 30% (v/v) ethanol at 65°C, repeating with fresh 30% (v/v) ethanol until clear. Gel was then washed with triethylammonium bicarbonate buffer (TEAB) in 50% acetonitrile (ACN) and incubated with 10 mM dithiothreitol (DTT) for 30 min at 55 °C. The DTT solution was decanted away and 30 mM iodoacetamide (IAA) in 50 mM TEAB added and incubated for 30 minutes at room temperature, vortexed and kept dark. IAA was removed and gel was washed with TEAB/50%ACN then 50 mM TEAB. Buffer was removed then slices were placed individually into a sterile petri dish and cut into 1x1mm pieces. Strips were transferred into a Low Bind tube using the scalpel then washed with TEAB/50% ACN then 100% ACN. Samples were then transferred to John Innes Centre Proteomics Facility for analysis.

Gel slices were washed, treated with trypsin, and extracted according to standard procedures[298]. The peptide solution resulting from the digest was mixed 1:1 with α-cyano-4-hydroxycinnamic acid (Sigma) as matrix in 50% acetonitrile/0.1% TFA and spotted on a steel target plate. The samples were analysed on an AutoflexTM Speed MALDI-TOF/TOF mass spectrometer (Bruker DaltonicsTM GmbH, Coventry, UK). The instrument was controlled by a flexControlTM (version 3.4, Bruker) using a method optimised for peptide detection and calibrated using peptide standards (Bruker). Data were processed in FlexAnalysis (Bruker) and the peak lists were submitted for a database search using an in-house Mascot[299] Server (Matrixscience, London, UK). The search was performed against a

Pseudomonas aeruginosa protein database downloaded from Uniprot (20161212, 5563 sequences) using trypsin/P as enzyme with maximum 1 missed cleavage, 100 ppm mass tolerance, carbamidomethylation (C) as fixed and oxidation (M) and acetylation (protein N-terminus) as variable modifications.

The samples for intact mass analysis were analysed by LCMS on a Synapt G2-Si mass spectrometer coupled to an Acquity UPLC system (Waters, Manchester, UK). Aliquots of the samples were injected onto an Aeris WIDEPORE 3.6µ C4 column, 2.1 mm x 50 mm, (Phenomenex, Macclesfield, UK) and eluted with a linear gradient of acetonitrile in water/0.1% formic acid from 5% to 90% in 5 min with a flow rate of 0.4 mL min^{-1}. The mass spectrometer was controlled by the Masslynx 4.1 software (Waters) and operated in positive MS-TOF and resolution mode with a capillary voltage of 2.5 kV and a cone voltage of 40 V. Calibration was performed in the m/z range of 50-2000 using sodium formate according to the manufacturer's recommendations. Leu-enkephalin peptide (0.5 µM in 50% methanol/0.1% formic acid, Waters) was infused at 10 µl min^{-1} as a lock mass and measured and applied every 30 s. The data were processed in Masslynx 4.1 after combining relevant spectra using the background subtract and smooth options. The protein mass was determined by deconvolution using the MaxEnt 1 option.

2.2.11 Gateway cloning of PelA and PelE constructs

Further cloning used the Gateway® expression system for *E.coli* using a standard procedure[300]. The gentamycin resistant pDONR207 plasmid was used as a donor and ampicillin resistant pDEST17 as the expression plasmid, containing an N-terminal hexahistidine tag. PCR was performed using the manufacturer's instruction with Phusion® High Fidelity DNA Polymerase (NEB) in GC buffer with 5% v/v DMSO in two stages. Adapter 1 primers incorporate an N-terminal 3C protease cleavage sequence and a C-terminal partial attB2 site. Adapter 2 primers used a 1/50 dilution of the Adapter 1 product as a PCR template and contain a standardised attB1-3C protease cleavage forward primer and completed attB2 reverse primer to complete the Gateway® recombinant site. Thermal cycling programs are illustrated in Table A3 (see Appendix).

Gene of interest was transferred into pDONR207 using a 5 µl BP Clonase® (Thermo Fisher) reaction as per manufacturer's instructions, incubating overnight at 25 °C following a brief vortex to mix. The BP reaction integrates the PCR product into the plasmid through homologous recombination. Reactions were terminated with the addition of 0.5 µl Proteinase K (Thermo Fisher) for 10 minutes at 37 °C.

For transformation, 2 µl of the BP reaction was added to 20 µl DH5α Library-Efficiency Competent Cells (Invitrogen) on ice in a 1.5 mL microcentrifuge tube. The mix was incubated for 30 minutes and water bath to prepared at 42 °C. SOC media (Sigma) was defrosted in the water bath. Cells were heat shocked for 35 seconds then placed back on ice for 2

minutes. 180 µl of the warm SOC media was added to the microcentrifuge tube then incubated (shaken) at 37 °C for 1 hour. Heated LB-agar containing 30 µg/mL gentamycin was added to CytoOne® 6 Well Plates (StarLabs) and allowed to set. Wells were then inoculated with 30 µl of transformed culture. The remaining cultures were then centrifuged at 3000 xg for 5 minutes using a benchtop centrifuge. 140 µl of supernatant was removed and cells resuspended in the remaining 30 µl and used to inoculate a second well for each sample. Three 6 mm sterile glass beads were added to each well and shaken to mix. Beads were then removed and plates incubated, with lids, at 37 °C overnight.

Colonies were assessed for the presence of the gene insert using colony PCR (C-PCR). This technique uses primers from the original PCR amplification with DNA from a single colony as a template; a colony that contains the gene of interest will therefore produce a band at the expected MW for that gene while one that does not will only have non-specific amplification. 25 µl of sterile water was added to (preferably) 8 PCR tubes per transformed construct using a P100 Eppendorf™ Research Plus® micropipette. Volume was reduced to 15 µl and then individual colonies were picked using sterile pipette tips, inoculated to the water and mixed by aspiration. 15 µl of diluted colony was then transferred to a fresh PCR tube; the remainder in the first tube now reflects a sample for growth should a positive result be returned. The 15 µl samples were denatured at 98°C for 10 minutes using a thermal cycler and then pelleted at 4000 xg for 5 minutes. 1 µl of supernatant was then used as a template for PCR using GoTAQ® G2 Flexi Polymerase (Promega) in Green Buffer following the manufacturer's protocol with the universal Adapter 2 primers and an independent thermal cycler program (see Appendix Table A3). Agarose gel electrophoresis was carried out as previously except that 5 µl of sample was run with no loading dye added, as loading dye was included in the Green Buffer. Colonies that produced bands at the correct molecular weight were deemed successful and the corresponding sequestered diluted *E.coli* samples were used to inoculate 10 mL LB containing 30 µg/mL gentamycin, which was then grown overnight at 30 °C, 200 rpm. Plasmid extraction was performed using the GenElute® Plasmid Miniprep Kit (Merck) as per the manufacturer's instructions.

Extracted plasmids underwent LR Clonase® (Thermo Fisher) reactions to transfer the gene of interest from pDONR207 to pDEST17. Reactions were performed in 5 µl using 150 ng masses of donor and acceptor plasmids at a 1:1 ratio for 1 hour at 37 °C then terminated as per the BP reaction. LR reactions were transformed, grown and extracted using the same protocols as for the BP reactions except 100 µg/mL ampicillin was used for selection. Plasmids were then sent for Sanger sequencing (Source Bioscience™, Cambridge) using appropriate Gateway® compatible primers supplied by that service. Plasmids found to contain the gene of interest were transformed into BL21(DE3) Rosetta2 (Agilent) and SoluBL21 (AMSBIO) for expression and grown overnight using the same protocol, with 30 µg/mL chloramphenicol additionally used for selection in Rosetta2. Glycerol stocks

were made by diluting 500 µl of overnight culture with 500 µl 60% glycerol and freezing in dry ice or liquid nitrogen. Glycerol stocks were stored at -80 °C.

2.2.12 Expression Trials of Gateway constructs.

These trials were performed to assess whether any of the cloned PelA truncations, or PelE, successfully produced soluble protein. 1.5 mL of preculture was used to inoculate to 50 mL LB antibiotic and overexpression was induced with 0.1 mM IPTG and temperature was reduced to 20 °C. 1mL samples were taken from the 50 mL after 20 hours and lysed by being sonicated twice for 30 seconds each with a 3 mm probe at 15% amplitude. Probe was cleaned with 20% v/v ethanol and cooled with ice slurry for 1 minute between sonication blocks. Lysed samples were fractionated at 20,000 xg using a benchtop centrifuge for 30 minutes at 4 °C. Samples of supernatant were taken with the remaining soluble fraction being discarded. Insoluble fraction pellets were resuspended in 1 mL distilled H_2O. Samples were then assessed by SDS PAGE.

To test whether IPTG concentration affected the solubility of the expressed proteins PelA 523-920 and 262-507 this protocol was repeated at IPTG concentrations at ten-fold dilutions from 1-100 µM.

PelA and PelE Rosetta 2 constructs were inoculated from glycerol stocks to 10mL LB containing 100 µg/mL ampicillin and 35 µg/mL chloramphenicol and grown overnight at 30 °C, 200 rpm overnight. 0.75 mL of these overnight cultures was inoculated to 25 mL each of LB antibiotics and grown at 37 °C until OD_{600} reached 0.60. Overexpression was induced with 0.1 mM IPTG and temperature was reduced to 20 °C. Cells were clarified at 4000 xg for 15 minutes after 20 hours and stored at -20 °C. Cell pellets were resuspended in 7.5 mL Lysis Buffer (50 mM Tris pH 8.0, 0.5 M NaCl, 10m M Imidazole, 10% w/v glycerol, 1 cOmplete protease inhibitor tablet/300 mL, a pinch of DNase) for 1 hour and then lysed at 30 kPSI using a Cell Disruptor® (Constant Systems). 1.5 mL of each lysate was clarified at 20 000 xg for 30 minutes at 4 °C using a benchtop centrifuge. Soluble fractions were applied to pre-equilibrated 50 µl Ni^{2+}-NTA Superflow beads in a 96 well format using a vacuum manifold. Wells were washed three times with 0.5 mL Wash Buffer (20 mM Tris pH 8.0, 0.5 M NaCl, 30 mM Imidazole, 10% w/v glycerol) and then wells was incubated in 50 µl Elution Buffer (20 mM Tris pH 8.0, 0.5 M NaCl, 300 mM Imidazole, 10% w/v glycerol) for 30 minutes at 4 °C and centrifuged at 4000 xg for 5 minutes. Eluate was reapplied to the beads and underwent a further 15-minute incubation and centrifugation to produce a more concentrated elution. Insoluble fractions were resuspended in 1 mL Lysis Buffer. Sample preparation and SDS PAGE was performed as previously.

2.2.13 Purification of PelA 46-507 from inclusion bodies

The purpose of this work was to determine whether a method of solubilising protein from the insoluble fraction using lightly denaturing conditions would produce protein viable for crystallisation. PelA 46-507

(pDEST17, Rosetta2) was cultured as per the standard protocol using 1L day culture and post-induction incubation temperature of 16 °C mL. The cell pellet was resuspended and lysed as per the standard protocol. During the remaining protocol, multiple rounds of centrifugation occur, all at 10, 000 xg for 10 minutes, 4 °C, using a JA25.50 rotor.

The lysate was clarified, then the supernatant was removed, and the pellet resuspended in 40mL Buffer B (20 mM Tris pH 8.0, 0. 5M NaCl, 1.75 M GdHCl, 1% v/v Triton X-100, 1 mM β-mercaptoethanol) using a homogeniser. The solution was clarified and resuspended in this manner a twice more. This was then repeated, resuspending in distilled water twice and then in Buffer C (20 mM Tris pH 8.0, 0.5 M NaCl, 30 mM Imidazole, 0.2% w/v sarkosyl, 1 mM β-mercaptoethanol) three times.

Buffer C supernatant was applied to a HisTRAP FF column (GE Healthcare) using a peristaltic pump P1 pump (GE Healthcare). The HisTRAP column was connected to an AKTA Pure (GE Healthcare) at 4 °C and washed with Buffer C before being switched to Buffer D (20 mM Tris pH 8.0, 30 mM imidazole, 1 mM β-mercaptoethanol). Buffer D was placed in line A and Buffer E (20 mM Tris pH 8.0, 300 mM imidazole, 1 mM β-mercaptoethanol) in line B using a pump wash. A linear gradient elution was then applied to elute protein from the column. Eluted fractions were pooled and placed in a dialysis membrane with Buffer D containing 3C protease at a 1:100 mass ratio overnight at 4 °C. Protein was removed from the dialysis membrane and concentrated using an Ampicon® Centrifugal Concentrator (Millipore) with a molecular weight cut-off of 15 kDa. Concentrated sample was injected to a HiLoad 16/600 Superdex 75 PG column (Ge Healthcare).

2.2.14 Expression Trials of PelA46-262 and PelA 46-507

The purpose of these trials was to determine whether the solubility of PelA 46-262 and 46-507 (pDEST17, Rosetta2) were affected by typical culture conditions. Overnight cultures of PelA 46-262 and PelA 46-507 (pDEST17, Rosetta2) were prepared as previously described for expression trials of Gateway constructs. Each of six flasks of 100mL LB containing 100 µg/mL ampicillin and 35 µg/mL chloramphenicol were inoculated with 3 mL of PelA 46-262 or PelA 46-507 and incubated at 37 °C, 200 rpm until an OD_{600} between 0.6 and 1.0 was observed. Half of the cultures were then induced to overexpress protein with 0.1 mM IPTG and the other half with 0.01 mM IPTG. Paired cultures were incubated at 37 °C for 4 hours, 20 °C for 20 hours or 16 °C for 20 hours. 1 mL samples of cultures were taken, and the remainder was harvested at 4000 xg for 15 minutes at 4 °C. Larger samples were stored at -20 °C.

1 mL samples were lysed, clarified, resuspended and run on SDS PAGE as described previously for the expression trials of Gateway constructs. This was later repeated using a Cell Disruptor® (Constant Systems) at 30 kPSI as the lysis method.

2.2.15 Expression and Purification of PelA 46-507

The purpose of this purification is to obtain PelA 46-507 from pDEST17 Rosetta2 cells for crystallisation experiments. This follows the standard protocols for cell culture and protein purification, except cells were grown for 20 hours at 18 ^{o}C following induction of overexpression.

2.2.16 Crystallisation of PelA 46-507

Protein from the peak identified as corresponding to the protein of interest was concentrated to 10 mg/mL and screened against the standard array of crystallisation screens as per standard protocol.

2.2.17 Crystal Structure Determination of PelA 46-507

Crystals were transferred to a droplet of cryoprotectant solution (0.1M Tris pH 8.5, 0.2M Trimethylamine N-oxide, 20% w/v PEG 2000 MME, 20% v/v glycerol), soaked for 30 seconds, then flash frozen in liquid nitrogen.

Crystals were sent to Diamond Light Source (UK), beamLine station i04. 3600 frames each of 0.1° rotation and 0.05s exposure were collected using a wavelength of 0.9795 Å. Diffraction data was processed using xia2 Dials to a resolution limit of 2.61 Å.

Phases were estimated using molecular replacement as part of the Phenix[261] programme suite with the crystal structure of the putative glycosidase tm1410 from *Thermotoga maritima* (PDB: 2AAM) used as a best fit search model identified by PHYRE2[125]. Refinement was performed using iterations of refinement using Phenix[261], CCP4[249] and manual model building with Coot[260].

2.2.18 Sequence and ligation independent cloning (SLIC) of PelA and PelB

As an established method with a more rapid turnover time, the SLIC technique[301] was selected for the final cloning of PelA and for constructs of PelB. To remove the thermolysin cleavage site at which PelA 46-948 is degraded to PelA 46-523, primers were designed to introduce a double glycine mutation at residues 523-524. A C-terminal hexahistidine tag was also to be added to the end of the protein in place of an N-terminal tag as stable PelA 46-948 has been reported to be expressed using such a construct[150]. The same paper identified two constructs of PelB that were capable of interacting with PelA; PelB 351-588 and PelB 47-880. The PelB constructs were cloned into *pEHisTEV* such that the N-terminal tag would be linked to the gene while PelA was cloned N-terminal to this tag with a stop codon at the end of the gene to prevent the tag's expression.

PCR was performed using the Q5 Hot-Start DNA Polymerase (NEB) reaction under the manufacturer's instructions. PCR products were digested with 0.5 units of T4 DNA Polymerase (Thermo Fisher) using the SLIC1 program (see Appendix Table A1) to introduce 5' overhangs. Inserts were annealed to *pEHisTEV* plasmids using the SLIC2 program (see Appendix: Table A3) at a

3:1 molar ratio in Annealing Buffer (10mM Tris pH 8.0, 100mM NaCl, 1mM EDTA). 10µl of the 20µl reaction was transformed into 200µl TOP10 *E.coli* competent cells using a 90 second heat shock transformation. Cells were incubated with 800 µl SOC media for 1 hour at 37°C then centrifuged for 4 minutes at 3000 xg. Supernatant was discarded and pellets resuspended and plated to LB agar plates containing 35µg/mL kanamycin which were incubated overnight at 37°C.

2.2.19 Semi-functional assignment of PelA 303-507

To assess the oligomerisation state of PelA 303-507, PelA 46-507 and PelA 46-948 *in vitro*, the proteins were expressed and purified as previously described. Additionally, PelA 46-303 was expressed and purified using the same method as for PelA 46-507. A Superdex 200 Increase 10/300 GL Column (GE Healthcare) was calibrated using a Gel Filtration Markers Kit for Protein Molecular Weights 12,000-200,000 Da (Sigma) using an AKTA Pure (GE Healthcare) at 4°C. Proteins were eluted using using SEC buffer (20mM Tris pH 8.0, 150mM NaCl, 2mM DTT). 500µl of PelA samples containing 1 mg of protein were passed through the column individually and in pairwise combination with other samples. Molecular weight of the peaks were estimated using the calibration of Kav vs. LogMW. Kav was calculated using the below formula:

$$Kav = \frac{Ve - Vo}{Vc - Vo}$$

Where Ve is the elution volume of the peak, Vo is the void volume and Vc is the volume of the column.

2.3 WaaB Methods (Chapter 4)

The methods described in this section are based upon on the basic protocols described in 2.1 with specific details for experiments that took place in place in Chapter 4.

2.3.1 Identification of Key Mutants of WaaB

Residues of interest were identified using the X-ray structure of WaaB co-crystallised with UDP (PDB: 5N80) superimposed with the structure of WaaG[302] (PDB: 2IV7). Residues V186, Q194, K195, I216, W243, E268, T273 and E276 were identified as having appropriate orientation and distance to the co-bound UDP molecule of the wild type WaaB structure to be involved in binding. Additionally, residues K195 and E268 occupy equivalent positions of the proposed catalytic residues of WaaG[302].

2.3.2 Cloning of the WaaB mutants

Site directed mutagenesis was performed using the Liu and Naismith method[303] (see Appendix: Table A1 for primer sequences) to introduce eight separate mutations (V186A, Q194A, K195A, I216A, W243A, E268A, T273A and E276A). PCR was performed using previously cloned *S. enterica (typhimurium)* WaaB in the pLou3 plasmid[304] and the Q5 HotStart polymerase reaction at annealing temperatures of 62°C and 53°C in Ram 1

and 2 respectively. Dpn1 digestion was completed using 20U enzyme/50µl reaction for 2hrs at 37°C. Mutated plasmids were then transformed into XL-10 Gold Ultracompetent cells at a 2:15 ratio using the standard heat shock protocol.

2.3.3 Expression and purification of the WaaB WT and mutants

The wild type and mutant plasmids were transformed into BL21 (DE3) cells and cultured using the standard protocol, except only 2L of day culture was grown for each variant and cells were grown for 4 hours at 37 °C, following induction, before harvesting.

Cells were thawed and resuspended in lysis buffer (20mM Tris pH 7.0, 150mM NaCl, 30mM imidazole) with 2µg DNase and lysed using a Constant Systems® Cell Disruptor at 30 kPSI. Soluble fractions were applied to 5mL HisTRAP FF column (GE Healthcare), washed with 10 CV lysis buffer then eluted with 2 CV elution buffer (20mM Tris pH 7.0, 150mM NaCl, 500mM imidazole). Fractions containing WaaB (as assessed by SDS PAGE) were pooled and concentrated to 0.5mL then loaded to a 16/600 HiPrep Superdex 200PG (GE Healthcare) column for gel filtration using SEC buffer (20mM Tris pH 7.0, 150mM NaCl). Fractions containing WaaB were identified by absorbance at 280nm and SDS PAGE then pooled. Subsequently, half of the pooled sample was concentrated and flash frozen in liquid nitrogen either with or without 33% v/v glycerol while the remainder was incubated overnight at 4°C with TEV protease at a 1:100 mass ratio for tag removal. Detagged protein was passed through 3mL Ni^{2+}-NTA Superflow gravity column, concentrated and then loaded to a 16/600 HiPrep 200PG column for gel filtration using SEC buffer. WaaB-containing fractions were pooled, concentrated and flash frozen with liquid nitrogen, either with or without 33% v/v glycerol.

2.3.4 Measuring glycosyltransferase activity of WaaB mutants

The hydrolytic activity of WaaB was measured using the UDP-Glo® Glycosyltransferase Assay Kit (Promega) as per the manufacturer's instructions. A UDP-galactose concentration of 100µM was found to provide the clearest signal. Wild type and mutant proteins were assayed in a 1 in 2 serial dilution from 200 to 0.775ng WaaB per well (n=4) alongside negative controls (n=3) for UDP-galactose, UDP-Glo® (luciferase), WaaB and empty wells. Measurements were performed in white polystyrene 384 well plates using a HIDEX Sense instrument in top read mode with a 1s count time at room temperature following a 60 minutes reaction and 60 minutes post-reaction incubation. Percentage activity of mutants was determined by dividing mean luminescence of the mutant by mean luminescence of the wild type and multiplying by 100 for each concentration of WaaB. Student's t-test of raw data found that WaaB concentrations greater than 6.25 ng/well were significantly different ($p<0.05$) from WaaB negative controls and so results at lower concentrations were excluded for the purposes of comparing activity. Means were taken of percentage activity calculated for

6.25-200 ng/well concentrations to determine the overall percentage activity compared to the wild type for each mutant.

2.3.5 Glycosyltransferase assay using different UDP-sugar donors

Assays testing the ability of both WT and T273A WaaB to hydrolyze varied UDP-sugar substrates used the same basic parameters as the assays with UDP-galactose except that only WaaB concentrations of 100ng/reaction and 200ng/reaction were explored. The sugars selected for investigation were UDP-galactose (to serve as a control), UDP-galactosamine, UDP-glucose, UDP-glucosamine and UDP-glucuronic acid as these are common sugars substrates for glycosyltransferases of the GT4 family and are readily available at the purity required for the assays.

Results are representative of four assays (n=15, 100ng WaaB/reaction). Baseline luminescence was calculated using ΔWaaB conditions containing the other assay components. Mean baseline luminescence for each UDP-sugar was subtracted from the experimental conditions with the corresponding UDP-sugars to produce true WaaB generated luminesce values.

2.3.6 Purification of WaaB mutants for crystallisation

Mutant proteins WaaB K195A, E268A and T273A were expressed and purified as previously described for purification with tag removal. Protocols were subsequently and sequentially adapted to improve outcomes; buffer pHs were changed from 7.0 to 8.0 and then 7.8, the 16/600 HiPrep Superdex 200 PG (GE Healthcare) was replaced with a 16/600 HiLoad Superdex 75 PG (GE Healthcare) to improve resolution between tag and WaaB, an MBPTRAP column (GE Healthcare) was also added to capture additional tag before gel filtration and finally, purifications were performed at room temperature rather than at 4^{o}C.

The final purification of WaaB T273A used the standard protocol with the following exceptions; Lysis Buffer was 20mM Tris pH 7.8, 300mM NaCl, 30mM imidazole with 1µg/L DNase, Elution Buffer was 20mM Tris pH 7.8, 300mM NaCl, 300mM imidazole, the TEV protease digest was incubated overnight at room temperature with gentle rocking, and a 5mL MBPTRAP column (GE Healthcare) was connected to the bottom of the 5ml HisTRAP FF column (GE Healthcare) to help capture the released tag and contaminant MBP. Fractions were concentrated to 10 mg/mL using an Ampicon® Centrifugal Concentrator (Millipore) with a molecular weight cut-off of 15 kDa. All procedures were performed at room temperature to avoid precipitation of WaaB.

2.3.7 Crystallisation of WaaB T273A

WaaB was previously crystallised at 10mg/mL in 50mM HEPES pH6.8, 16% W/V PEG 3350, 1mM sodium azide, 1% tryptone. Purified WaaB K195A, E268A and T273A mutants were screened for crystallization against this condition and the standard optimisation screens, as per the standard protocol. WaaB T273A was additionally screened against these conditions

with the addition of 1mM UDP to the protein droplet. Crystallisation conditions were optimised as per standard protocol.

2.3.8 Solving the structure of WaaB T273A

Crystals were harvested using mounted LithoLoops (Molecular Dimensions) and transferred to droplets of fresh well solution containing a cryoprotectant selected by comparison with cryoprotectants used for existing crystals[243]. Crystals were incubated in the droplet at 16°C for up to 1 minute and then flash frozen in liquid nitrogen. Crystals were sent to Diamond Light Source (UK), beamLine station i03 or i04. X-ray diffraction data was collected by the rotation method at the Diamond Light Source (UK) on beamLines i03 and i04. 3600 x 0.1° images were collected using a wavelength of 0.9795Å and exposure time of 0.05 seconds. One crystal of WaaB T273A in complex with UDP diffracted to 2.6 Å.

The crystal structure of WaaB T273A was solved using molecular replacement in the Phenix software suite[261] using the structure of wild type WaaB as a model. The structure was manually rebuilt in Coot[260] and refined using Phenix and REFMAC as part of the CCP4[249] programme suite.

2.4 GtfC Methods (Chapter 5)

The methods described in this section are based upon on the basic protocols described in 2.1 with specific details for experiments that took place in place in Chapter 3.

2.4.1 Crystallisation of $GtfC_{100\text{-}23}$ and $GtfC_{ATCC\ 53608}$

Protein was supplied by our collaborators from Dr Nathalie Juge's group at the Quadram Institute at 10mg/mL in 20mM Tris pH 8.0, 150mM NaCl. $GtfC_{100\text{-}23}$ and $GtfC_{53608}$ were screened against using the standard crystallisation screen and optimisation protocols.

Crystals were harvested using mounted LithoLoops (Molecular Dimensions) and transferred to droplets of fresh well solution containing a cryoprotectant selected by comparison with cryoprotectants used for existing crystals[243]. Crystals were incubated in the droplet at 16°C for up to 1 minute and then flash frozen in liquid nitrogen. Crystals were sent to Diamond Light Source (UK), beamLine station i03 or i04. Images were collected using the rotation method with 3600 frames collected of 0.1° rotation with an exposure time of 0.05 s at an X-ray wavelength of 0.9795Å.

2.4.2 Solving the structure of $GtfC_{100\text{-}23}$

The initial crystal structure of apo-$GtfC_{100\text{-}23}$ was solved at 2.8 Å resolution from plates grown in 0.1M Bis-Tris pH 7.5, 0.2M potassium thiocyanate, 20% w/v PEG 3350 at 16°C after 14 days using 30% v/v glycerol as a cryoprotectant. This structure was solved using molecular replacement in the Phenix software suite[261] using the structure of the homologous glycosyltransferase 3 from *Streptococcus parasanguinis* (PDB: 3QKW)[274] as a search model, as identified using PHYRE2[125]. The structure was further

refined using Phenix and Buccaneer from the CCP4 suite[249] and with manual rebuilding in Coot[260]. This structural model contains four copies of the enzyme in the asymmetric unit in space group P 2_1 2 $_12_1$.

Later, a second rod-like crystal appeared after seven days in a condition containing 0.02 M calcium chloride dihydrate, 0.1 M sodium acetate pH 4.6, 30% v/v hexylene glycol, 1 mM UDP at 16 °C. Crystals were harvested using well solution containing no UDP and no additional cryoprotectant and diffracted to 2.4 Å resolution. These crystals were found to be perfectly twinned and therefore were not used further.

A third crystal, grown in the same condition as the first but with 1mM UDP added, was found to diffract to 2.5 Å in space group P 2_1 1. This structure was solved using molecular replacement in the Phenix software suite[261] using a monomer of the previous solved structure of apo-GtfC as a search model. The structure was further refined using Phenix, CCP4[249] and manually built in Coot[260]. The structural model contains four copies of the enzyme in the asymmetric unit, three of which bind UDP.

2.4.3 Determining the biological unit of $GtfC_{100\text{-}23}$

A Superdex 200 Increase 10/30 GL Column (GE Healthcare) was calibrated using a Gel Filtration Markers Kit for Protein Molecular Weights 12,000-200,000 Da (Sigma) using an AKTA Pure (GE Healthcare) at 4 °C. A 500 µl sample of $GtfC_{100\text{-}23}$ at 1 mg/mL was passed through the column. Molecular weight of the peaks were estimated using the calibration of Kav vs. LogMW as previously described (section 2.2.19).

Chapter 3 - Proteins of the *pel* operon

3.1 Cloning of the *pel* operon

Initial analysis of the proteins from the Pel operon provided a map of predicted domain boundaries (see Figure 24) that were used to generate constructs for protein expression. A broad, rather than targeted, approach to cloning was chosen due to the high likelihood of components of the operon being non-viable at the expression, purification or crystallisation stages of the pipeline, therefore a diverse focus is more likely to return a positive result than a singular focus upon a single gene. It was decided that the full-length genes would be suitable for expression as the structure of a full length protein is more valuable than a truncated one. There are two exceptions where the organisation of the proteins means that it would be unlikely to express well in *E.coli* as a full length protein. PelC is predicted to be a soluble periplasmic protein that would be attached to the OM through lipation; as soluble proteins are typically easier to deal with than membrane proteins it was decided that removing the N-terminal signal peptide and lipation site would likely improve the expression of the protein. Additionally, the N-terminal glycoside hydrolase domain of PelA was flagged as a potential transmembrane region and so an additional construct C-terminal to this domain was selected for cloning and expression.

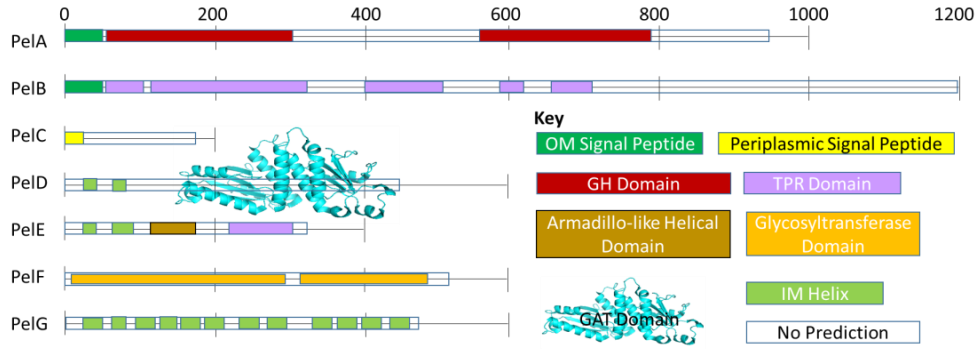

Figure 24: Bioinformatic domain boundary predictions in PelA-G used to determine constructs. Scale shows length in amino acids with a major unit of 200. Coloured bars show predicted domain boundaries within the proteins, with the pre-determined cytoplasmic domain of PelD represented as a cyan ribbon diagram. White bars show regions of the protein for which the amino acid sequence lack sufficient similarity to predictively assign a domain. Full length constructs were selected to be generated for all seven proteins, other than PelD, in addition to constructs of PelA with signal peptide and N-terminal glycoside hydrolase domain removed and PelC with signal peptide removed.

Expression plasmids to house these constructs were chosen based upon the predicted solubility of the protein. The pEHisTEV and pLou3 vectors were chosen for the expression of the soluble proteins PelA (1-948, 288-948), PelC (20-172) and PelF (1-506). Both of these vectors contain an N-terminal hexahistidine tag with a Tobacco Etch Virus protease cleavage sequence for removal of the tag under an isopropyl β- d-1-

thiogalactopyranoside (IPTG) inducible promoter. The pLou3 vector also contains MalE[287] between the histidine tag and TEV site which has been reported to improve protein solubility as well as allowing purification using a maltose binding column should purification using immobilised metal affinity chromatography (IMAC) prove problematic. The pBAD24 vector was chosen for expression of predicted membrane proteins PelA (1-948), PelB (1-1192), PelD (1-455), PelE (1-329) and PelG (1-456) under an L-arabinose inducible promoter that helps to reduce toxic leaky expression that might be damaging to cells and allow rapid expression in a late log phase[305]. During PCR, an octahistidine tag was added to the C-terminus of these membrane proteins to allow for purification using IMAC.

PCR reactions incorporating vector-compatible restriction sites at the gene ends were used as the basis for ligation-based cloning and successful amplification was achieved for all targets other than PelB, which consistently encountered issues with non-specific amplification. This is not entirely surprising given the length of the PelB gene and the GC-rich codon bias of *P. aeruginosa*. PCR restriction sites and vectors were digested with appropriate restriction enzymes then ligated following denaturation of the enzymes and purification and extraction from agarose gels. Ligated plasmids containing the genes for the proteins of interest were transformed into *E.coli* TAM1 or Top10 for antibiotic selection and plasmid amplification then extracted. Extracted plasmids underwent double-digests to determine the success of the ligation reaction. A double digestion was considered successful if the plasmid separated into two bands, one of which was at the correct molecular weight for the gene of interest. Sequencing was used to confirm that the plasmid insert was correct, using BLAST, and that the protein of interest was in frame. PelA 1-948, 288-948, PelC 20-172, PelD 1-455, PelE 1-329 and PelG 1-456 successfully passed double digest, with PelD 1-329 failing at the sequencing stage. Double digestion and sequencing of PelF 1-506 identified an error in primer design, namely that the forward primer's restriction site was present within the gene; the resulting cloning produced a construct containing the C-terminal fragment 761-1524 of the gene in frame 3. The resulting sequencing introduced a STOP codon after 123 bases making it of little use.

3.2 Initial constructs do not express in *E.coli*

Plasmids were cloned into *E.coli* C43 cells for protein expression. C43 cells are optimised for the expression of potentially toxic proteins, such as membrane proteins or proteins of unknown function, as is the case here. Initial expression trials were performed using standardised conditions (0.02% L-arabinose/0.1 mM IPTG at OD 0.980 for 4 hour incubation at 37 ^{o}C) then purified in one-step Ni^{2+}-NTA columns for analysis (see Figure 2.3A). Potential bands were observed for PelA 288-948 (pBAD24) near 73 kDa and for PelC 20-172 around 64 kDa (although a similar band is seen in other lanes). Significant bands were not observed at the expected molecular weights for PelA 288-948 in pLou3 (120 kDa), PelA 1-948 in pBAD24 (105 kDa), PelE 1-329 (37 kDa) or PelG 1-456 (52 kDa). A large

overexpression band was observed around 50 kDa in the pLou3 control condition that corresponds to the N-terminal tag from the pLou3 plasmid.

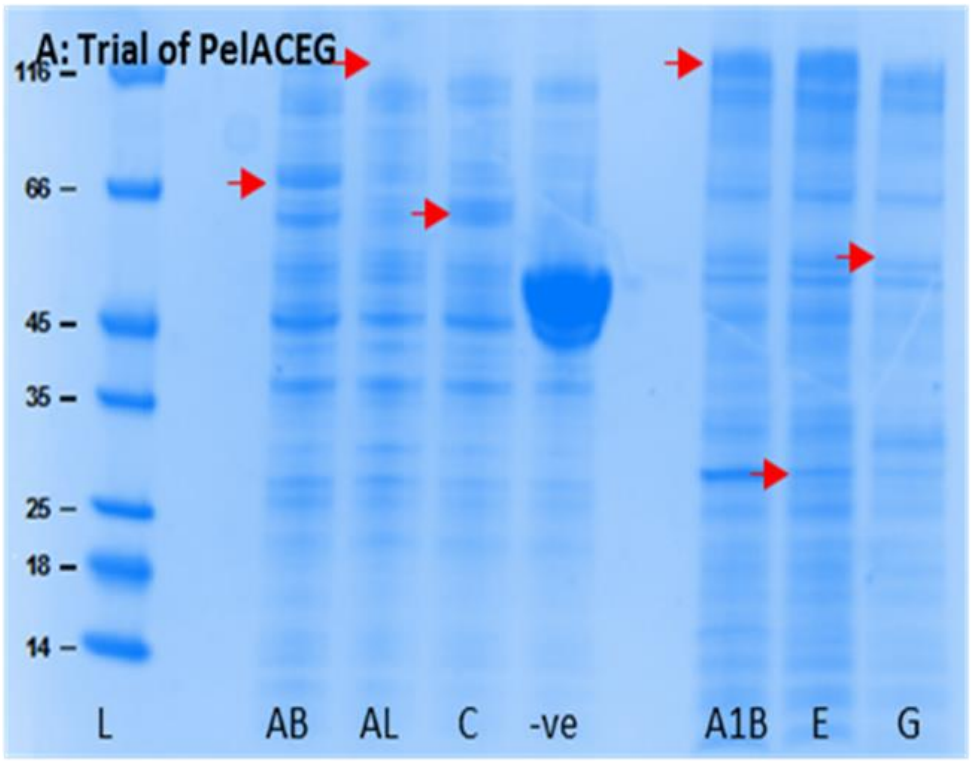

Figure 25: SDS PAGE gel of expression trials of Pel A, C, E and G. Red arrows mark expected molecular weights of the proteins of interest. **Single step his tag purified Pel expressions** after 4 hours at 37 °C. The negative control illustrates the expected proportion of an overexpressed protein. Protein at the expected molecular weights shown do not correspond to the cloned genes. **Key:** Ladder (L), PelA288-946 in pBAD24 (AB), PelA288-946 in pLou3 (AL), PelC20-172 in pLou3 (C), pLou3 control (-ve), PelA1-946 in pBAD24 (A1B), PelE1-329 in pBAD24 (E), PelG1-456 in pBAD24 (G).Samples in lanes 3-6 were purified as soluble proteins and 8-10 as membrane proteins.

To test whether expression of the potential hits could be improved, pLou3 PelC 20-172 underwent further trials using varying IPTG (0.1, 0.5 and 1.0 mM), and temperature (20, 30 and 37 °C). This trial produced the previously observed bands at 64 kDa, however, the intensity of these bands was not relative to IPTG concentration and showed no specific variance with temperature, supporting an interpretation that this protein is not an expression protein but one native to *E.coli*. This was then confirmed for both pBAD24 PelA288+ and pLou3 PelC 20-172 during a second trial which made use of the His-tag binding InVision® stain as well as monitoring expression over 16 hours at 20 and 37 °C. It was therefore concluded that Pel A (1-948, 288-948), PelC (20-172), PelE (1-329) and PelG (1-456) do not express stably in *E.coli* under standard conditions.

3.3 Refining Cloning and Expression

Given the lack of expression from the previous cloning attempt it was obvious that a new strategy needed to be employed. The Pel proteins were submitted to the PHYRE2[125] homology modelling server to reassess appropriate domain boundaries. Rather than trying to work on all proteins simultaneously, it was also determined that a more focused approach may be required to successfully express these proteins therefore the soluble proteins PelA and PelF were selected as the most suitable targets given the relative difficulty of working with membrane proteins. PelC was eliminated as a target due to the publication of a structure by another group[152].

In these cases, PHYRE2[125] analysis identified a larger region corresponding to the N-terminal glycoside hydrolase domain, 46-303 rather than 52-288, and comparison with homologous structures revealed that the domain was unlikely to be integral to a membrane. Using this information, it was decided that removing the signal peptide sequence should allow for best expression of the protein as the larger cytosolic compartment is more ideal for protein expression than the periplasmic compartment it was being directed to. Primers for PelF were redesigned using a different restriction site.

New constructs were successfully cloned and confirmed using double digest and sequencing. They were cloned into three strains of *E.coli* BL21 (DE3): pLysS, Rosetta 2 and Rosetta-2-gami. The pLysS cells show improved cell lysis, Rosetta 2 cells contain two pRARE plasmids that encode additional tRNAs that are not commonly produced by *E.coli* while Rosetta-2-gami cells have the two pRARE plasmids as well as producing a more oxidising cytosolic environment to replicate the periplasmic environment for disulphide bond formation. The pRARE plasmids would account for the codon bias issue or producing a *P. aeruginosa* protein, which has a strong GC bias, in *E.coli*, which does not. The Rosetta-2-gami cells may be important for expression of PelA as the protein is nominally a periplasmic protein and the C-terminal region contains two cysteines which may form a structurally important disulphide bond. The pLysS strain is unlikely to have a large impact upon expression but is expected to provide a higher yield than regular BL21 (DE3) and illustrates the regular level of expression that would be produced by a non-specialist cell line.

PelA and PelF pLou3 constructs were trialled for expression in the three different cell lines in the soluble fraction, insoluble fraction and elution from an IMAC column (see Figure 26). Soluble expression and IMAC fractions of PelA 46-948 was observed in both Rosetta 2 strains. Insoluble expression of both proteins was also observed in both Rosetta 2 strains and in the pLysS strain of PelF, suggesting that 1mM IPTG might be inducing the formation of inclusion bodies. Focusing upon PelA (46-948) in pLou3, reduced IPTG concentrations were trialled in the range of 1 µM - 1 mM to determine whether a higher proportion of soluble expression could be achieved. This found that PelA was associated with the insoluble fraction at concentrations 0.1-1.0 mM IPTG but that the PelA found in the soluble

fraction was proportional to this. The highest yields of PelA would therefore be found at the highest IPTG concentrations of 0.5-1 mM.

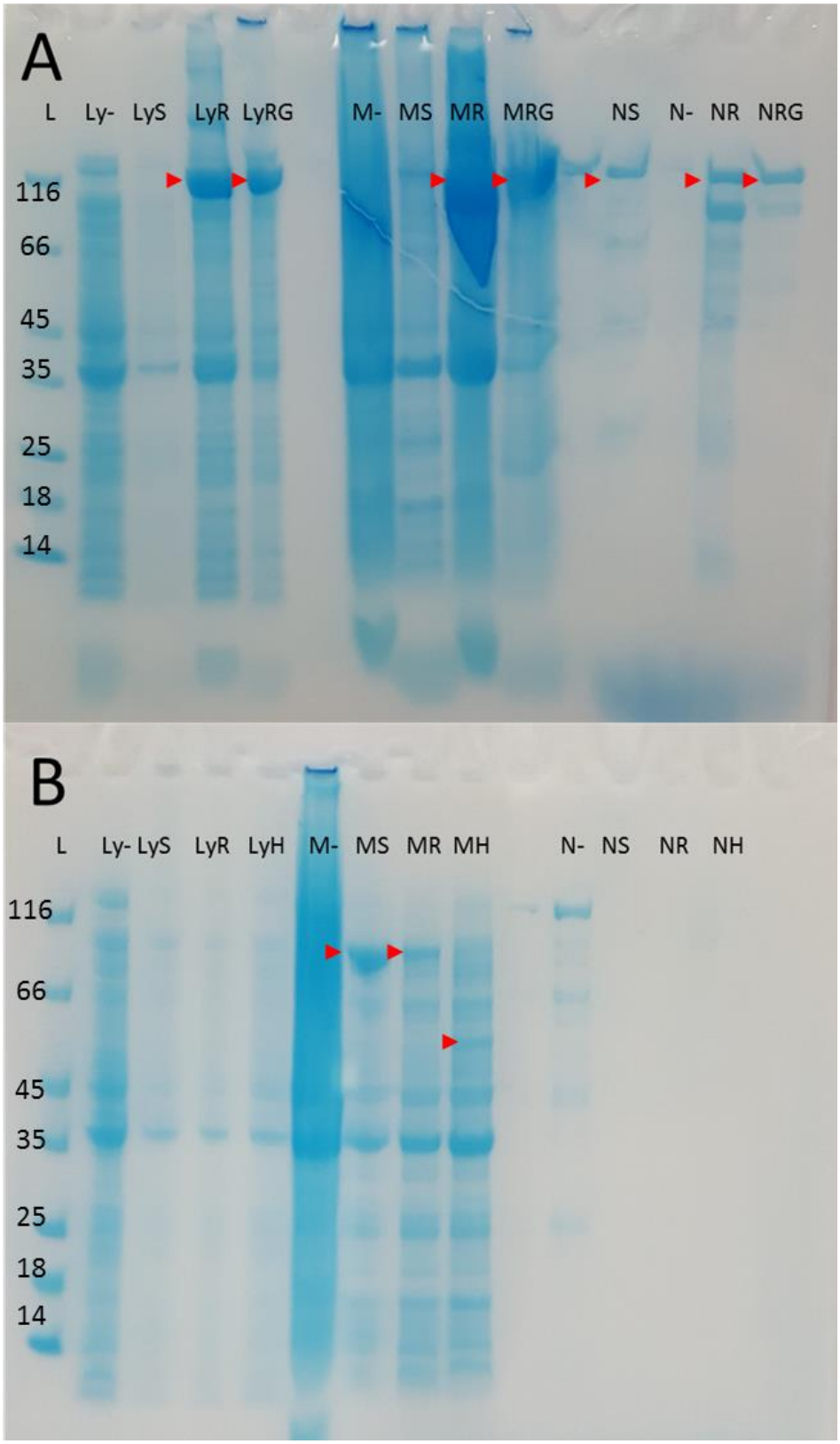

Figure 26: SDS PAGE gels of expression trials of PelA 46-948 pLou3 (A) and PelF 1-506 pLou3 and pEHisTEV (B). Trials were carried out at 1mM IPTG with 20-hour incubation at 20^0C. Red arrowheads show recombinant protein

expression. PelA expresses as a soluble protein while PelF is trapped in the membrane fraction. **Key:** Ladder (L, kDa), cell lysate (Ly), membrane fraction (M), elution from IMAC (N) pLou3 (-), pLou3 pLysS (S), pLou3 Rosetta2 (R), pLou3 Rosetta-2-gami (RG), pEHisTEV Rosetta 2 (H). Expected molecular weight of proteins: PelA 46-948 pLou3 (140kDa), PelF pLou3 (103kDa), PelF pEHisTEV (57kDa).

The possibility that PelF might express better at lower concentrations of IPTG or under different incubation parameters was also considered. It was found that IPTG concentrations of 1-100 µM did not improve the solubility of PelF, nor did using a 4 hour incubation at 37°C (data not shown). The additive trehalose has been reported to improve the solubility of some proteins[306] when present during lysis but addition of 25% trehalose did not visibly improve the solubility of PelF (data not shown). Some specialised strains of *E.coli* are also known to improve protein solubility when overexpressed. Attempts were made expressing PelF pLou3 and pEHisTEV in ArcticExpress (DE3) RT and SoluBL21™ cells with similar results (data not shown). It was therefore concluded that both PelF constructs produce nonviable protein.

3.4 Cloning and Expression of PelBC

As PelB was previously not successfully cloned, it was decided that this project might best be advanced through gene synthesis. As reports suggest that full length PelB protein is difficult to express[150] and that PelC interacts to form a complex[152], it was decided that co-expression of the two might produce a viable complex. The two genes were commercially synthesised and codon optimised for expression in *E.coli* by GenScript on the pET22b(+) plasmid. The pET22B(+) plasmid contains an N-terminal signal peptide for pectate lyase 2 (PelB) from *Pectobacterium carotovorum*, a strong OM signal peptide replacing the PelB signal peptide, and a C-terminal hexahistidine tag was incorporated onto PelC.

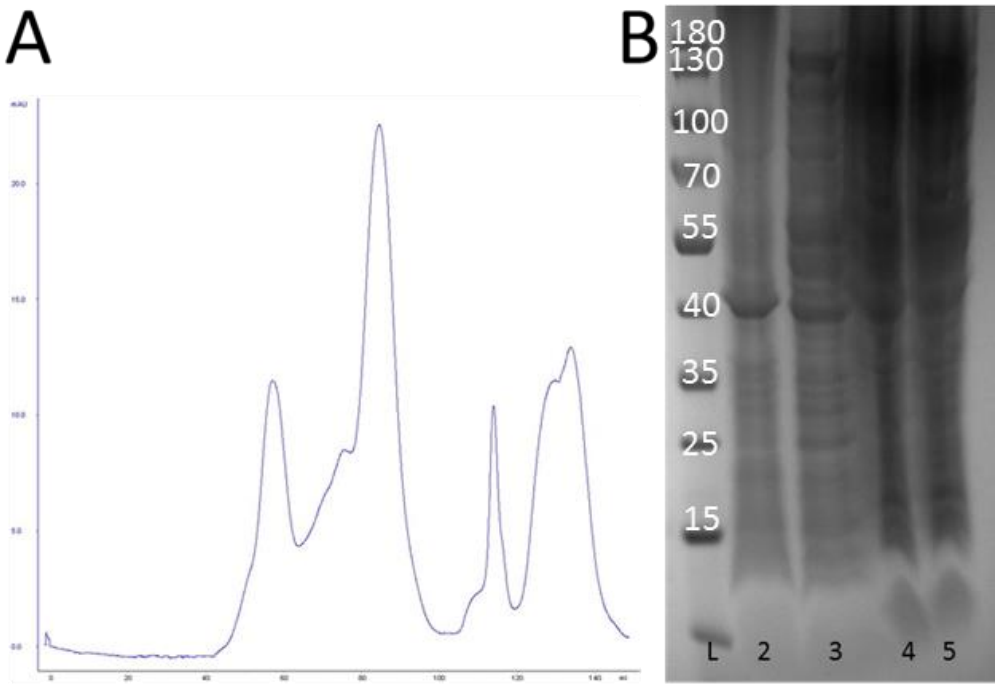

Figure 27: A: Size Exclusion Chromatograph of PelBC. ***20 milliabsorbance units is insufficient for accurate detection using SDS PAGE and insufficient for crystallography. B: SDS PAGE Expression Trial of PelBC.*** *PageRuler™ 10-180kDa Ladder (L), uninduced control (2), 37°C for 4 hours (3), 30°C for 20 hours (4), 16°C for 20 hours (5). Expected MW of PelB is 135 kDa and PelC is 19kDa. No visible overexpression of PelB or C.*

As the OM space is significantly smaller than the cytoplasmic space the overexpression of membrane proteins often produces a smaller yield of protein. It was therefore decided to attempt expression with a large scale (12 L) culture of the plasmid and determine degree of expression via purification. This method found little protein was bound to the IMAC column used to bind the tail of PelC and hopefully co-elute PelB (data not shown). The following gel filtration produced insignificant amounts of protein (see Figure 27A). A second attempt was therefore taken to assess ideal conditions for the expression of PelBC at temperatures from 16-37°C through SDS PAGE of lysed cells. Similarly, this showed no evidence of the overexpression of either PelB or PelC (see Figure 27 B). Further work by others within the lab group has similarly not found no expression of PelBC (results not shown).

3.5 PelA 46-948 co-purifies with a contaminant

The expression trials of PelA 46-948 pLou3 showed that the highest level of expression was with 0.1-1 mM IPTG and in Rosetta2 cells (see Chapter 2). It was also observed that a large portion of the expressed protein was found in the insoluble fraction. The additive trehalose has previously been identified as increasing the solubility of some proteins[306]; it was therefore determined that as part of the initial purification of PelA 46-948 that it would be divided into equal parts resuspended in a standard lysis solution and in the same solution containing 0.75 M trehalose. Following subcellular fractionation, the supernatants of the two samples could be combined to

continue the purification while the relative levels of PelA 46-948 found in the insoluble fractions could be compared to determine whether this additive was able to unlock the insoluble store of PelA. Contrary to expectation, it was found that a larger portion of PelA was in the insoluble fraction with this additive (see Figure 28), suggesting that a proportionately lower amount of PelA was soluble and therefore that trehalose should not be used in further purifications.

While the purification did successfully isolate PelA, there was a notable contaminant at the tail end of the peak that seemed to appear after digestion with TEV protease (see Figure 28). This timing suggested that the contaminant might be excess His_6-MBP tag that overloaded the HisTRAP column, an interpretation strengthened by an elution volume consistent with that expected of MBP. While this issue was observed, it was determined that the leading edge of the PelA peak should be sufficiently pure for crystallisation studies. While some small spheres and needles were produced, optimisation of these conditions did not produce diffracting crystals.

3.6 PelA 46-948 contaminant consistently co-purifies with PelA

A possible explanation for the lack of viable crystals was the presence of the contaminant. It was therefore decided to attempt expression and purification of PelA 46-948 pEHisTEV Rosetta2 as this plasmid is functionally identical but removes the MalE protein that produces the majority of the tag's mass; it should be comparatively simple to separate a 101 kDa protein from a 3 kDa tag using size exclusion chromatography. Using the sample purification protocol as previously, minus trehalose, it was found that a contaminant was present which eluted at the same volume as previously (see Figure 29). This suggests that the contaminant may be a stable degradation product of the full-length protein, so attempts were made to crystallise both PelA 46-948 and this 55 kDa fragment (referred to subsequently as PelA55). While no significant diffraction was observed from crystals of the cloned protein, modest diffraction was achieved from several PelA55 crystals. This limited diffraction may be due to contamination of the fragment by PelA 46-948.

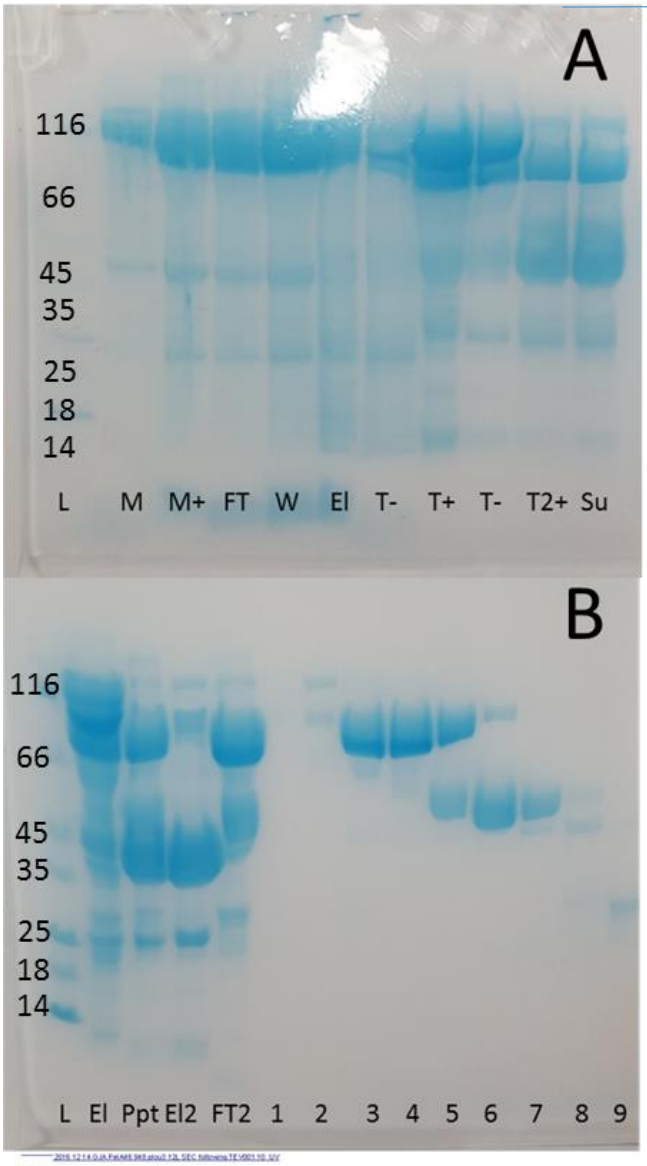

Figure 28 Purification trial of PelA 46-948 pLou3 Rosetta2**. A, SDS PAGE of Day 1 Purification:** Ladder (L), Membrane fraction (M), Membrane fraction with 25% w/v Trehalose (M+), HisTRAP flowthrough (FT), Column Wash (W), Column Elution (El), PreTEV digest (T-), TEV digest after 2 hours (T+), TEV digest overnight (T2+), Supernatant from TEV digest (Su).The recombinant protein is expressed as the high MW band around 116 kDa. The lower band between 116 and 66kDa after TEV protease digestion corresponds to PelA with its N-terminal tag removed. The fresh band in the T2+ and supernatant lanes around 45 kDa is consistent with the hexahistidine-MalE affinity tag. **B, SDS PAGE of Day 2 Purification:** (L), HisTRAP column elution from previous gel (El), TEV digest precipitate (Ppt), HisTRAP elution (El2, tag removal), HisTRAP Flowthrough (FT2, untagged PelA), elution from gel filtration at 48ml (1), 69ml (2), 72ml (3), 75ml (4), 78ml (5), 81ml (6), 84ml (7), 87ml (8) and 90ml (9). Both supernatant (see A) and precipitate fractions show similar contents. El2 removes his-tagged species via affinity chromatography

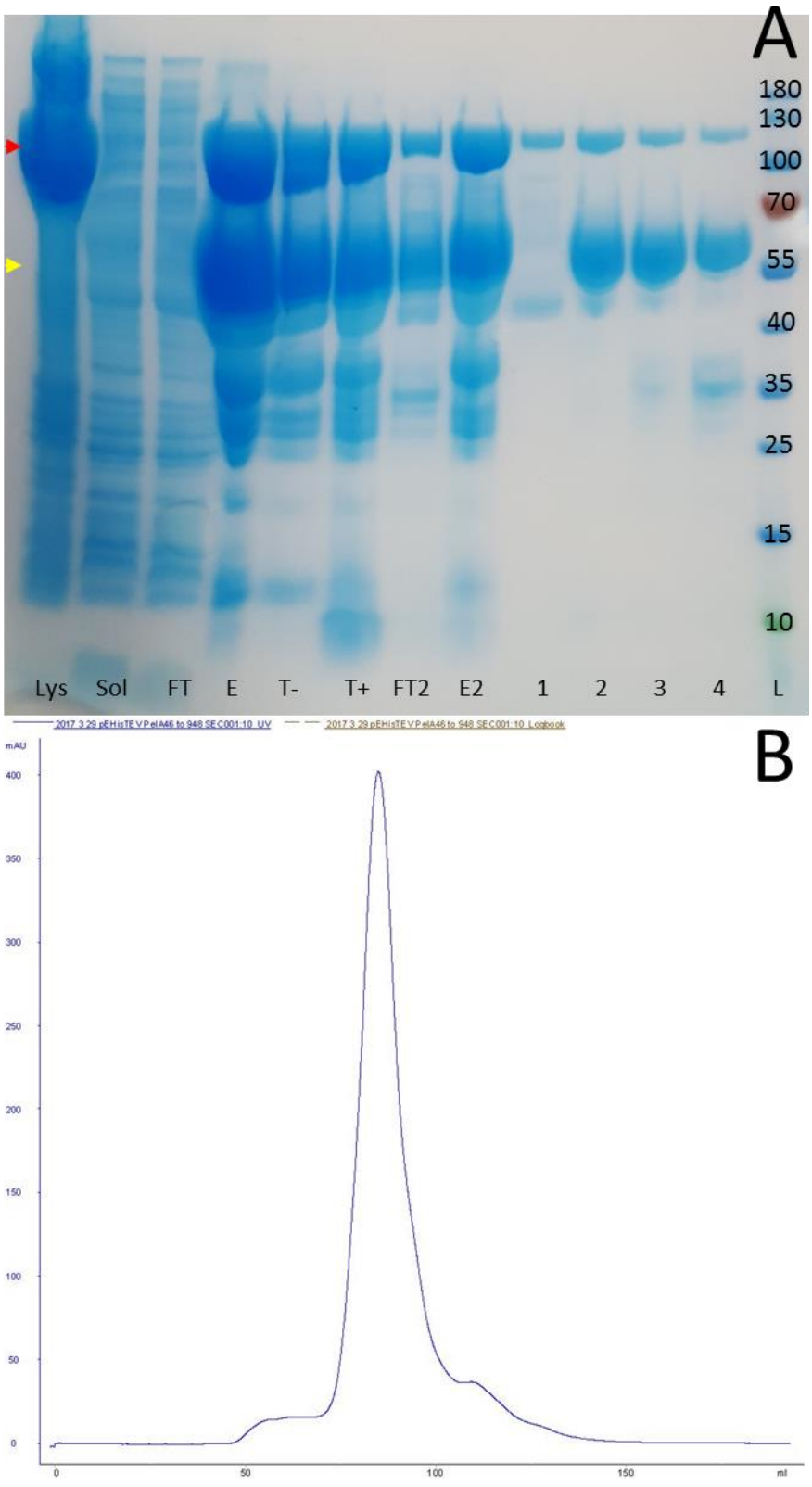

Figure 29: Purification trial of PelA 46-948 pEHisTEV Rosetta2**. A, SDS PAGE Overview of Purification:** Cell Lysate (Lys), Soluble fraction (Sol), Flow through (FT), HisTRAP elution (El), Pre-TEV digest (T-), Post-TEV digest (T+), 2nd IMAC flowthrough (FT2), 2nd IMAC elution (El2, tag), size exclusion fractions at 80 ml (1), 86ml (2), 89ml (3) and 92ml (4), PageRuler™ 10-180kDa Ladder (L,). Expected MW of PelA 46-948 is 105

kDa (red arrow). Contaminant has a MW of ~55 kDa (yellow arrow). **B, Size Exclusion Chromatograph of PelA.** Chart shows UV. The single peak represents elution of both PelA 46-948 and the contaminant protein.

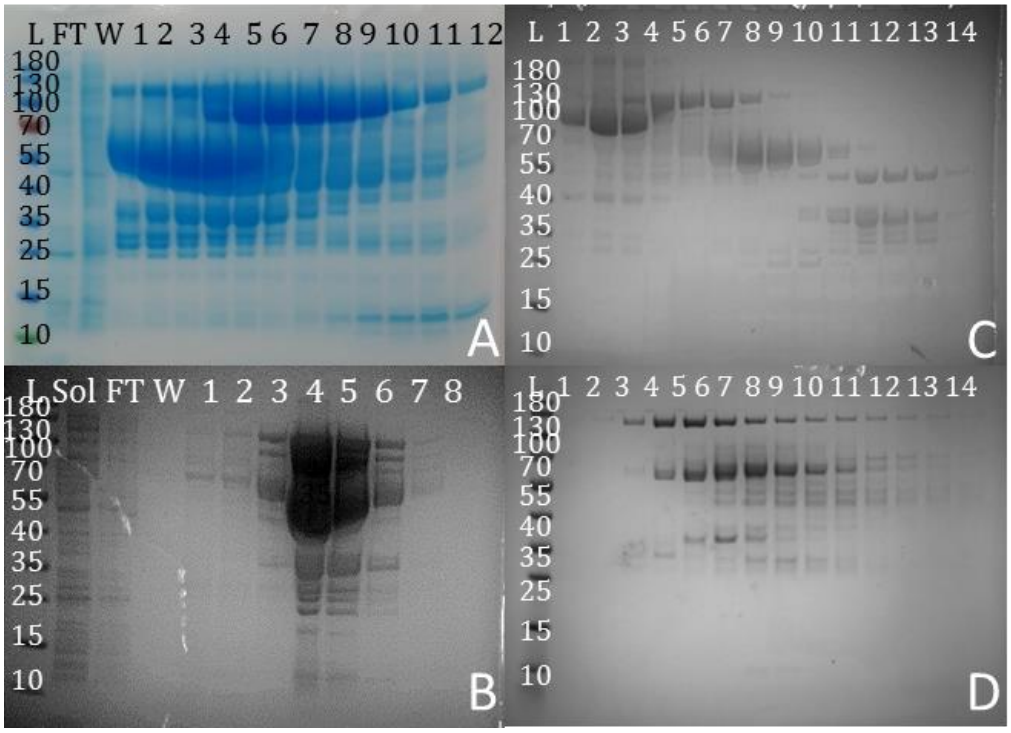

Figure 30: SDS PAGE gels showing separation of PelA 105 from PelA 55**. A: 5 column volume HisTRAP gradient elution.** PageRuler™ 10-180kDa Ladder (L), HisTRAP flow through (FT), column wash (W), elution from HisTRAP column at 12 (1) to 34 (12) ml in 1 ml increments following start of gradient. While PelA 55 elutes earlier than PelA 105, the residual is more than sufficient to contaminate PelA 105. **B: 30 column volume HisTRAP gradient elution.** PageRuler™ 10-180kDa Ladder (L), Soluble Fraction (Sol), HisTRAP flowthrough (FT), Column Wash (W), HisTRAP elution at 10 ml (1) to 80ml in 10ml increments following the start of the gradient. No separation observed between PelA 55 and PelA 105. **C: Size Exclusion using a 75 PG Column.** PageRuler™ 10-180kDa Ladder (L), fractions from the start of the peak at 55 ml (1) to the end at 88ml (14) in 3ml increments. Despite some separation the two species remain contaminated with one another. **D: 20 column volume Anion Exchange gradient elution.** PageRuler™ 10-180kDa Ladder (L), fractions from the start of the peak at 54ml (1) to the end of the peak at 87ml (14) in 3 ml increments. No apparent separation of species.

It was therefore decided that better separation of PelA 46-948 and PelA55 might produce well diffracting crystals of either species. The appearance of PelA55 coincides with the elution from the HisTRAP column (see Figure 29), therefore this might be a viable stage at which to separate the two species. This was attempted using a 5 CV gradient elution from a HisTRAP column (see Figure 30A), which caused some separation as PelA55 eluted at greater concentration at lower imidazole while PelA 46-948 experienced a delay in elution to higher concentrations of imidazole. This was not enough

to generate pure samples later in the purification so a 30CV gradient was also attempted (see Figure 30B) but failed to resolve the species. Raising imidazole concentration to the point at which PelA55 begins to elute, maintaining this until this protein was removed and then raising to elute PelA 46-948 was considered but both species begin to elute at same point and so this would produce the same result. For the final size exclusion chromatography step, the 200 PG column was exchanged for a 75PG column as this should provide greater resolving power between the molecular weights of 55 and 101 kDa but, while an improvement, this was still unable to resolve the two species fully (see Figure 30C). These improvements, combined with an increase to a 50CV IMAC elution gradient and the addition of a 20CV anion exchange step before size exclusion chromatography, were unable to prevent co-elution of these proteins (see Figure 30D). Removal of glycerol in the SEC buffer also failed to improve crystal formation.

3.7 The contaminant PelA55 is PelA 46-523

The lack of success in resolving these proteins suggests that there may be an interaction between PelA55 and PelA 46-948 and illustrates the need for a new approach. A sample of PelA55 was analysed by mass spectrometry to determine its identity. MASCOT[299] search of the peptide fragments generated by trypsin digest matched PelA55 to PelA with a confidence of 1.9×10^{-19} over the region of 88-535, with a further single peptide covering 827-838. This is strong evidence that PelA55 is an N-terminal segment of PelA. The later peptide might be explained by residual PelA 46-948 within the sample and suggests that the 827-838 region may be involved with an interaction with PelA55. Intact mass spectroscopy estimated the mass of the PelA55 fragment as 52752 Da (see Figure 31), a perfect match for single-ionised PelA 46-523.

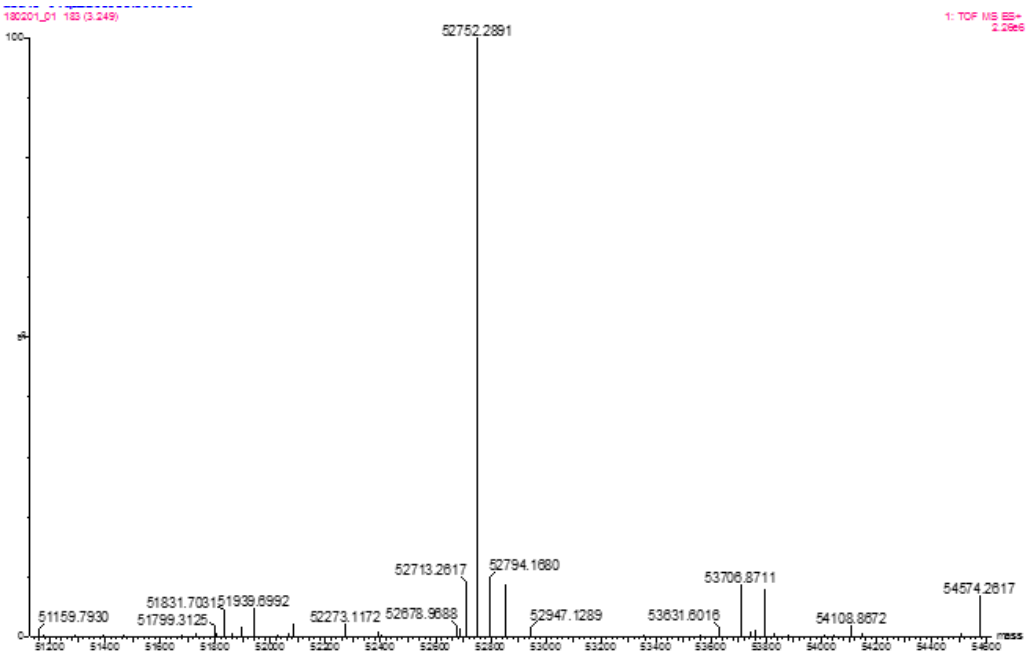

Figure 31: Mass spectrometry of PelA55. The fragment with the highest incidence is 52752 Da, matching to the molecular weight of PelA 46-523. Higher molecular weight masses detected may correspond to post translational modifications of the protein or mild contamination with PelA 46-948.

Disorder prediction using DISOPRED2[296] identifies three regions of potential disorder within PelA 46-948; in the region between 507-523, 523-540 and within 920-948. As we have identified a stable region of PelA in PelA 46-523 and disordered regions tend to inhibit crystal formation, it was decided to generate ten partial constructs of PelA to express the N and C-terminal regions of the protein (see Figure 32) in addition to the C-terminus of PelE following the end of the second predicted transmembrane helix (94-329). Small disordered regions are predicted at either side of the 523 domain boundary and at the C-terminus of the full length protein that would interfere with the crystallisation of the protein due to their flexibility. N-terminal PelA constructs were therefore generated terminating at residue 507, before one of these regions. The stability of the C-terminal domains, those following the 523-domain boundary, is less clear, therefore a combination of constructs that respectively omitted or contained these three predicted disordered regions were made. As PelA 46-523 has been observed in my purifications it can be proposed that this represents a stable section of the protein containing the predicted N-terminal glycoside hydrolase domain and an additional domain of unknown function. Isolation of this second domain, predicted to occupy the region of PelA 262-523, could therefore be very informative about the functionality of PelA.

Residues	507-920	507-948	523-920	523-948	544-920	544-948	262-707	46-262	46-507	46-303
Construct	1	2	3	4	5	6	7	8	9	10

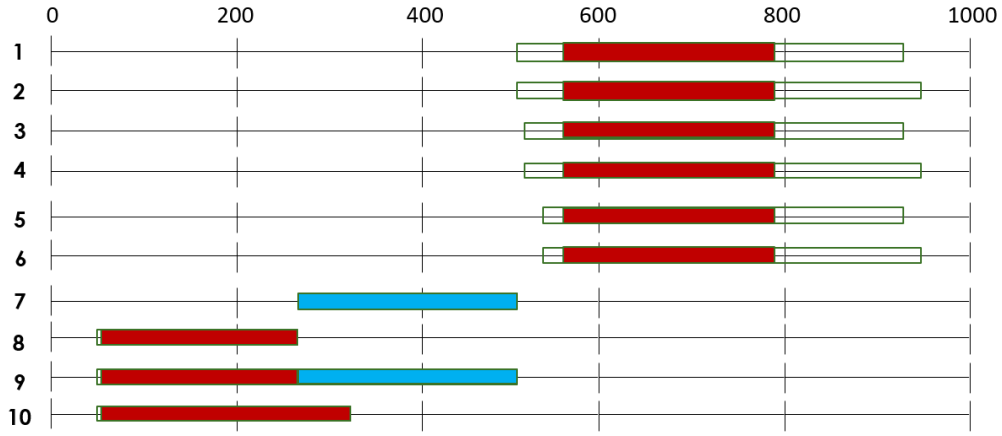

Figure 32: Constructs of PelA for expression of fragments of the protein. Red bars indicate established glycoside hydrolase and de-N-acetylase domains of PelA, blue bars represent the theoretical domain boundary identified by mass spectrometry, open bars represent protein with no predicted tertiary structure. Key represents position of amino acid residues relative to the full sequence. C-terminal constructs (1-6) begin at residues 507 (1, 2), 523 (3, 4) or 544 (5, 6) and terminate at residues 920 (1, 3, 5) or 948 (2, 4, 6). N-terminal constructs (7-10) begin at residue 262 (7) or 46 (8-10) and terminate at residues 507 (7, 9), 262 (8) or 303 (10). Constructs 1-9 were cloned together while 10 was cloned later for assays.

3.8 Expression of PelA Constructs

The expression of the ten constructs was trialled and it was found that none of the six C-terminal constructs were soluble (see Figure 33 A and B). Of the three N-terminal constructs, PelA 262-507 and PelA 46-262 were

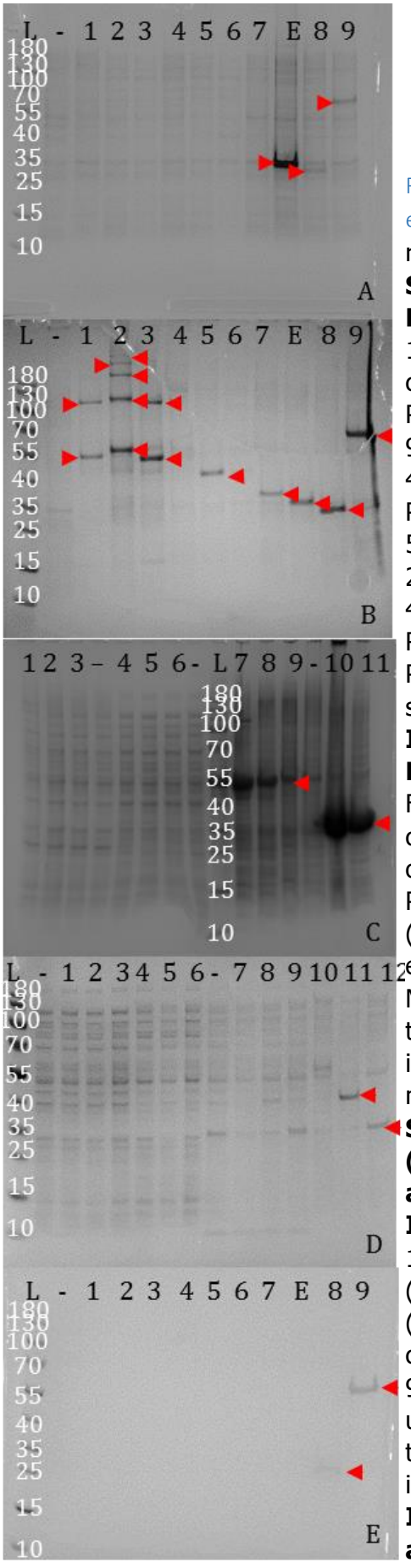

Figure 33: SDS PAGE of PelA construct expression trials**.** Red arrows show recombinant protein bands. **A: Soluble fractions of PelA (and PelE) constructs.** PageRuler™ 10-180kDa Ladder (L), No plasmid control (-), PelA 507-920 (1, 46kDa), PelA 507-948 (2, 49kDa), PelA 523-920 (3, 45kDa), PelA 523-948 (4, 48kDa), PelA 544-920 (5, 42kDa), PelA 544-948 (6, 46kDa), PelA 262-507 (7, 27kDa), PelE 94-329 (E, 26kDa), PelA46-262 (8, 24kDa), PelA 46-507 (9, 51kDa). Expression in Rosetta2 cells. Only PelE 94-329 (E), PelA 46-262 (8) and PelA 46-507 (9) show evidence of solubility. **B: Insoluble fractions of PelA (and PelE) constructs.** Labelled as per Figure 3.6A. Expression in Rosetta2 cells. Strong insoluble expression observed for all constructs other than PelA 523-948 (4) and PelA 544-948 (6) suggesting that proteins are expressing as inclusion bodies. Multiple bands are present for C-terminal constructs, suggesting incorrect disulphide bond formation may be leading to folding issues. **C: Soluble (left) and insoluble (right) fractions of PelA 523-920 and PelA 262-507 at different IPTG concentrations**. PageRuler™ 10-180kDa Ladder (L), PelA 523-920 (1-3, 7-9, 45kDa) and PelA 262-507 (4-6, 10-11, 27kDa) at IPTG concentrations 100-1μM (1-3, 4-5, 7-9) and 100-10μM (10-11) with uninduced controls (-). This confirms that IPTG concentration did not influence protein solubility. **D: Low IPTG Expression of select Pel A and E constructs.** PageRuler™ 10-

507 was soluble. The insoluble expression of PelA 262-507 and 523-920 were confirmed to not be IPTG dependent (see Figure 33C). The PelE 94-329 construct was initially found to be insoluble (see Figure 33D) but was later found to be soluble (see Figure 33A), but to not bind to IMAC beads (see Figure 33E). Interestingly, a small amount of PelA 46-262 may also be present in the IMAC fractions (see Figure 33E), suggesting that this may be recoverable. Further expression trials of PelA 46-262 and 46-507 under varying IPTG, temperature and time conditions curiously found both to be insoluble (see Figure 34A and B), in contrast to the earlier trial where PelA 46-507 to be soluble. The defining difference between these trials, and between those where PelE 94-329 was found to be varyingly soluble or insoluble, is that in the trials where proteins were soluble the lysis method used was pressurised homogenisation whereas in the trials where they were insoluble sonication was used. This suggests that overheating of the sonication probe may be causing protein misfolding resulting in sequestration or aggregation into the membrane fraction. When these trials were repeated using pressurised homogenisation as the lysis method it was found that PelA 46-262 remained insoluble while PelA 46-507 was soluble (see Figure 34C and D). This also provides limited evidence that PelA 46-507 and PelE 94-329 are heat sensitive.

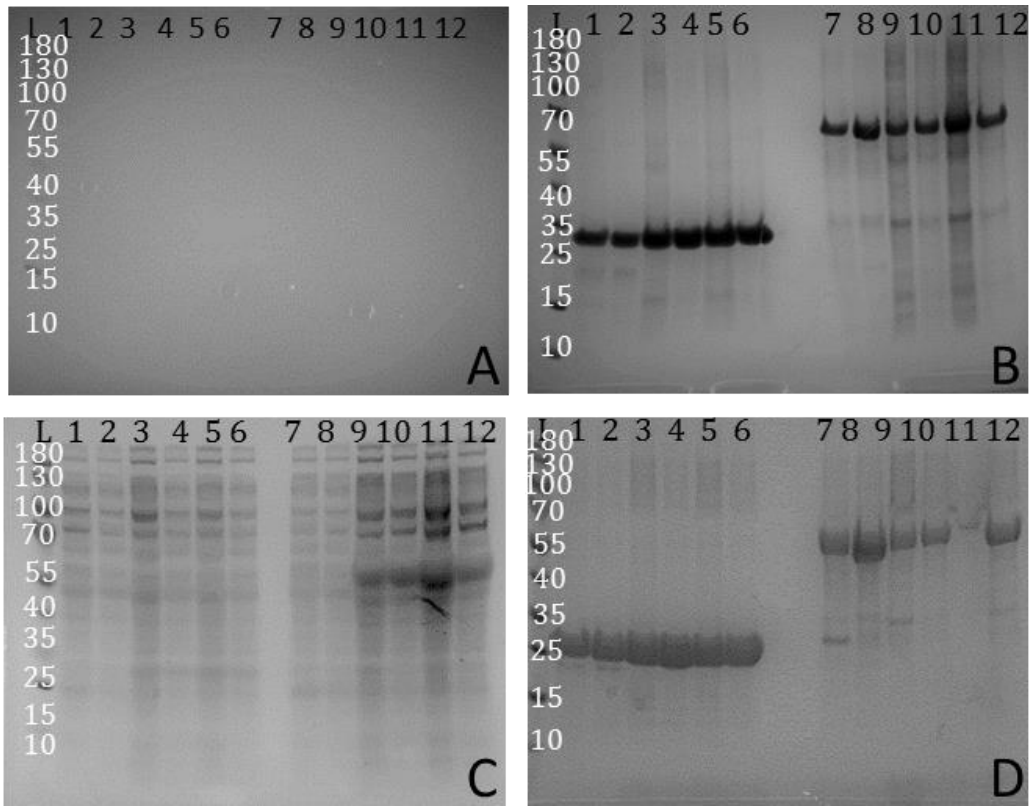

Figure 34: SDS PAGE of Expression Trials of PelA 46-262 and 46-507. A: Soluble fractions of PelA 46-262 and 46-507 lysed by sonication showing no soluble expression. B: Insoluble fractions of PelA 46-262 and 46-507 lysed by sonication showing internalisation into inclusion bodies. C: Soluble fractions of PelA 46-262 and 46-507 lysed by pressurised homogenisation. D: Soluble fractions of PelA 46-262 and 46-507 lysed by pressurised homogenisation showing soluble expression. ***Key:*** *PageRuler™ 10-180kDa Ladder (L, kDa), PelA46-262 (1-6, 24kDa), PelA 46-507 (7-12, 51kDa), 10-100μM IPTG (1-2, 3-4, 5-6, 7-8, 9-10, 11-12), 4 hour incubation at 37^{0}C (1-2, 7-8), 20 hour incubation at 16^{0}C (3-4, 9-10) or 20^{0}C (5-6, 11-12).*

3.9 Purification from inclusion bodies does not produce well folded protein

As PelA262-507 represents a product of interest, it was decided to further investigate whether this represented correctly folded protein. Proteins that are over-expressed are often trapped in membranous vesicles known as inclusion bodies (IB) that form as a cellular response to misfolded or

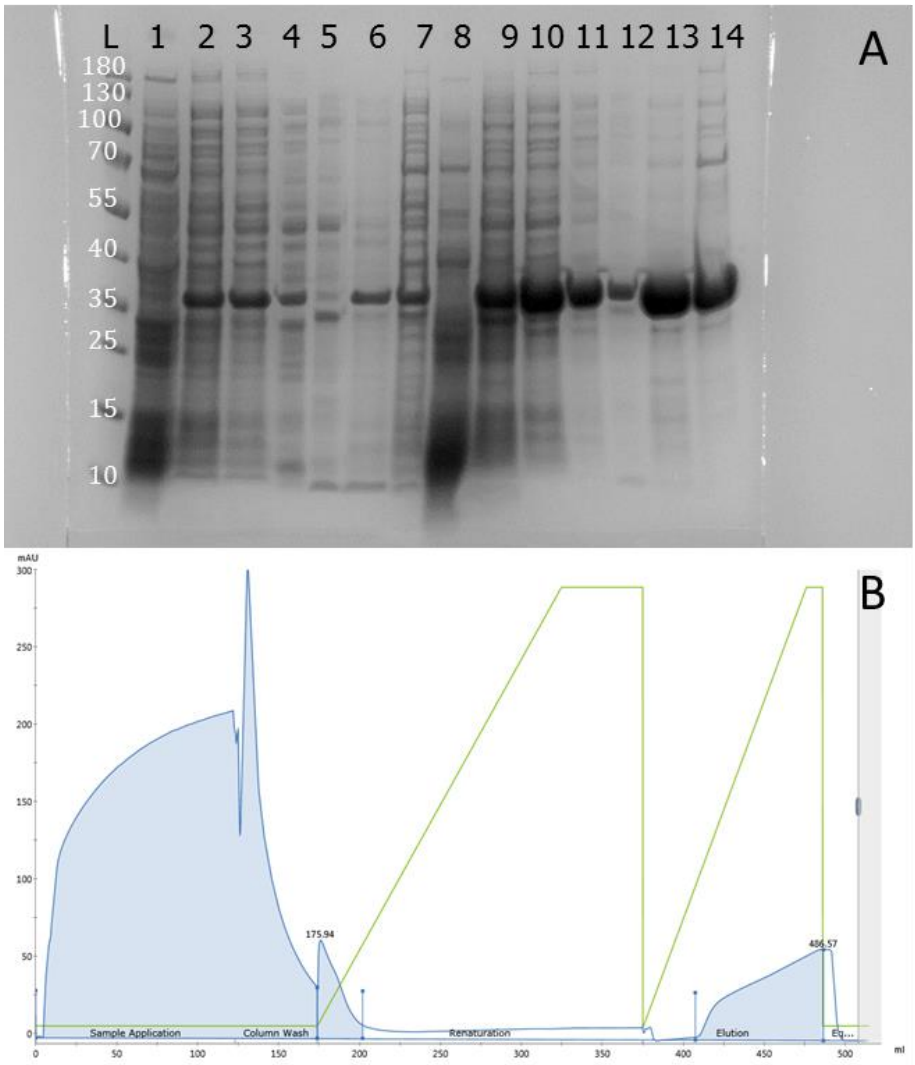

Figure 35: Purification Trials of PelA 262-507 from Inclusion Bodies. ***A, Solubilisation of PelA 262-507 using mildly denaturing conditions:*** *PageRuler™ 10-180kDa Ladder (L), uninduced control (1-7), PelA 262-507 (8-14, 27kDa), Soluble fraction (1, 8), Membrane washes (2-3, 9-10, 2M urea, 1%v/v triton X-100 buffer), Water washes (4-5, 11-12), Buffer wash (6, 13, 1mM β-mercaptoenthanol and 0.2% w/v sarcosyl buffer) to remove sample, 8M urea wash (7, 14) to see what's left. A large portion of PelA 262-507 can be isolated using this weakly denaturing protocol.* ***B: Loading, refolding and elution from 5ml HisTRAP affinity column.*** *X-axis shows volume in ml. Y-axis shows UV absorbance at 280nm (blue, mAU). Green line represents change in buffer concentration from 0.2% sarcosyl to 0% and then from 30mM imidazole to 300mM imidazole in buffers. The broad, misshapen peak during elution is indicative of heterogeneous binding of protein to the column associated with misfolded protein.*

aggregate protein. These are typically considered to contain denatured or unfolded protein, however, more recent evidence[307] suggests that proteins found within IBs may be present in a near-natively folded state, extremely pure and can often biologically active. To test whether this might be the

case for this protein, it was purified using mildly solubilising conditions that were able to extract the protein from their containing IBs (see Figure 35A). While extraction using this method was successful at isolating the protein, and illustrating the leaky expression of the pDEST17 plasmid, once the protein was applied and eluted from an IMAC column the protein produced a broad cliff-face rather than the expected sharp peak, indicating that the protein is most likely not sufficiently homogenously folded for the purposes of crystallisation (see Figure 35B). Furthermore, the protein precipitated overnight during tag removal, producing a final peak of less than 5 mAU (data not shown), indicating that further technical problems would need to be overcome before this method would be productive.

3.10 PelA 46-507 produces diffracting crystals

When expressed at large scale, PelA 46-507 remained soluble and purified free of contaminants (see Figure 36). PelA 46-507 crystallised in 0.1M Tris

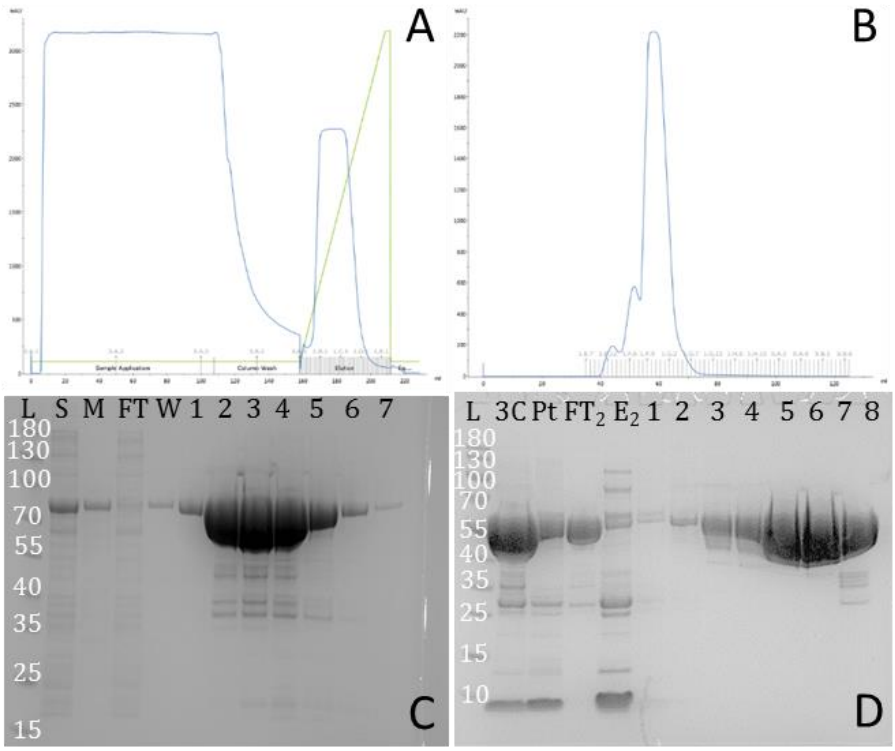

Figure 36: Purification of PelA 46-507. ***A: Chromatograph showing loading and elution of PelA 46-507 from a 5ml HisTRAP Affinity column.*** *X-axis shows volume in ml and the y-axis shows absorbance in mAU from 0 to 3000. The blue line shows UV at 280nm and the green shows % of 30-300mM imidazole. Phases and fractions are labelled. The peak between 165 and 200ml corresponds to PelA 46-507.* ***B: Chromatograph of size exclusion chromatography of PelA 46-507.*** *X-axis shows volume in ml and the y-axis shows absorbance in mAU from 0 to 2300. The blue line shows UV at 280nm. Fractions are labelled. The highest peak corresponds to purest PelA 46-507.* ***C: SDS PAGE gel of day 1 of the PelA 46-507 purification:*** *PageRuler™ 10-180kDa Ladder (L), soluble fraction (S), membrane pellet (M), HisTRAP flowthrough (FT), HisTRAP wash (W), HisTRAP elution from 160 (1), 168 (2), 176 (3), 184 (4), 192 (5), 200 (6) and 208 (7) ml. These fractions correspond to the peak in A with protein of interest at 70kDa. Contaminants are present but minor.* ***D: SDS PAGE gel of day 2 of the PelA 46-507 purification:*** *PageRuler™ 10-180kDa Ladder (L, 180, 130, 100, 70, 55, 40, 35, 25, 15, 10), Supernatant from overnight 3C protease digest (3C), precipitate from digest (Pt), untagged flow through 5ml HisTRAP column (FT_2), tag, contaminant and undigested protein eluted from column (E_2), fractions from HiLoad 16/600 Superdex 75 PG column (see B) at 45 (1), 48 (2), 51 (3), 54 (4), 57 (5), 60 (6), 63 (7) and 66 (8) ml. Fractions 57-63ml were concentrated for crystallisation.*

pH 8.5, 0.2M Trimethylamine N-oxide, 20% w/v PEG 2000 MME at 21°C. Crystals were slow growing, being observed after 42 days. Crystals were sent to the Diamond Light Source synchrotron and X-ray diffraction data was collected to 2.61 Å resolution. The structure was solved using molecular replacement using the crystal structure of the putative glycosidase tm1410 from *Thermotoga maritima* (PDB: 2AAM) as the strongest predicted homology modelling match[125]. Model data is shown below in Table 4, as generated by the Phenix software suite.

Table 4 PelA N1N2 data collection and refinement statistics.

Wavelength	
Resolution range	59.68 - 2.609 (2.703 - 2.609)
Space group	P 62 2 2
Unit cell	119.358 119.358 355.012 90 90 120
Total reflections	1778681 (176213)
Unique reflections	46546 (4519)
Multiplicity	38.2 (39.0)
Completeness (%)	99.27 (99.38)
Mean I/sigma(I)	9.92 (1.61)
Wilson B-factor	37.99
R-merge	0.406 (3.562)
R-meas	0.4114 (3.609)
R-pim	0.06603 (0.5747)
CC1/2	0.995 (0.494)
CC*	0.999 (0.813)
Reflections used in refinement	46238 (4520)
Reflections used for R-free	2300 (213)
R-work	0.2720 (0.3897)
R-free	0.3107 (0.4373)
CC(work)	0.900 (0.560)
CC(free)	0.860 (0.460)
Number of non-hydrogen atoms	6631
macromolecules	6497
solvent	134

Protein residues	824
RMS(bonds)	0.006
RMS(angles)	0.93
Ramachandran favored (%)	90.55
Ramachandran allowed (%)	7.09
Ramachandran outliers (%)	2.36
Rotamer outliers (%)	0.30
Clashscore	10.77
Average B-factor	41.66
macromolecules	41.82
solvent	33.80
Number of TLS groups	1

Statistics for the highest-resolution shell are shown in parentheses.

The resolution of the diffraction is slightly below average, at 2.6 Å, which has contributed to a less than ideal model. While R-work is approaching an acceptable level for this structure (ideally less than 0.26), R-work and R-free have begun to diverge, suggesting that a flaw is present which it has not been possible to identify. This flaw is most likely present in the C-terminal region of the model where electron density was weakest. There are slightly more Ramachandran outliers present than would be ideal (2.36% vs. <2.0%) but is within acceptable bounds. All other statistics support that this is a generally strong model from which valid conclusions can be drawn.

3.11 PelA 46-507 is a multi-domain protein

The crystal structure contains two molecules in the asymmetric unit that each spans residues 47-507 and contains two domains (see Figure 37). The N-terminal domain (N1) starts at residue 46 and ends at residue 300 while the C-terminal (N2) domain is made up of residues 305 to 507. There are multiple potential flexible regions within the protein, as illustrated by breaks in the chain in the regions 103-128, 314-323, 413-423 and 485-489. While these are likely to represent disordered, and therefore undetected, regions

of protein, it cannot be ruled out that these gaps in electron density might be related to the long growth period of the crystal and be the

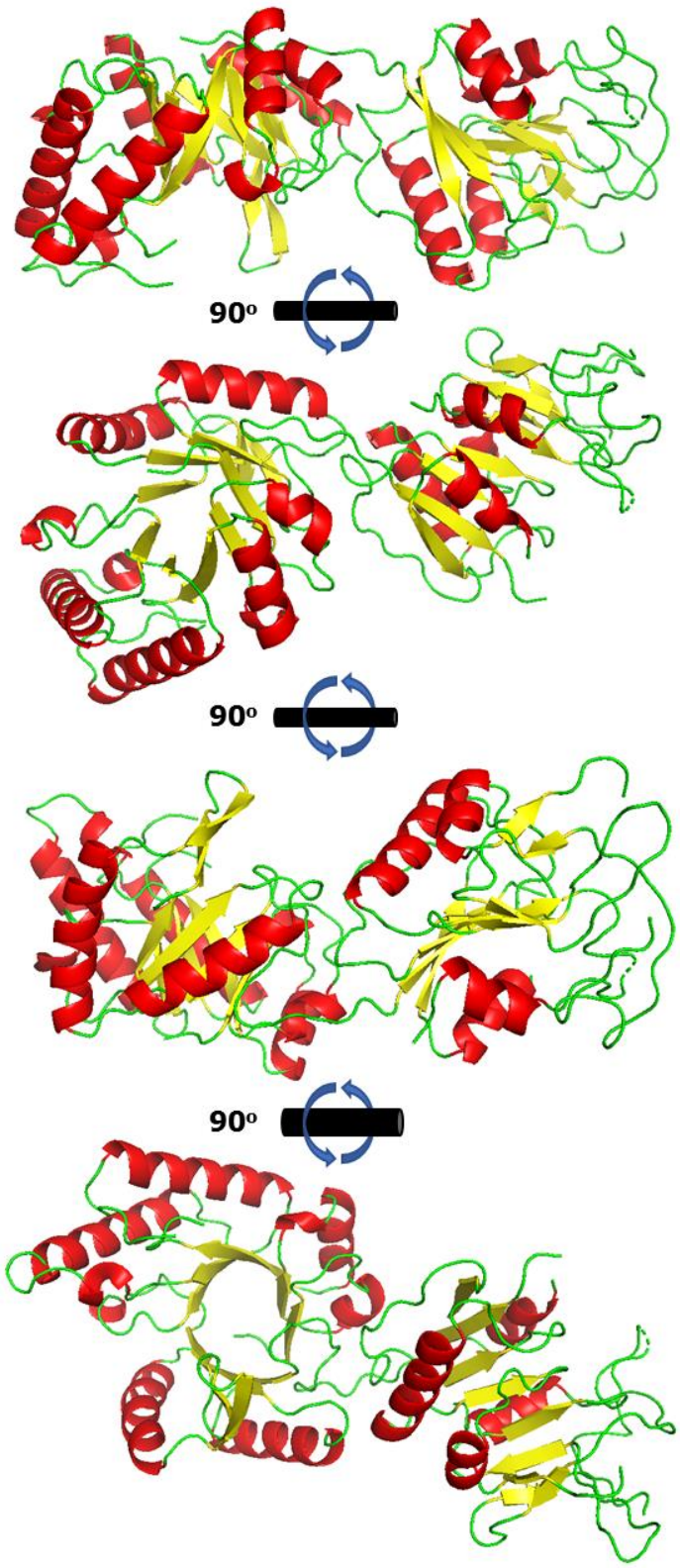

Figure 37 : Panoramic view of PelA 47-507. The N1 TIM Barrel domain is displayed on the left with the N2 GATase-like domain displayed on the right. The protein is displayed as a ribbon diagram with α-helices in red, β-sheets in yellow and loops in green. The protein is rotated by 90^0 on the x-axis between images.

results of some form of limited proteolysis.

3.12 PelA contains an unusual N-terminal TIM Barrel

The N-terminal region shows some of the conserved characteristics of a triosephophosphate isomerase (TIM) barrel with 7 α-helices sandwiching 8 β-strands to form a closed β-barrel. This is somewhat unusual as an eighth α-helix is not formed at the C-terminus, presumably as this is not necessary to enclose the β-barrel; this space is instead occupied with loops leading to the second domain of the protein. The sequence follows a canonical β/α folding pattern, with two additional antiparallel β-sheets forming a β-hairpin beneath the domain between β5 and β8 (see Figure 38A) which are unusual. β6 can be considered an extension of β5 but β7 is unusual in that it pushes outwards into a region that would expect to be occupied by stabilising α-helices and may serve a similar function.

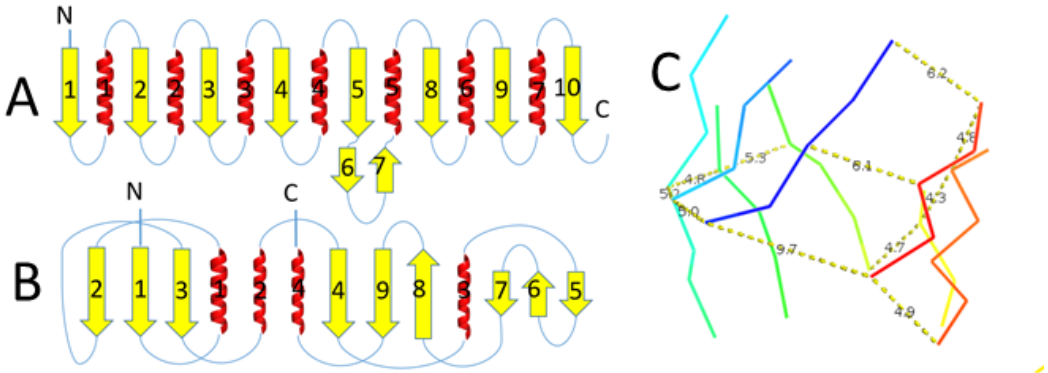

Figure 38: Topology of the crystal structure of PelA 47-507**. A: Topology diagram of the N1 domain of PelA.** The N1 domain takes on a β/α folding pattern characteristic of TIM barrels with two deviations; the β67-hairpin and lack of α8 at the C-terminus of the domain. B-sheets are displayed as yellow arrows, α-helices as red helices and connecting loops as blue lines. **B: Topology diagram of the N2 domain of PelA.** The glutamine amidotransferase-like family domain of PelA has a more complex folding pattern than the N1 domain and is predominantly formed of β-sheets and loops. See A for the key. **C: Bond distances between β-strands of PelA N1.** Strands are displayed as Cα backbones in Jones Rainbow going from blue at the N-terminus through the visible spectrum to red at the C-terminus. Cα-Cα interatomic distances are typically around 5 Å between sheets, which allows distances of ~3 Å between potential hydrogen bonding pairs within the barrel. The distance between β1-β10 are 6 Å and higher which prevents hydrogen bonds and opens the barrel at this point.

The ~75° tilt of the barrel is also non-canonical, compared to the expect 35°[308]. This is because the distance between equivalent α-carbons of β1 and β10 (equivalent to β8) vary from 6.1 - 9.7 Å whereas similar measurements between other sheets were in the range of 4.3 – 5.3 Å (see Figure 38C). A similar disparity in distance is apparent in potential

hydrogen bonding partners between sheets with the distances between β1 and β10 being in the range of 4.5 – 9.6 Å while remaining sheets are typically within 3.0 Å of one another, indicating that β1 and β10, unlike the remaining sheets, do not hydrogen bond with each other. This unusual arrangement causes the barrel to open up to with distance between α-carbons opposing sheets at the top of the barrel being ~15 Å apart and those at the bottom reaching distances of 20.5 Å whereas a typical β-barrel would have an expected diameter of approximately 14 Å[308]. The increased openness of the base of the barrel and associated unusual higher slant prevents a measure of the shear number[308], a measure of the extent the barrel is staggered, for PelA as there is no equivalent positioning of residues between sheets 1 and 10 and positioning does not carry beyond sheet 5 from sheet 1. This also suggests that this structure represents an open conformation of the enzyme for substrate binding at the base and raises the possibility that ligand binding would induce a conformational change that would restore hydrogen bonding between β1 and β10, as a shift of less than 2 Å would promote this. A similar mechanism has been shown to occur with methylmalonyl CoA mutase with the open conformation of the enzyme losing hydrogen bonding between β1 and β2 which is then restored in the closed conformation[309].

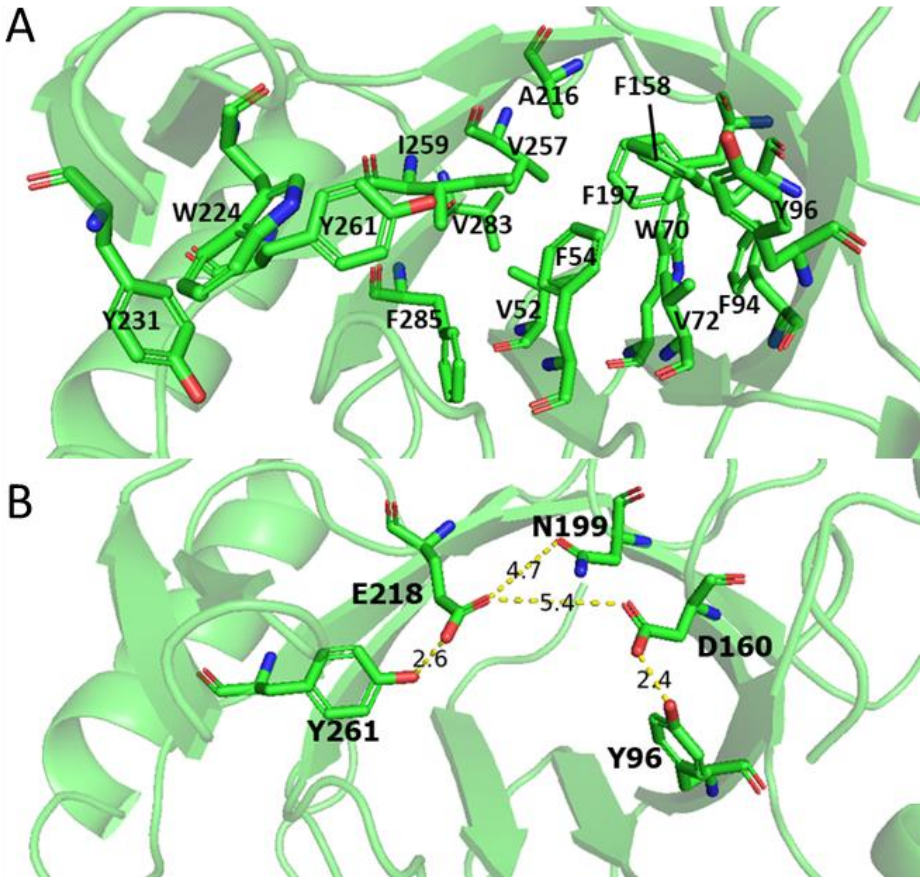

Figure 39: The active cleft of the PelA TIM barrel. **A: The hydrophobic core of the N1 TIM barrel.** The hydrophobic core of PelA's N1 TIM barrel stabilises the protein and forms the surface of the active site. It is

composed of V52, F54, W70, V72, F94, Y96, F158, F197, A216, W224, Y231, V257, I259, Y261, V283 and F285. The contribution of alternating residues is typical of a TIM barrel. **B: The putative catalytic site of the TIM barrel.** Residues E218 and D160 are likely to be catalytic and are supported by hydrogen bonds with Y96, N199 and Y261. Distances shown are in Å.

The active sites of TIM barrel proteins are positioned at the C-terminal ends of the β-strands. The interior of the PelA barrel forms a hydrophobic pocket due to the positioning of V52, F54, W70, V72, F94, Y96, F158, F197, A216, V257, I259, Y261, V283, and F285, which extends to β6 with W224 and Y231 above the lip of the barrel (see Figure 39A). A notable exception to this this trend is a small hydrophilic pocket made up of D160, N199 and E218 at the C-terminal tips of β4, β5 and β6 respectively which might reflect a catalytic site (see Figure 39B). Asp160 forms hydrogen bonds with Tyr95 and Asn199 while Glu218 is similarly supported by hydrogen bonds with Tyr 260. Such an arrangement has been observed to be typical of the active sites of glycoside hydrolases with the bulk of the tyrosine aiding sugar binding while the hydrogen bonds encourage an appropriate charge state[310]. A distance of 4.8 - 5.3 Å between carboxyl oxygens of general acid and base in retaining glycoside hydrolases would also support the interpretation of these residues forming an active site as the distance between D160 and E218 is 5.4 Å in what is proposed to be an open conformation of the enzyme.

The CAZy database[119] identifies five clans (GH-A, D, H, K and R) that adopt a $(\beta/\alpha)_8$ fold. Given the fold, it is likely that PelA belongs to one of the aforementioned clans and will therefore share a catalytic mechanism. These clans typically use an Asp or Glu residues in acid-base catalysis of the substrate, with the exception of GH-K which uses the carbonyl oxygen of the C-2 acetamido group of the substrate as the catalytic base. In brief, nucleophilic attack by the general base at the linking carbon occurs simultaneously with protonation of the glycosidic oxygen by the general acid forms a glycoenzyme transition state. The general acid then dissociates from this complex to form a covalent glycosyl-base enzyme intermediate then abstracts a proton from water to regenerate and produce a free solvent nucleophile which attacks the linking carbon of the enzyme intermediate to produce a cleaved saccharide retaining its original stereochemistry and a regenerated general base[310,311].

Interestingly, PelA is predicted to belong to the GH-114 family of endo-α-1,4-polygalactosaminidases[137] which has yet to be assigned a clan and for which the catalytic base and acid has not been identified[119]. This is functionally consistent with the known activity of the N1 domain[71,72,150] and this domain shows significant sequence identity (up to 74.4%) with numerous proteins that are predicted to belong to this family[248], including 38.1% sequence identity to the prototypical characterised example of this family[119,312], endo-α-1,4-polygalactosaminidase from *Pseudomonas sp. 881* (Uniprot ID: Q52423). This supports that this prediction is accurate and the

N1 domain represents the first structural characterisation of a GH-114 family protein at the time of this analysis.

3.13 PelA contains a glutamine amidotransferase-like domain

The second domain of PelA (PelA N2) contains 7-stranded β-sheets sandwiched between 4 α-helices. An additional 2 β-strands, β8 and β9, stack along a solvent exposed region of the protein alongside α2 and α3. Unlike the N1 domain, there is no consistent pattern of folding between the two elements of secondary structure (see Figure 38 B). B1-β5 show a parallel arrangement in a similar orientation to the β-barrel of N1 whereas the remaining two pairs of strands adopt an anti-parallel arrangement. Around these elements of secondary structure are a large number of loops, one of which (G428-V435) shows partial α-helical structure. The C-terminus of the protein exits the domain near its centre and is oriented such that the following domain would stack between and above N1 and N2 towards the closed face of the N1 β-barrel. Such an arrangement would place the carbohydrate esterase domain in this position which may allow the N1 domain to occlude the active site of the third domain until repositioned in response to interaction with PelB. While a potentially attractive mechanism for the regulation of PelA, there is as yet insufficient evidence to support this hypothesis as the orientation of the active site may well face away from the β-barrel.

The Dali server[313] can be used to interrogate similarity between a protein structure and existing structures. The recommended criteria for a significant match are a Z-score of (n/10)-4, where n is the number of residues in the query structure, or a Z-score greater than 2 with sequence identity of 20% or greater[314]. For the N2 domain, where (n/10)-4 is 13.7, the highest Z-score of 11.9 fails to meet this standard, although several matches were identified that shared sequence identity of 20% or more. Refining these weighting criteria, structures with a Z-score of 10.0 and sequence identity of 10 or higher and sequences with a Z-score of a minimum of 5.0 and sequence identity of 20% or higher were selected (see Appendix Table A4). Common themes amongst these proteins is that the high Z-score proteins show glutamine amidotransferase (GATase) related folds whereas the high sequence identity proteins tend to be the transcriptional element of a two-component signalling response regulator. This latter activity seems unlikely as the size and periplasmic localisation of PelA would prevent it from interacting with cytoplasmic DNA, this would however support a role in the binding a polymeric sugar. Oligomerisation is a common theme amongst both lists.

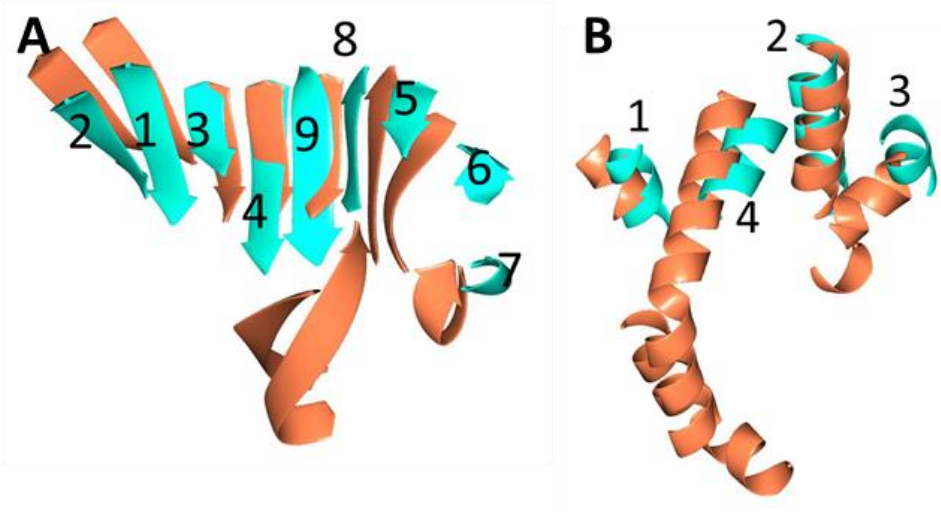

Figure 40: The N2 domain of PelA. A and B: The N2 domain of PelA (cyan) showing β-sheets (A) and α-helices (B) superimposed with the equivalent secondary structural elements from GATase TM1158 (Orange, PDB: 1O1Y). Secondary structural elements of PelA N2 are labelled.

Comparison with the GATase TM1158[315] reveals that the first 6 β-strands of PelA N2 (β2-1-3-4-9-8) are conserved while the remaining 3 β-strands (β7-6-5) are not (see Figure 40). There is conservation of α1, α2 and α4 but α3 adopts an almost perpendicular arrangement compared to the homologous structure and the active site is not conserved. Based on these similarities, it can be concluded that they adopt related folds and therefore that the N2 domain appears to adopt a class 1 glutamine amidotransferase-like superfamily fold. A notable feature of many of the proteins within this superfamily is an active site reliant upon a catalytic triad containing cysteine, a residue not found in the N2 domain. Domains similar to the N2 domain associated with oligomer formation have previously been reported in structures containing an N-terminal glycoside hydrolase domain[316–318].

3.14 The PelA N2 domain is required for oligomer formation

Given that the PelA N2 domain has a GAT-like fold but lacks the residues needed for GATase activity, a functional assignment for the domain was needed. One possibility is that it is involved in oligomerization of PelA in the bacterial periplasm. We therefore decided to investigate whether PelA 46-507 (PelA N1N2) was able to form oligomers *in vitro* using aSEC. This produced two peaks with elution volumes of 12.8 and 14.7mL corresponding to a 2.5-mer and monomer respectively (see Figure 41A), with the majority of the protein eluting as a monomer. That the oligomer peak is substantially smaller than that of the monomer suggests that interactions involving residues C-terminal to residue 507 may be important for stable oligomer formation in the full-length PelA protein. Indeed, formation of a stable complex in the *P.aeruginosa* periplasm may require further stabilising interactions with other proteins, for example, as has been demonstrated for PelA with the TPR domains of PelB[150].

To determine whether the presence of the PelA N2 domain is a necessary prerequisite for dimer formation it is necessary to firstly assess whether the PelA N1 domain alone is capable of oligomerization. The construct of the PelA N1 domain was found to be insoluble, however, it has recently been reported[72] that PelA 46-303 can be stably expressed, therefore an eleventh construct was prepared using these domain boundaries. This protein was expressed and purified as for PelA 46-507 (Figure 42). PelA 46-303 was found to produce a single peak from aSEC with an elution volume of 15.33 mL corresponding to a monomer (Figure 41A), thus confirming domain N2 as a necessary factor in determining the oligomerization activity observed for PelA N1N2 *in vitro*. To test whether the presence of the PelA N2 domain is a necessary and *sufficient* condition for oligomerization will require aSEC analysis of the N2 domain alone. However, all attempts to express recombinant N2 domain in *E.coli* led to formation of inclusion bodies.

We next decided to test to see if the crystal contacts between the N1 and N2 domains were biologically relevant. Combining fractions of PelA 46-303 (PelA N1) and PelA 46-507 (PelA N1N2) produced two peaks with elution volumes of 12.8 and 14.7mL. These peak volumes are representative of dimeric (2.5-mer) and monomeric PelA N1N2. Deformation of the tail end of the second peak reveals the location of PelA 46-303 (see Figure 41B)

and indicates that the elution of PelA N1 is not altered by the addition of PelA N1N2 and therefore that the two species do not interact.

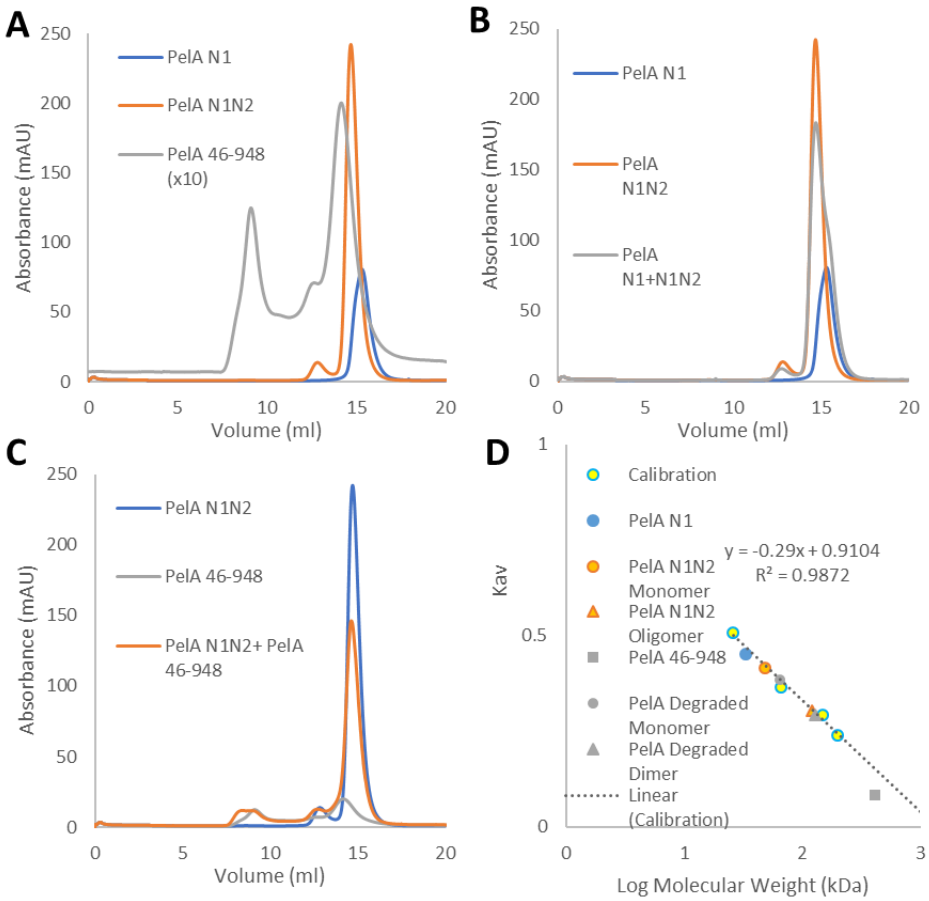

Figure 41 : Oligomerisation states of PelA. **A: Analytical gel filtration trace of PelA species.** PelA 46-303 (N1), 46-507 (N1N2) and 46-948 were independently passed through a gel filtration column. The trace for PelA 46-948 has been amplified 10x to allow visibility of the second peak. **B: Co-elution of PelA N1 and N1N2.** PelA N1 and N1N2 were run on gel filtration independently and combined at a 1:1 ratio. Deformation of the PelA N1 + N1N2 monomer peak suggests the two do not interact with one another. **C: Co-elution of PelA N1N2 and PelA 46-948.** PelA N1 and N1N2 were run on gel filtration independently and combined at a 1:1 ratio. Deformation of the peak at 9.11ml and corresponding lowering of the N1N2 peak at 14.7ml suggests an interaction between PelA N1N2 and PelA 46-948. Extension of this peak to the void volume suggests that such an interaction may be resulting in aggregation of N1N2. **D: Calibration of Superdex 200 Increase 10/30 GL column and fit of proposed oligomerisation states to calibration.** X-axis shows the $\log_{10}$ of the molecular weight of the standards (calibration) and the proposed complexes of PelA. The y-axis shows Kav, as calculated in chapter 3.2.11 and represents the point at which a peak eluted as a fraction of the volume of the column.

Having confirmed that an oligomer is formed by PelA 46-507 *in vitro*, we next wished to determine whether full length PelA complex is also able to form oligomers. aSEC experiments using PelA 46-948 resulted in the observation of three peaks with elution volumes of 9.11, 12.64 and 14.17

mL, where the first of these corresponds to an oligomer of PelA 46-948. This PelA 46-948 oligomer is outside of the range of accurate calibration but can tentatively be assigned a molecular weight of ~390-420 kDa which would presumably correspond to a tetramer. The remaining two elution peaks correspond to the major degradation product of PelA 46-948 (PelA55). This proteolytic fragment was the major contaminant of the sample injected to the chromatographic column. The peaks eluting at 12.64 mL and 14.7 mL correspond to a dimer and a monomer of PelA55, respectively (Figure 41A). The monomer form of PelA55 elutes at a slightly higher molecular weight than expected (observed at approximately 65 kDa rather than 55 kDa as expected) and the oligomer of this fragment elutes with an apparent molecular weight of ~130 kDa, indicating that this section of the protein is able to form dimers. It is therefore possible to tentatively assign dimerization or trimerization as a possible non-enzymatic function of the N-terminal domains of PelA.

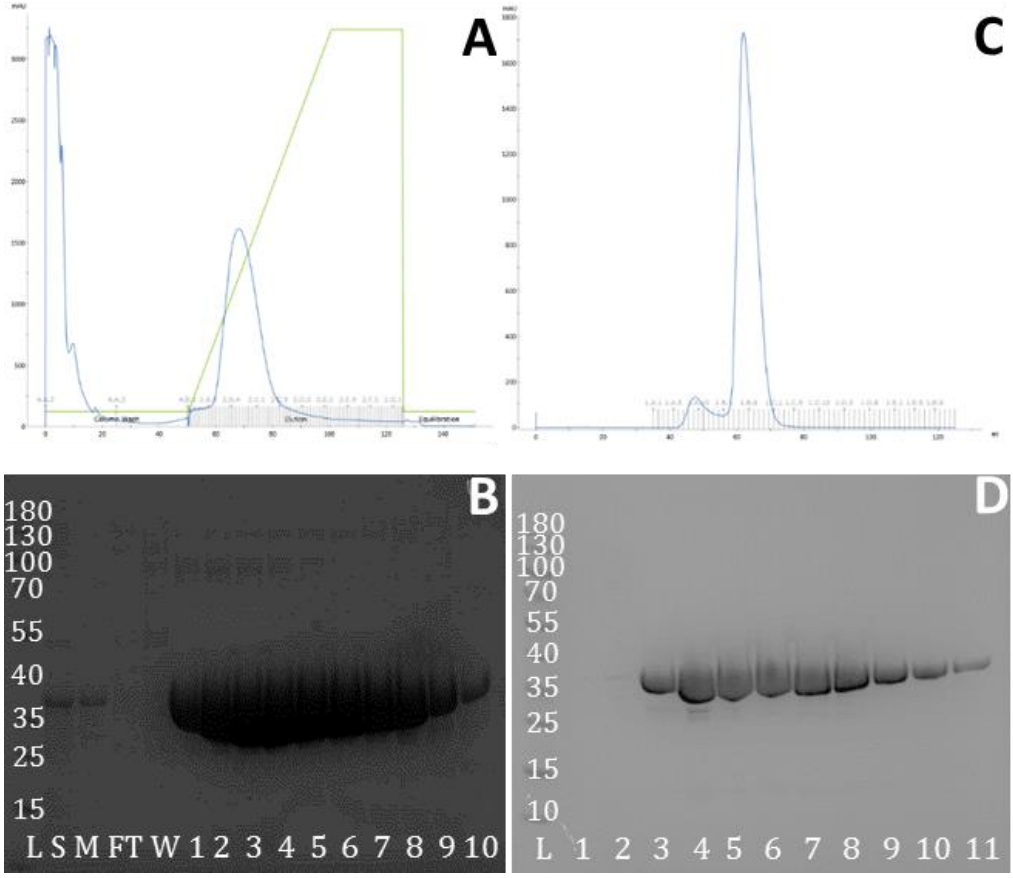

Figure 42 : Purification of PelA 46-303. **A: Affinity chromatograph of PelA 46-303.** The blue trace represents absorbance at 280nm from 0 to 3000 mAU in increments of 500. The green trace shows percentage of imidazole from 30-300mM. PelA46-303 elutes between 60 and 80ml. **B: SDS PAGE of PelA 46-303 Affinity Chromatography.** PageRuler™ 10-180kDa Ladder (L), soluble fraction (S), membrane pellet (M), HisTRAP flowthrough (FT), HisTRAP wash (W), HisTRAP elution from 58 (1), 61 (2), 64 (3), 67 (4), 70 (5), 73 (6), 76 (7), 79 (8), 82 (9) and 85 (10) ml. Expected molecular weight of PelA 46-303 is 33kDa. **C: Size exclusion chromatograph of PelA 46-303.** The blue trace represents absorbance

at 280nm from 0 to 1800 mAU in increments of 300. The first peak corresponds to aggregate protein and the second to PelA 46-303. **D: SDS PAGE of PelA 46-303 Size Exclusion Chromatography.** PageRuler™ 10-180kDa Ladder (L), 44 (1), 47 (2), 50 (3), 53 (4), 56 (5), 59 (6), 62 (7), 65 (8), 68 (9) and 71 (10) ml. Expected molecular weight of PelA 46-303 is 33kDa and can be seen at high purity from 50-71ml.

Evidence has been presented above suggesting that PelA 46-948 forms an oligomeric complex. To test whether regions of PelA C-terminal to the N2 domain contribute towards this complex, we incubated PelA 46-948 and PelA N1N2 and analysed the mixture by aSEC (see Figure 41C). We found a small, broad peak at the void volume of the column. This peak is most likely a result of protein aggregation but is not observed in other sample runs. This peak of high apparent molecular weight may correspond to aggregates of the complex between PelA 46-948 and PelA N1N2 and implies that regions of PelA C-terminal to N2 may also contribute towards the oligomerisation of the protein.

Chapter 4 - WaaB

4.1 V186 is important for the structural stability of WaaB

WaaB had previously been cloned into the pLou3 plasmid containing an N-terminal hexahistidine-MalE tag and the structures of the wild type protein have been solved in the apo and UDP-bound form using this construct[304]. Examination of these two structures found that the conformations of the side chains of V186, R188, K195, W243, P247, T273 and E276 changed upon UDP binding. Additionally, V186, R188, K195, I216, W243, P247, T273 and E276 are highly conserved in WaaB enzymes, suggesting a potential important role within in UDP-galactose binding and hydrolysis. Superposition of WaaB with the related glucosyltransferase WaaG which shows a high degree of similarity (RMSD of 3.26 Å over 358 aligned Cα atoms), revealed K195 and E268 to be potential catalytic residues and that G15, F120, Q194, G269 and F270 are equivalent to the residues of the glucose binding site of WaaG[302].

From this list the residues V186, Q194, K195, I216, W243, E268, T273 and E276 were selected for mutagenesis to determine their functional contribution to enzymatic activity and UDP-galactose binding. The pLou3 construct used to express WaaB for crystallisation was used as the template for site directed mutagenesis, the success of which was determined by Sanger sequencing performed by Source Bioscience. All eight mutants of interest (V186, Q194, K195, I216, W243, E268, T273 and E276) were successfully generated.

Initially, the mutant proteins and WT WaaB were expressed and purified using the protocol previously used to purify WaaB for crystallisation. In brief, this included a 4hr incubation at 37 ^{o}C following overexpression induction. Proteins were purified using a Ni^{2+} affinity column that bound the N-terminal His_6-MalE-TEV tag. The tag was then removed using His_6-tagged TEV protease and the protein passed through a Ni^{2+} affinity column again to remove tag and protease. Purity was further enhanced using size exclusion chromatography.

In this work, this method was found to produce surprisingly low yields of protein compared to the original purifications of the wild type protein. The mutant proteins, as well as the wild type, were found to precipitate once the MalE tag was removed. To prevent protein degradation from negatively affecting the assay, it was decided that the tag would not be removed for the assay preps; the overnight digestion and second IMAC step were removed to account for this. The protein produced was found to be pure and relatively stable (see Figure 43), except for the V186A mutant which was insoluble.

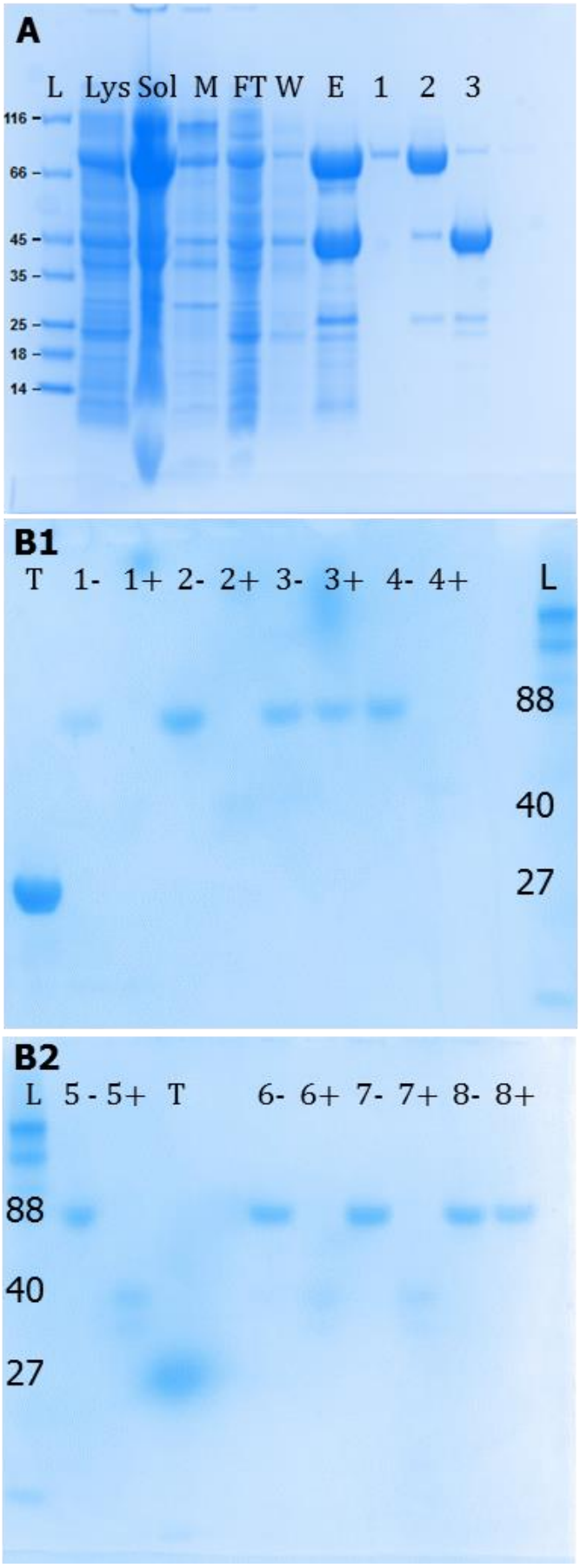

Figure 43: Purification of WaaB for biochemical assays. ***A: SDS PAGE of a typical WaaB purification (K195A).*** *Ladder (L), cell lysate (Lys), soluble fraction (sol), membrane fraction (M), HisTRAP flow through (FT), column wash (W), column elution (E), size exclusion chromatography peaks of aggregate WaaB (1, 88 kDa), tagged WaaB (2, 88 kDa) and maltose binding protein (3, 47 kDa). The resulting tagged WaaB is sufficiently pure for biochemical studies.* ***B1 and B2: SDS PAGE TEV digestion of WaaB.*** *TEV protease (T), conditions without (-) and with (+) TEV protease incubation for 2 hours. WaaB mutants Q194A (1), K195A (2), I216A (3), W243A (4), Wild Type (5), E268A (6), T273A (7), E276A (8). Ladder (L, in kDa). The presence of a His-tag was confirmed for all 27 and 80kDa bands using the InVision® His-tag stain. Bands correspond to tagged WaaB (88kDa), HisMalE tag (47kDa), untagged WaaB (41kDa) and TEV protease (27kDa). I216A and E276A show incomplete TEV digestion. Faint bands for both tag and WaaB are visible for the WT, K195A, E268A and T273A. No bands are visible for Q194A or W243A, indicating relative instability. The disappearance of the 88 kDa bands on addition of TEV protease shows that the purified protein contains the desired tag and the faintness of the digested band indicates instability of the untagged protein.*

4.2 Glycosyltransferase assay using all mutants

Glycosyltransferase assays were performed using the UDP-Glo® Glycosyltransferase Assay kit. This kit is an enzyme-linked assay that converts UDP, produced as a result of the hydrolysis of a UDP-sugar, into

ATP. ATP is then used by a luciferase enzyme to produce luminescence. Glycosyltransferases frequently hydrolyse their donor-substrates in the absence of an acceptor; this has been observed in the crystals of WaaB which were soaked with UDP-galactose as only enzyme and UDP were found in the final structure.

To determine whether the tag affected activity, the wild type enzyme was purified with and without the tag to serve as a control. Optimal temperature and time for TEV protease digestion were determined to help improve stability and used to confirm the identity of the mutant proteins (see Figure 43B) as TEV protease has a highly specific cleavage recognition sequence[319]. It was found that 2 hours at room temperature was ideal for incubation of WT WaaB.

Before performing the assay, several rounds of optimisation were performed to determine an appropriate concentration of WaaB and of UDP-galactose. These experiments were done alongside determinations of whether the MalE tag affected the activity of the WT enzyme. Comparisons were also made between protein flash-frozen in 30% w/v glycerol and protein flash-frozen without a cryoprotectant as prior investigation had found that glycerol interfered with the ability of the protein to form crystals and might therefore impact enzyme activity. The ability of D-glucose to act as an antagonist of UDP-glucose hydrolysis was also investigated as a potential additive for co-crystallisation.

It was found that the untagged enzyme was significantly less active than the tagged enzyme (see Figure 44A). This increase in activity is so substantial that it was deemed to more accurately reflect an increase in the relative stability of the enzyme than the results of a random co-operative enhancement of WaaB's activity. Given the low activity of the untagged enzyme it was deemed that it would be necessary to use the tagged enzyme for further work as the loss of stability under the *in vitro* conditions is more likely to represent a deviation from biological activity than the addition of the tag. A further experiment where purified tag was added to untagged WaaB showed no difference from the untagged WaaB alone, in keeping with the interpretation that the tag is only affecting the stability of the enzyme. In keeping with this decision, the highest recommended concentration of UDP-galactose (100 µM) and up to double the highest recommended mass of WaaB (200 ng/reaction) were selected to attempt to further boost the signal in later experiments.

Somewhat surprisingly, it was found that the presence of glycerol significantly reduced the hydrolytic activity of WaaB in comparison to WaaB flash-frozen without a cryoprotectant (see Figure 44B); preparations of enzyme for future assays therefore did not use a cryoprotectant. D-glucose

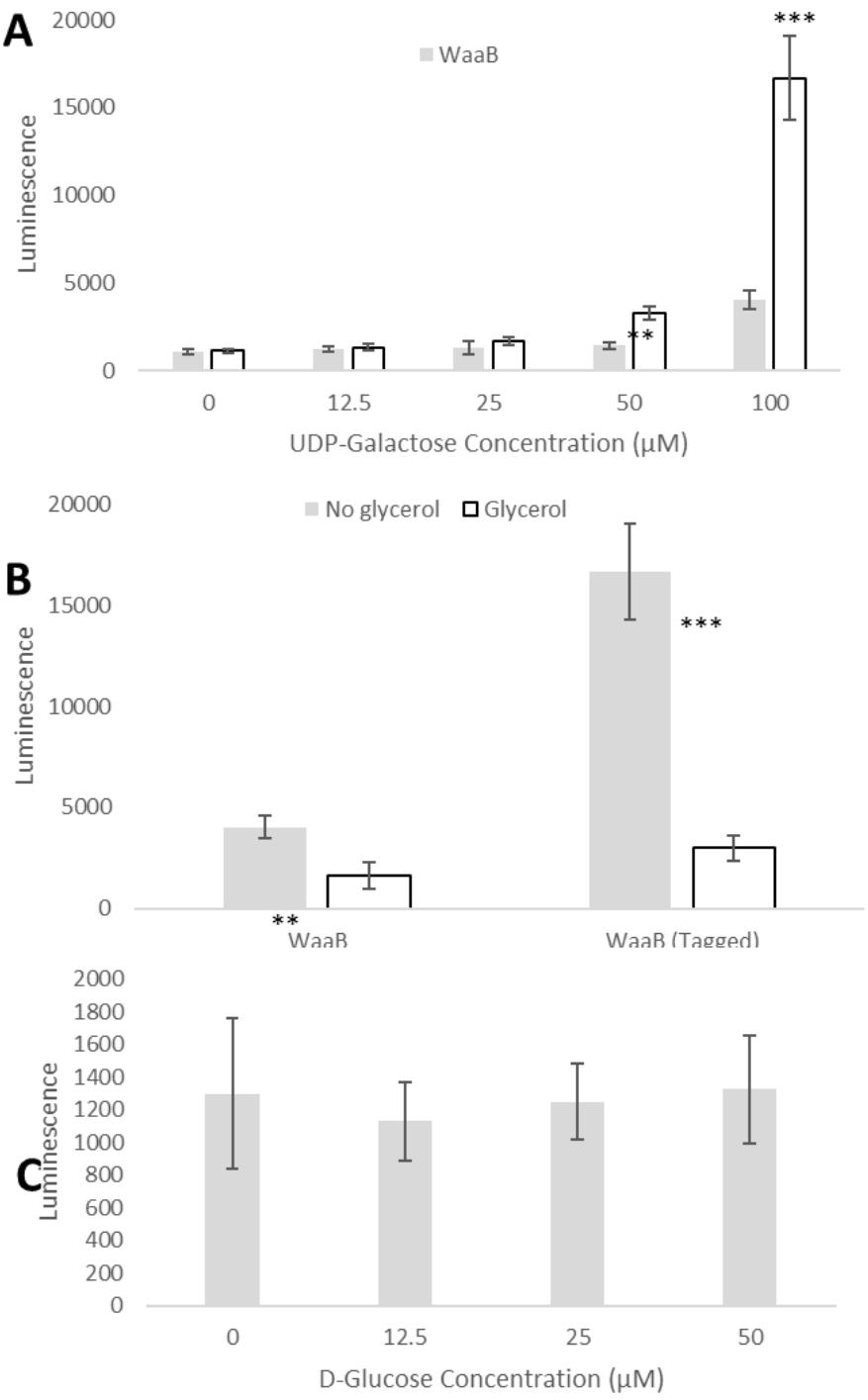

Figure 44 : Optimisation of conditions for UDP-Glo® assays of WaaB. ***A: Optimal UDP-galactose concentration and tag-state of WaaB***. *Columns show untagged wild type WaaB alongside wild type His_6MalE-WaaB. 50-100µM UDP-galactose is required to produce luminescence easily differentiated from the baseline. Error bars show standard deviation (n=5). * indicates significant difference between WaaB and WaaB (tagged) conditions with *=P<0.05, **=P<0.01 and ***=P<0.001.* ***B: Storage conditions of WaaB.*** *WT tagged and untagged WaaB was stored at -80^{0}C before assay use either in purification buffer or buffer containing 30% v/v glycerol as a cryoprotectant. Error bars show standard deviation (n=4). * indicates significant difference between WaaB with and without glycerol with *=P<0.05, **=P<0.01 and ***=P<0.001. This indicates that glycerol interferes with the activity of the enzyme.* ***C: D-glucose as an antagonist of UDP-galactose.*** *WT WaaB was incubated with different concentrations of D-glucose at 50µM UDP-galactose. Error bars show standard deviation (n=4). No significant differences were observed.*

was found to have no effect upon activity (see Figure 44C), indicating that it is not donor-substrate binding recognition is mediated by the UDP moiety and the D-glucose would be unlikely to reveal the relative position of UDP-glucose within the active site.

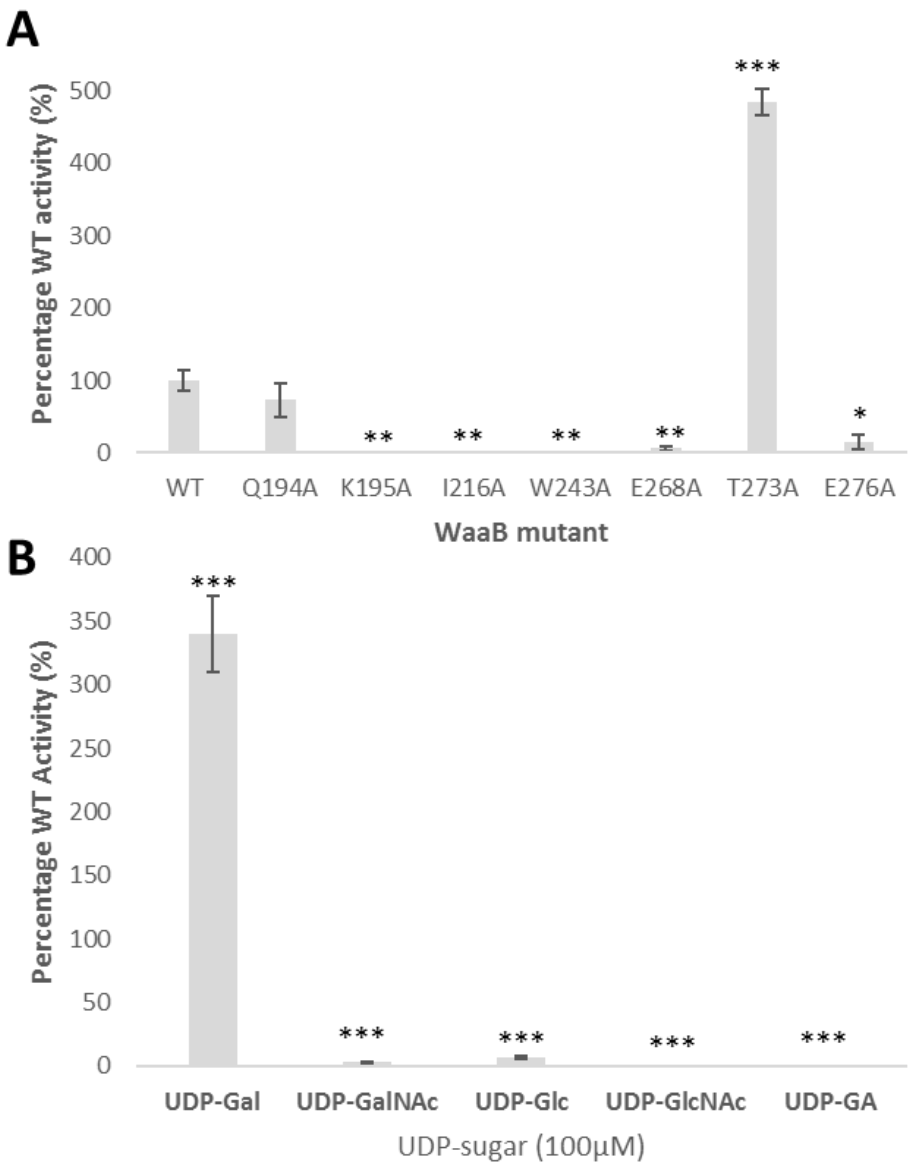

Figure 45 : Glycosyltransferase assays of WaaB mutants. ***A: Activity of WaaB mutants as a percentage of WaaB WT activity****. Error bars represent standard error (n=4). Activity at 200ng of protein is shown. * indicates significant difference between wild type WaaB and mutant WaaB conditions with *=P<0.05, **=P<0.01 and ***=P<0.001.* ***B: Activity of WaaB T273A with varied UDP-sugar donors as a percentage of WaaB WT activity with UDP-galactose.*** *Error bars represent standard error (n=15). These results are representative of four individual assays. * indicates significant difference between the activity of wild type WaaB with UDP-galactose and WaaB T273A with varied sugars (*=P<0.05, **=P<0.01 and ***=P<0.001.)*

Using this newfound information, the assay was repeated using WT WaaB and the 7 stable mutants Q194A, K195A, I216A, W243A, E268A, T273A and E276A along a WaaB serial dilution of 200 to 0.39ng (see Figure 45A).

It was found that the activity of K195A, I216A, W243A, E268A and E276A was significantly reduced, indicating that I216, W243 and E276 are invaluable in UDP binding and provides strong evidence for K195 and E268 being important for catalysis. An interesting and unexpected finding was that the T273A mutant shows a 4.8-fold increase in activity compared to the wild type. Several explanations for this phenomenon exist; it may the active site is more accessible, that the increased hydrophobicity of the hydrophobic pocket increases the strength of the hydrophobic interactions that draw UDP-galactose into the enzyme's active site or that the loss of hydrogen bonding from the threonine causes the strength of binding within the pocket to be weaker, allowing easier dissociation of the product from the active site.

4.3 Wild-type and T273A mutant WaaB are UDP-Gal selective

To determine whether the T273A mutant causes the active site to become more accessible, it was decided to repeat the assay for this mutant and the WT enzyme using different UDP-sugars. The existing assay kit is compatible with four other UDP-sugars that are common sugar-donors in glycosyltransferases; UDP-glucose (UDP-Glc), UDP-N-acetylglucose (UDP-GlcNAc), UDP-N-acetylgalactosamine (UDP-GalNAc) and UDP-glucuronic acid (UDP-GA). Both the wild type and T273A mutant of WaaB were unable to perform significant hydrolysis upon any of the tested UDP-sugars (see Figure 45B) other than its native substrate, UDP-galactose (UDP-Gal).

4.4 Optimisation of the WaaB purification for crystallisation

The most accessible method to determine the cause of the unexpected efficiency of WaaB T273A is X-ray crystallography, as a solution for the WT protein has already been solved then repeating this method by which this was obtained should provide a rapid and easy route to solve this problem. Interesting crystallographic data might also be found through the co-crystallisation of the catalytic mutants K195A and E268A with UDP-Gal as the degenerate active site should be less able to hydrolyse the substrate and a more complete picture of donor-substrate binding by WaaB might be identified.

Initial crystallisation trials using the WaaB produced for glycosyltransferase assays was not successful, with no growth of protein crystals. As previously discussed, the WaaB produced by my previous purifications is unstable in comparison to the protein originally produced for crystallisation. Examination of the thesis where this structure is described identified a discrepancy in that the buffer system described did not match the one in use, notably the addition of 10% w/v glycerol in Lysis, Wash and Elution buffers and a change in pH from 7.0 to 8.0. The purification of WaaB was carried out as previously described but continued to produce little enzyme once the MalE tag was removed. Examination of the publication manuscript for the crystal structure of WaaB identified a further discrepancy, with a buffer pH reported as 7.8. This was tried but, as expect for a minor

change, failed to alleviate the problem. It was, however, possible to produce a single low-resolution crystal of WaaB K195A at 6.11 Å.

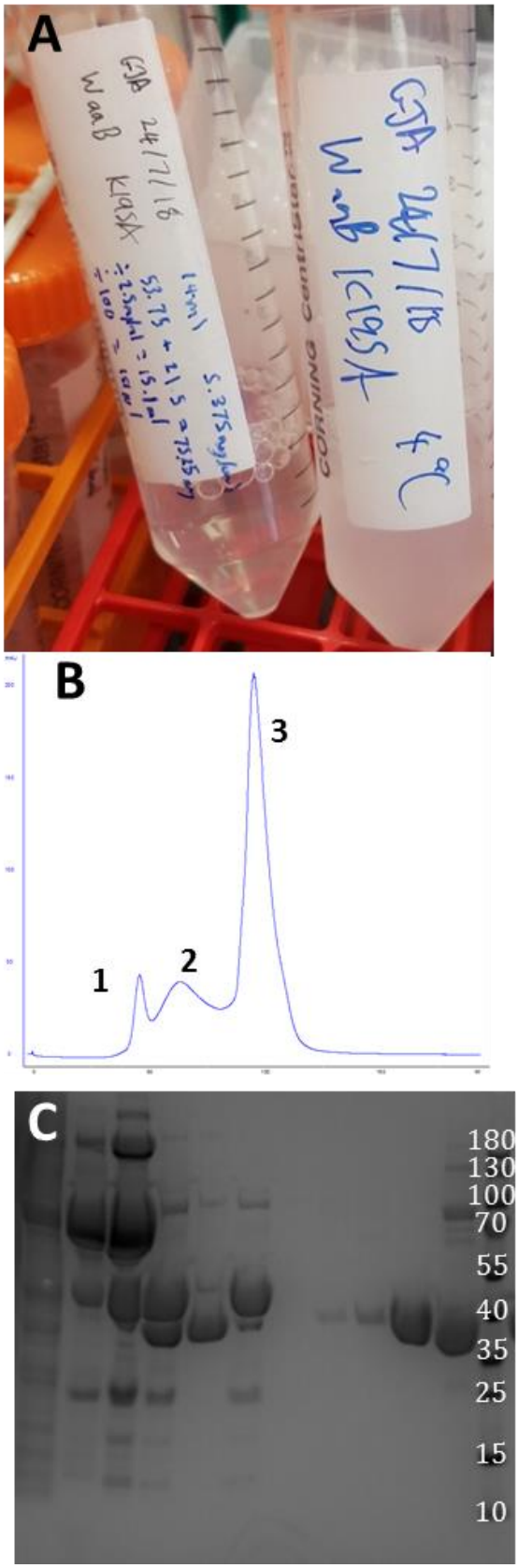

Figure 46: Improved purification of WaaB. ***A: WaaB is temperature sensitive.*** *Photograph of two equal volumes of WaaB K195A following an overnight digest with TEV protease at room temperature (left) and at 4^{0}C (right). The relative turbidity is visibly greater in the sample incubated at lower temperature reflecting a greater degree of protein precipitation.* ***B: Chromatograph of WaaB T237A gel filtration.*** *Chromatograph is of a Hiload 16/600 Superdex 75 PG column. WaaB elutes as three peaks with 1 representing aggregate protein, 2 leftover MBP tag and 3 as untagged WaaB. X-axis shows elution volume in increments of 50ml, y-axis shows absorbance at 280nm in increments of 50 mAU.* ***C: SDS PAGE of WaaB T273A purification.*** *HisTRAP column wash (W), elution from HisTRAP column (El), desalted protein before (T-) and after (T+) overnight TEV protease digestion at room temperature, HisTRAP-MBPTRAP flow-through containing untagged WaaB (FT), protein retained by HisTRAP-MBPTRAP columns (E2), peak 1 (1), 2 (2) and 3 (3), PageRuler™ 10-180kDa Ladder (L).*

This crystal was produced in the same condition as the original crystals, 50mM HEPES pH6.8, 16% W/V PEG 3350, 1mM Sodium Azide, 1% Typtone, and used the same cryoprotectant (20% v/v glycerol). The crystal itself appeared slightly cracked and only mildly iridescent, suggesting that the protein's purity might not have been sufficient. WaaB elutes from gel filtration as three peaks during this purification with a first peak corresponding to a void volume aggregate, the second to the MalE tag and the final peak to WaaB. The molecular weight of the tag and of WaaB is very similar at 46.6 and 40.9 kDa respectively. To help increase the proportion of pure protein the 200 PG column was exchanged for a 75 PG column to improve the resolution between the final two peaks. An MBP-TRAP column was also attached to the bottom of the HisTRAP column used to remove the tag to ensure a smaller second peak.

A breakthrough moment occurred when further examination of the TEV protease digestion was carried out. Comparison of two equal volumes of protein digested overnight at room temperature and at 4^{o}C found that the level of precipitation in the sample purified at low temperature was visibly higher than that left overnight at room temperature (see Figure 46 A). Further experimentation found that exposure to 4^{o}C for a period of an hour was sufficient to cause significant loss of protein to precipitation. The AKTA system on which the purifications were taking place was maintained at 4^{o}C and the final gel filtration step of the procedure took two hours to complete, meaning an alternative was necessary.

Further purifications took place solely at room temperature and produced stable protein using an AKTA Prime unit (see Figure 46 B and C). WaaB T273A was screened initially against the original crystallisation condition but, curiously, produced no crystals. Screening against the same broad screens as PelA also failed to identify any crystals. Additional screens adding 1mM UDP to the protein droplet produced several conditions that developed crystals. One condition (0.2M sodium citrate tribasic dihydrate, 20% w/v PEG 3350, 15% v/v glycerol as cryoprotectant) produced a crystal of WaaB T273A which diffracted to 2.6 Å resolution, sufficient that the changes to the WaaB binding site should be readily apparent. Model data is shown below in Table 5 as generated by the Phenix software suite.

Table 5 WaaB T273A in complex with UDP data collection and refinement statistics.

Wavelength	
Resolution range	54.16 - 2.61 (2.703 - 2.61)
Space group	C 2 2 21
Unit cell	132.873 177.946 56.0194 90 90 90
Total reflections	274516 (28105)
Unique reflections	20708 (2051)

Multiplicity	13.3 (13.7)
Completeness (%)	99.77 (99.61)
Mean I/sigma(I)	5.40 (1.13)
Wilson B-factor	49.25
R-merge	0.3425 (2.387)
R-meas	0.3562 (2.479)
R-pim	0.09712 (0.6654)
CC1/2	0.991 (0.582)
CC*	0.998 (0.858)
Reflections used in refinement	20665 (2045)
Reflections used for R-free	1996 (199)
R-work	0.1910 (0.3242)
R-free	0.2321 (0.3773)
CC(work)	0.968 (0.772)
CC(free)	0.943 (0.787)
Number of non-hydrogen atoms	2978
macromolecules	2876
ligands	55
solvent	47
Protein residues	360
RMS(bonds)	0.007
RMS(angles)	0.83
Ramachandran favored (%)	96.37
Ramachandran allowed (%)	3.63
Ramachandran outliers (%)	0.00
Rotamer outliers (%)	0.00
Clashscore	6.17
Average B-factor	56.69
macromolecules	56.04
ligands	93.51
solvent	53.53

Statistics for the highest-resolution shell are shown in parentheses.

The resolution of the diffraction is slightly below average, at 2.6 Å, and significantly below the resolution of the wild-type structure at 1.92 Å; it is likely that with more time dedicated towards optimisation that this resolution could be improved. All other statistics are well within ideal parameters, indicating that this is a very strong model from which reliable conclusions can be drawn.

4.5 Crystal structure of WaaB T273A

The initial solving of the structure of WaaB T273A used WT WaaB as a molecular replacement model but produced an initial R-Free greater than 0.5 and diffraction data was in the $C222_1$ space group, in contrast to the $P4_32_12$ space group of the WT crystals. Following significant rebuilding and refinement to an R-Free of 0.23, it was found that the WaaB T273A crystal has adopted a similar but different conformation from that of the wild type protein, which accounts for the difference in space group. This structure contains a monomer within the asymmetric unit with no breaks in the chain and no residues from the tag.

As expected, WaaB T273A in complex with UDP adopts the characteristic GT-B fold of Glycosyltransferase Family 4[119]. This characteristic fold includes two (βαβ) Rossman-like domains with an N-terminal presumed acceptor substrate binding domain and a C-terminal donor-substrate binding domain in complex with UDP (see Figure 47 A). The N-terminal domain comprises Met1-Val161 and contains 7 parallel β-sheets sandwiches arranged in a β3214567 pattern between 6 α-helices, with four, α2-5, on the outer face of the molecule and the remaining two, α1α6, occupying the rear of the molecule adjourning the donor-substrate domain (see Figure 47 B). The 6th α-helix of the N-terminal domain is the C-terminus of the protein, Aasp342-Leu357. The donor-substrate domain comprises Tyr162-Tyr341 and contains 6 parallel β-sheets packed between 6 α-helices. The β-sheets are arranged in a β654123 pattern which partially mirrors the N-terminal domain. The α-helices pack around the β-sheets with α6α4α3 packing against the N-terminal domain and α5α1α2 on the outer face of the molecule. The two domains are connected by a long loop consisting of Val161-Pro180 and the two α6 helices comprised of His330-Leu357.

Using the Dali server[313], the highest quality match to the mutant protein was unsurprisingly the structure of WT WaaB in complex with UDP (PDB: 5N80). The RMSD for this match was only 2.3 over 357 residues which, given that they are the same enzyme, is strong evidence that a different conformation has been adopted as an identical conformer should provide a match of 1.0 or less. Other strong hits included BshA from *Staphylococcus aureus* complexed with UDP and N-acetylglucosamine (PDB: 6N1X)[320], with an RMSD of 3.0 over 335 residues and a sequence identity of 20%, the crystal structure of Anabaena Alr3699/HepE in complex with UDP and glucose (PDB: 4XSU)[321], with an RMSD of 2.7 over 325 residues and a

sequence identity of 21%, and PglH from *Campylobacter jejuni* in complex

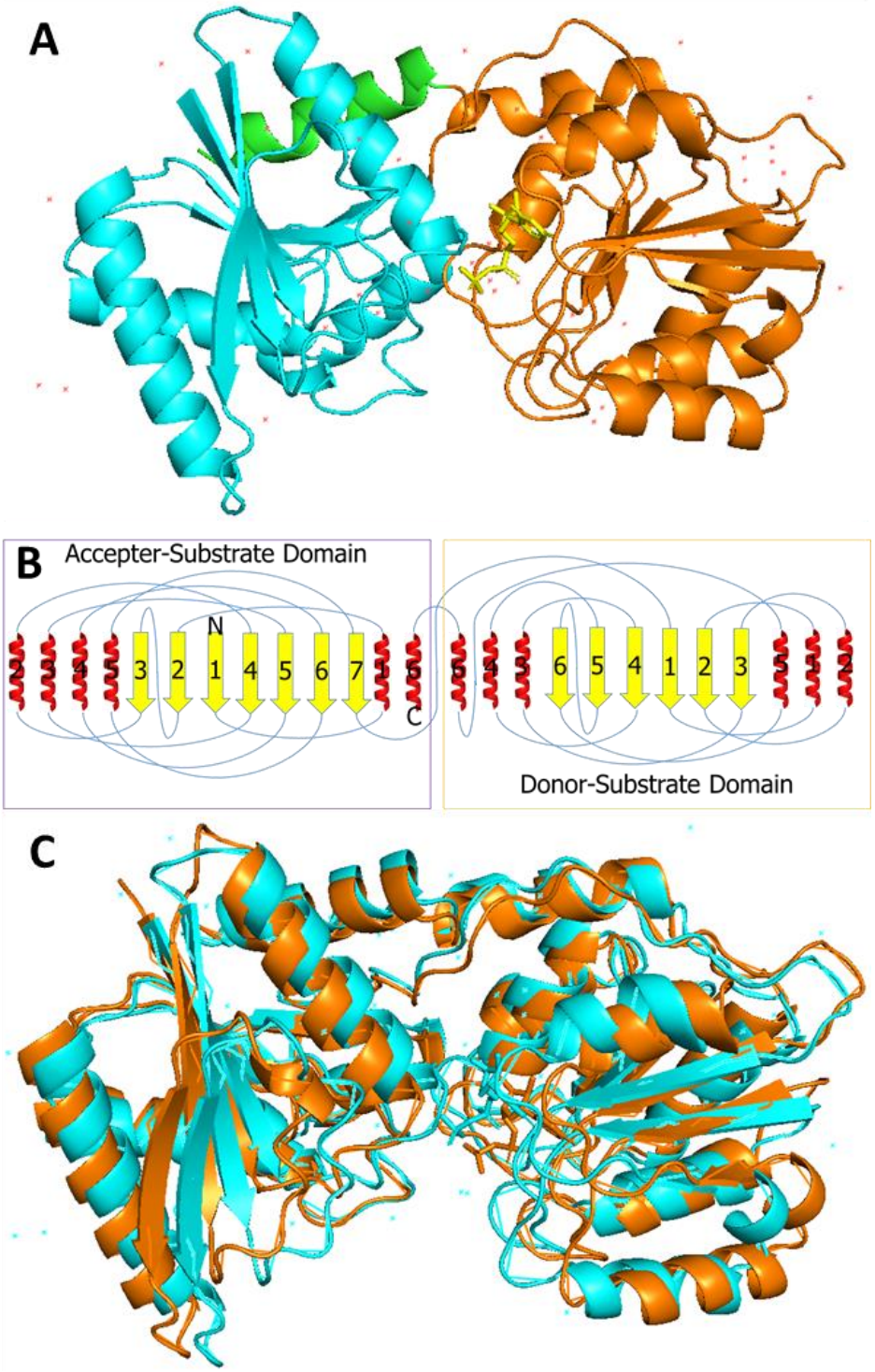

Figure 47 : The structure of WaaB T273A in complex with UDP. ***A: Ribbon diagram of WaaB in complex with UDP.*** *The acceptor-substrate domain (cyan), including the C-terminal α6 helix (green), and the donor-substrate binding domain (orange) adopt the "Rossman-like" GT-B fold typical of glycosyltransferase family 4. UDP (yellow) is bound at the donor-substrate domain.* ***B: Topology diagram of the domains of WaaB.*** *Secondary structural elements are mapped with α-helices in red, β-sheets in yellow and loops in blue.* ***C: Superimposition of the WT and T273A structures of WaaB in complex with UDP.*** *Ribbon diagrams of WT WaaB (orange) superimposed over the T273A structure (cyan). The WT structure adopts an open conformation while the T273A structure adopts a closed conformation.*

with UDP-N-acetylgalactosamine (PDB: 6EJI)[322], with an RMSD of 2.9 over 331 residues and sequence identity of 20%. All these enzyme adopt a GT-B fold from GT family 4[119,248], acting as further confirmation for the designation of WaaB as a GT 4 enzyme.

All three of the listed strong hits had Z-scores of 33.1 or greater, indicating stronger matches than those found for the WT structure, with 32.4 being the highest reported. The structural matches reported for the WT WaaB[304] (PDBs: 4X7M, 4XSO, 3MBO and 5E9U) did not appear as hits for the mutant, although several are related structures that lack donor-substrates. Interestingly, the matches for the mutant enzyme all correspond to structures with the UDP-donor substrate bound to the enzyme and therefore might represent a closed-active conformation of those enzymes, although the same is true for two of the four structures identified for the wild type.

4.6 WaaB T273A adopts an open conformation

Comparison of the WT and T273A mutant structures of WaaB shows that the WT enzyme adopts an open conformation while the mutant has adopted the closed-active conformation (see Figure 47 C). Examination of the site around T273 shows that the threonine found in the WT would sterically hinder the V186 in the closed conformation (see Figure 48). It can therefore be proposed that the increased activity of the T273A mutant is caused by a reduction in the energy barrier needed to adopt the closed conformation required for hydrolysis of the donor-substrate due to increased flexibility of the β1-α1 loop. It would be interesting to test this hypothesis by mutagenesis of V186 to alanine, but that has previously been shown to produce an insoluble product in chapter 4.1.

The regional equivalent of Phe61-Tyr76 has been associated with membrane anchoring in the related enzyme WaaA[323] and is well conserved between the WT and mutant structures. It has been suggested that this area of the protein may cause the conformational shift between open and closed conformations in response to membrane association. Alignment without superposition of this region of the WT and T273A mutant proteins has shown an RMSD of 1.44 Å, while a similar alignment for the entire structures presents an RMSD of 2.49 Å, indicating that the conformational shift at Phe61-76Tyr is substantially below average for the molecule between the open and closed conformations. Measuring the distances within this region shows that the mutant opens by 0.4 Å compared to the wild type (see Figure 49). These results imply that this region does not control conformational shift between the active and inactive states of the enzyme as a larger conformational change would most likely be needed to alter the conformation of the protein in this way.

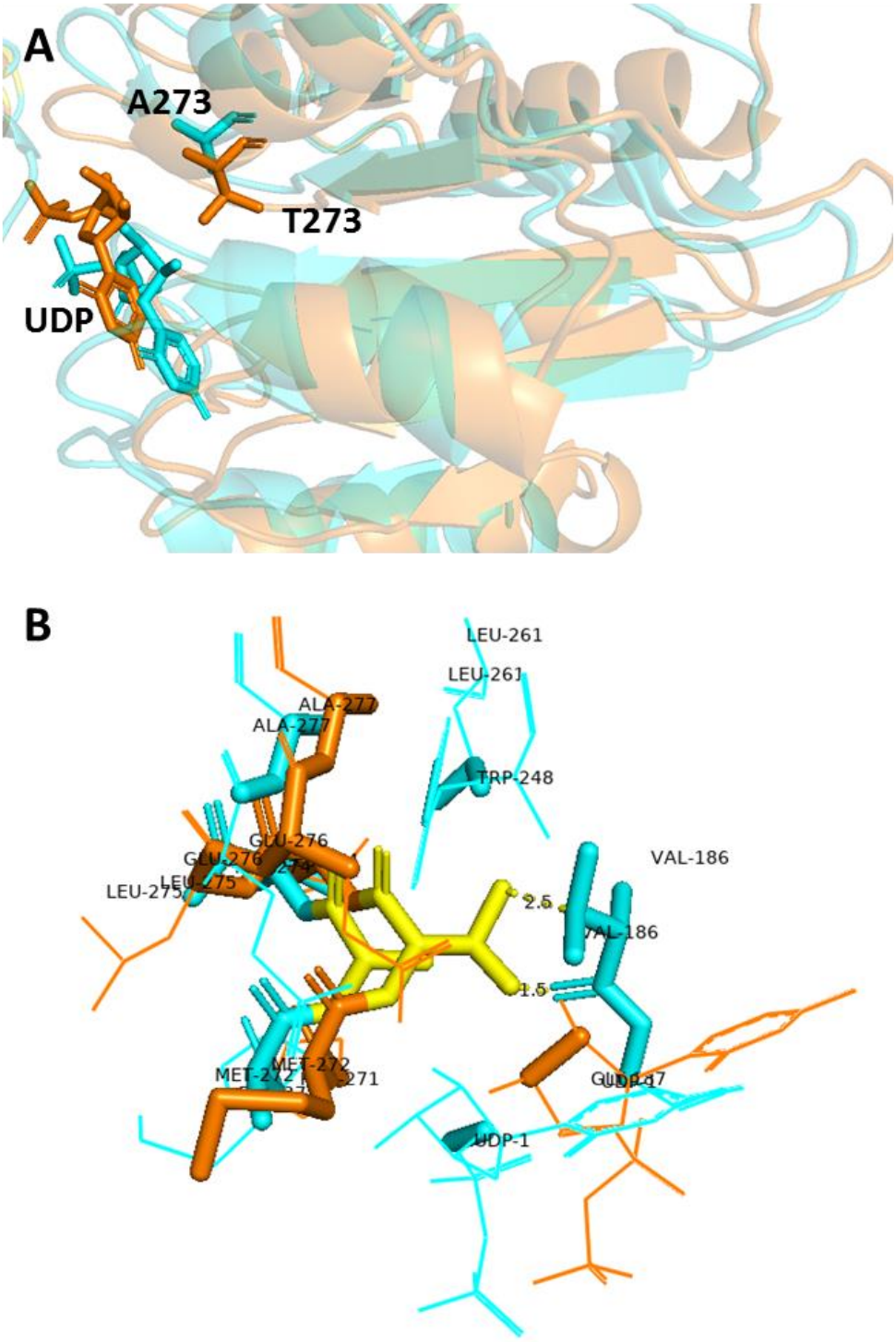

Figure 48 : The mutant site of WaaB T273A. ***A: The UDP binding site around residue T273A.*** *Ribbon diagrams of WT (orange) and T237A (cyan) WaaB in complex with UDP. UDP and residue 273 are shows as sticks.* ***B: Potential close contacts with residue 273.*** *Residue 273 (yellow) and atoms within 4 Å are shown as sticks. The remainder of those residues or molecules is displayed as lines. Distances between the Thr 273 sidechain and the sidechain and carboxyl group of Val 186 are shown in Å. The T273A mutation leaves space allowing a closer conformation of the enzyme more representative of its active state.*

The DynDom Server[324] was used to interrogate the differences between the WT and T273A structures of WaaB. This identified residues His119-Asp123, Pro164-Val165 and Asp432-Ile353 as bending residues that introduce a

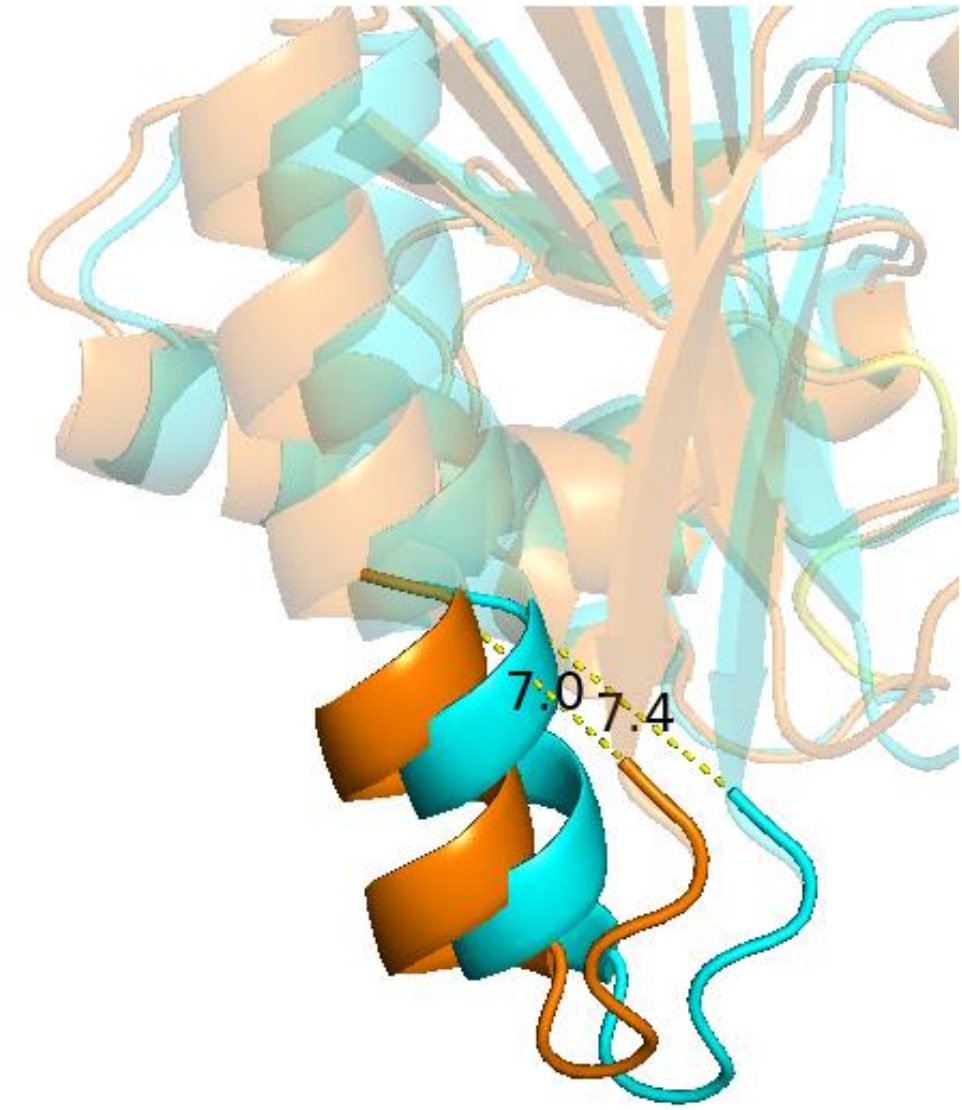

Figure 49 : The putative membrane anchoring region of WaaB. Ribbon diagrams of Phe61-Tyr76 of WT (orange) and T273A (cyan) WaaB in complexes with UDP. The shift in the predicted membrane anchor is small relative to that of the average for the protein and the relative conformation of this region remains unchanged.

rotational angle of 16.8°, well within the 10-25° expected of a change from an open to closed conformation[236,237]. The hinge region is located within the N-terminal domain and consists of segments of the β5-α4 loop, the transdomain β7-β1 loop and a large portion of α6 (see Figure 50). The closed conformation adopted causes an increase in the number of close contacts between residues, especially between the domains. At the base of the molecule, His124 is drawn 0.2 Å closer to Ser121 and 0.5 Å closer to Phe267 of the opposing domain (see Figure 51 A). Met272 is drawn 0.7 Å closer to Val19 and 0.5 Å closer to Gly15. The most dramatic shift is the closing of distance between Phe13 and Trp243 at the centre of the active site cleft (see Figure 51 B) which are drawn 6.8 Å closer together and illustrates the importance of these residues in maintaining this conformation. This identifies an additional role for Trp243 within the

protein and suggests that a mutation of Phe13 might have a similar effect upon catalytic activity.

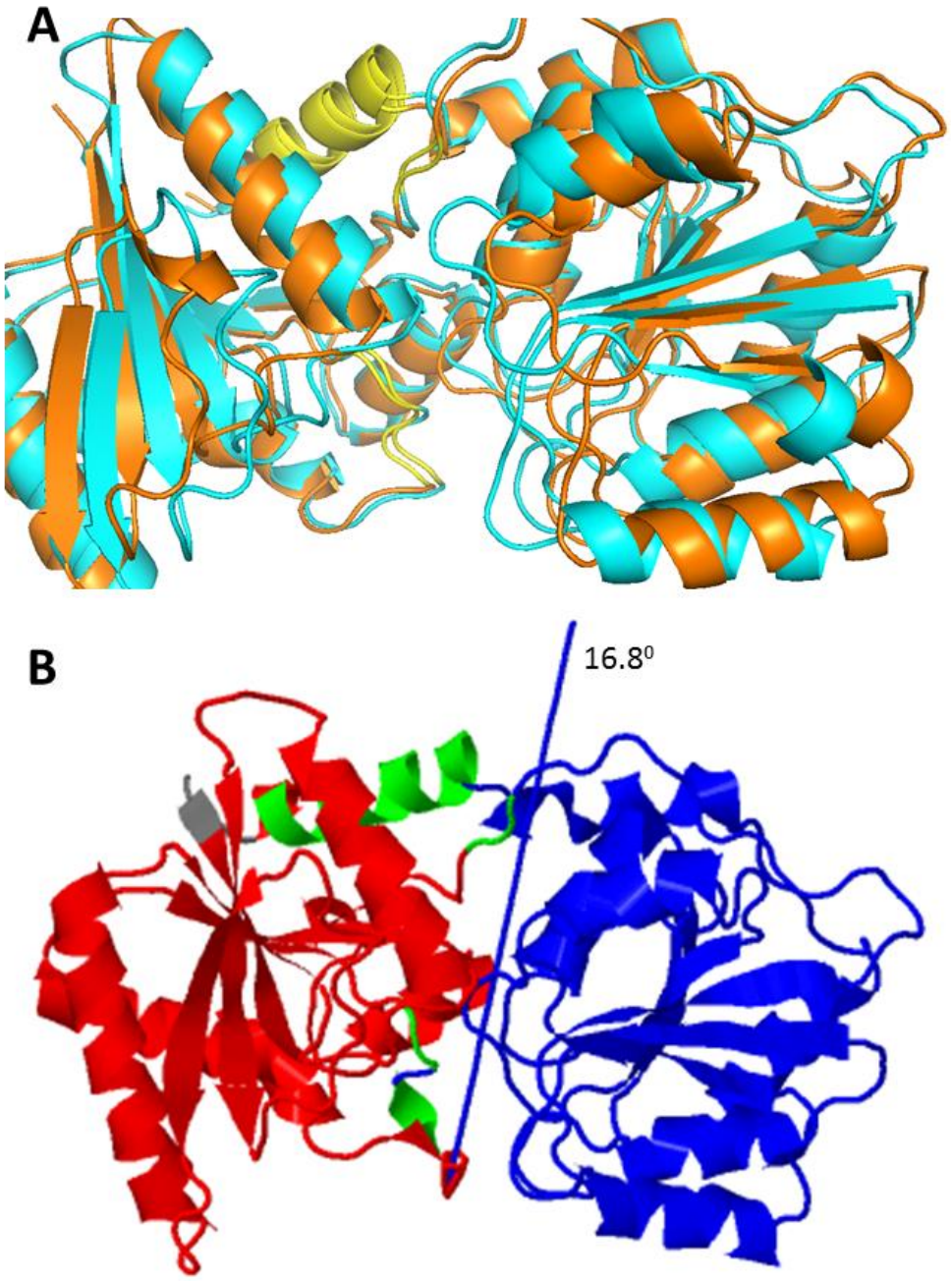

Figure 50 : The hinge region of WaaB. ***A: The hinge region of WaaB in open and closed conformations.*** *Ribbon diagrams of WT (orange) and T273A (cyan) WaaB in complex with UDP in the open and closed conformations of the enzyme, respectively. The hinge region of both structures is displayed in yellow. UDP not shown.* ***B: The hinge region of WaaB in the open conformation showing the axis of rotation.*** *Ribbon diagram of WaaB WT in complex with UDP. N-terminal acceptor-substrate domain (red), C-terminal donor-substrate binding domain (blue), hinge region (green), axis of rotation (blue arrow with red point).*

The positioning of the residues investigated for UDP binding properties and as part of the active site are conserved relative to UDP between the two structures despite the difference in conformation, providing further support

for their importance in UDP binding (see Figure 52) and catalysis (see

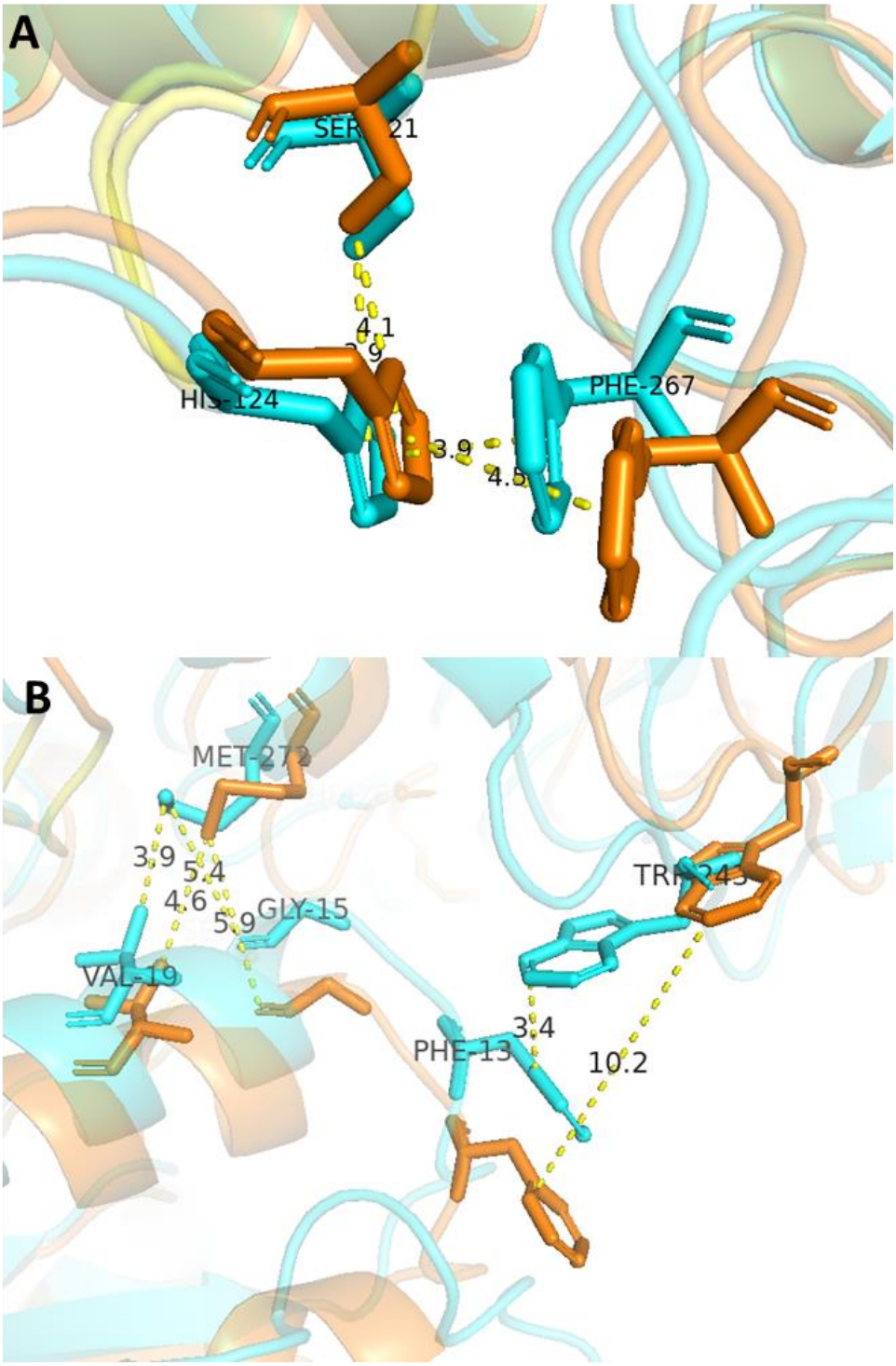

Figure 51 : New contacts formed in the closed conformation of WaaB. ***Ribbon diagrams showing WT (orange) and T273A (cyan) WaaB in complex with UDP. Hinge regions are displayed in yellow. Residues forming new contacts are shown as sticks. Interatomic distances are measured between equivalent atoms in the different models and are displayed in Å. A: Contacts at the base of the active cleft.*** *Contacts are formed between His125 with Ser121 and with Phe267.* ***B: Contacts at the top and centre of the active cleft.*** *Contacts are formed between Phe 13 with Trp243, and between Met272 and Gly 15 and Val 19. The Phe13-Trp243 contact is displays a particularly large shift and may correspond to a molecular clasp.*

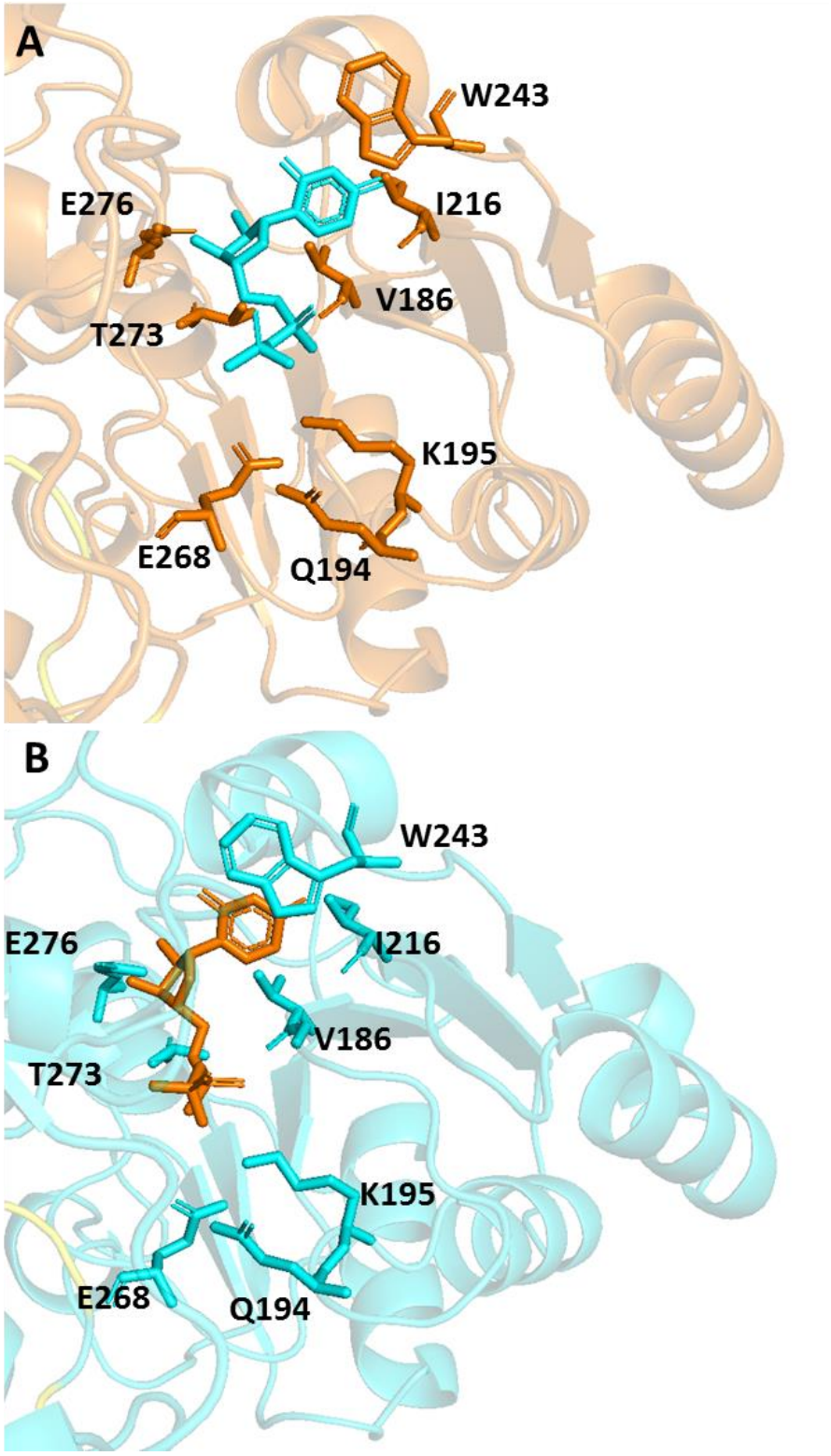

Figure 52 : Positions of mutated residues relative to UDP. ***Cartoon diagrams showing WaaB in complex with UDP taken from the same perspective. Hinge regions are displayed in yellow. Mutated residues and UDP are shown as sticks. UDP is coloured differently for contrast. Mutated residues are V186, Q194, K195, I216, W243, E268, T273 and E276. A: Positions of mutated residues relative to UDP in the open conformation of WaaB.*** *WaaB WT (orange) in complex with UDP (cyan).* ***B: Positions of mutated residues relative to UDP in the closed conformation of WaaB.*** *WaaB T273A (cyan) in complex with UDP (orange). The change in conformation does not alter the secondary structural content but shifts residues in towards the active site. W243 in particular hooks around UDP at the previously open interface.*

Figure 53). Of the residues shown, mutation to alanine was found to significantly decrease activity in all but Val186, which was insoluble, Gln194, which was less active but not significantly so, and in Thr273, in which activity was increased. Gln194 is positioned such that it might have supported the positioning of the catalytic residue Glu268, however, the new structure reveals that in the closed conformation the two sidechains are not sufficiently close to allow this and furthermore reveals a potential role for Ser265 in positioning Glu268 for catalysis.

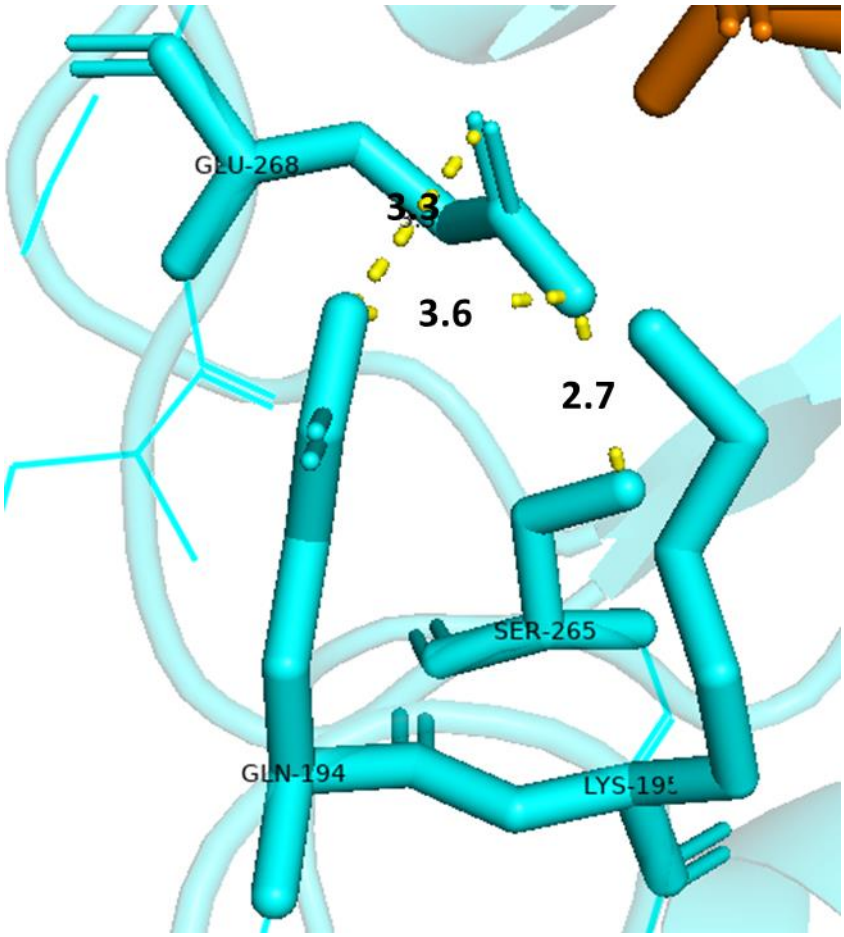

Figure 53 : Position of the catalytic site in the closed conformation. ***Cartoon diagram of WaaB T273A*** *(cyan) in complex with UDP (orange). Catalytic site residues Lys195 and Glu268 are shown with potential hydrogen bonding pairs within 4 Å. Interatomic distances are shown in Å. Gln194 was predicted to be either catalytic or support catalysis but has been shown to not be needed for catalysis and is not close enough (3.3-3.6 Å) to hydrogen bond with Glu268. The orientation of Glu268 is instead stabilised by hydrogen bonding with Ser265. This area provides an abundance of charged residues, providing support for this region being the catalytic microenvironment for a same-side S_ni reaction mechanism.*

Chapter 5 – GtfC

5.1 Crystallization of $GtfC_{100\text{-}23}$

Both $GtfC_{100\text{-}23}$ and $GtfC_{ATCC\ 53608}$ were supplied by our collaborators and crystallisation was trialled against a diverse range of crystallization screens at 16 and 4 °C. $GtfC_{100\text{-}23}$ produced an abundance of plate-shaped crystals (see Figure 54) in numerous conditions. A second round of screening in which 1mM UDP was added to the crystallisation droplet identified an even more diverse selection of crystals, including rods and square bipyramidal crystals (see Figure 54). The originally identified optimal condition for growing plates without UDP also produced plates with UDP. Through both rounds of screening, no crystals were found to grow at 4 °C or for $GtfC_{ATCC\ 53608}$. SDS PAGE analysis of the two provided enzymes identified a contaminant in the $GtfC_{ATCC\ 53608}$ sample (see Figure 55) which may account for the lack of success in crystallising that enzyme. Our collaborators found that this contaminant was persistent through metal affinity, anion exchange and size exclusion chromatography (data not shown).

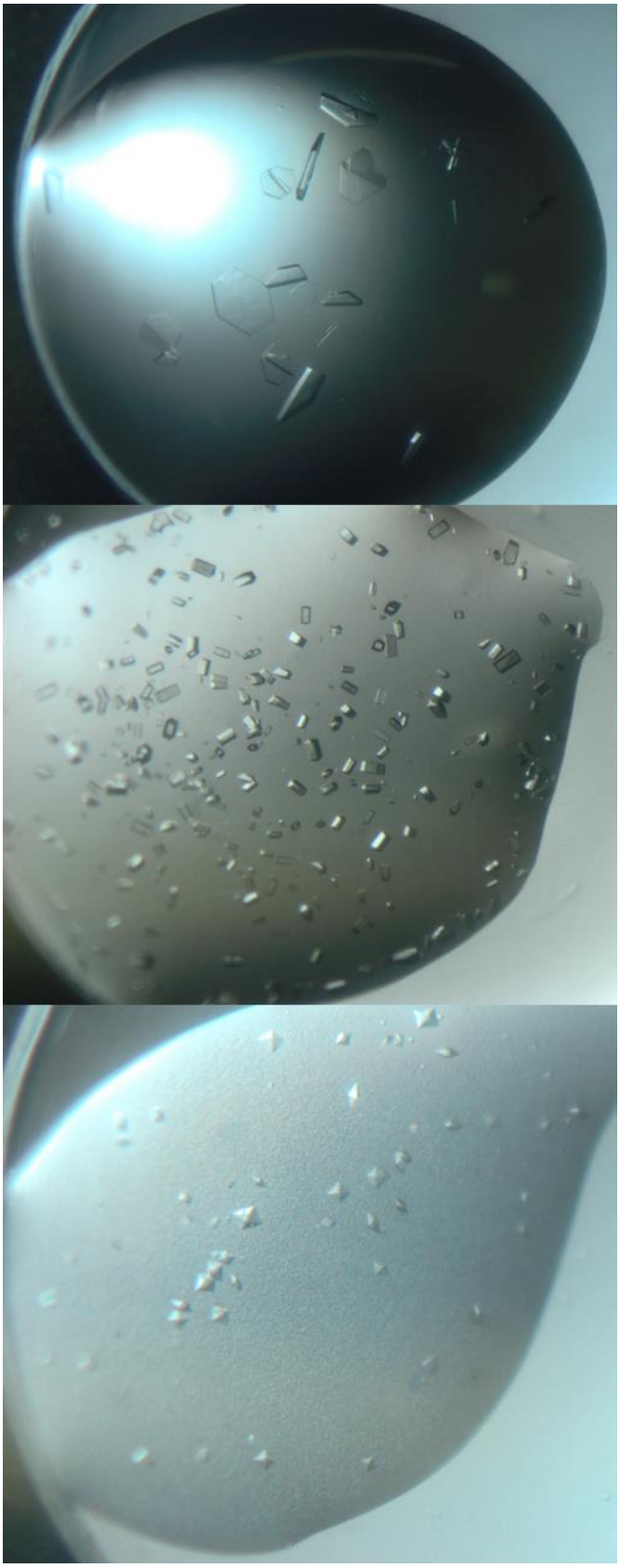

Figure 54: Droplets began at 1µl in size with a 1:1 ratio between well solution and 10mg/ml protein. ***A: Plates.*** *0.2M Potassium Thiocyanate, 0.1M Bis-Tris Propane pH 7.5, 20% w/v Polyethylene Glycol 3350 after two days.* ***B: Rods.*** *0.1M Sodium Acetate pH 4.9, 0.02M Calcium chloride 33% V/V Hexylene Glycol, 1mM UDP after 7 days.* ***C: Bipyramidal.*** *0.15M Potassium Bromide, 33% W/V Polyethylene Glycol Mono-methyl Ether, pH 6.5 after 7 days.*

More than 100 crystals of $GtfC_{100\text{-}23}$ were sent to Diamond Light Source for diffraction experiments. Of these, it was found that the plates originally identified were able to diffract to a resolution of 2.8 Å while retaining a completeness of >95%, additionally, while these crystals were visibly fully formed after two days a longer incubation time of 2-4 weeks produced crystals able to diffract to higher resolutions. The rod-shaped crystals were able to diffract to a resolution of 2.4 Å but produced twinned data sets that were not amenable to processing. The bipyramidal crystals did not diffract. Interestingly, addition of UDP to the plate-producing crystal generated plates that were visually indistinguishable but that when exposed and processed were found to produce a dataset at 2.5 Å and in the $P2_1$ space group rather than the previously observed P 2_1 2_1 2_1 space group of the UDP-free crystals. We selected this dataset to solve and refine the crystal structure of the complex of $GtfC_{100\text{-}23}$ with UDP. Model data is shown below in Tables 6.1 and 6.2, as generated by the Phenix software suite.

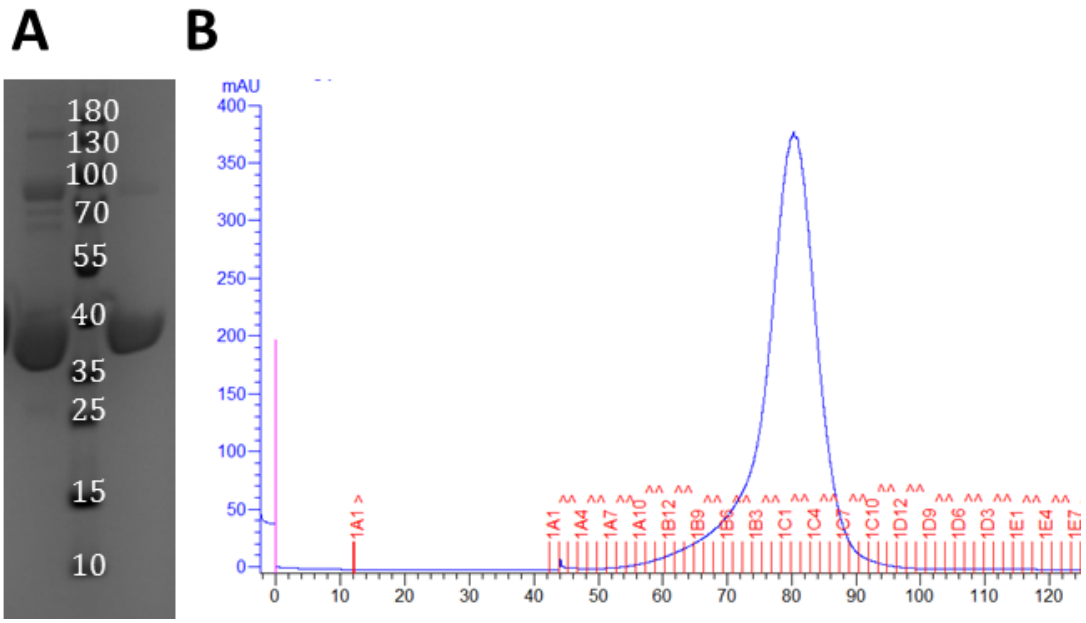

Figure 55: Sample quality of GtfC preparations. ***A: SDS PAGE of GtfC samples.*** *$GtfC_{ATCC\ 53608}$ (MW 38kDa), PageRuler™ 10-180kDa Ladder (L), $GtfC_{100\text{-}23}$ (38kDa). The $GtfC_{ATCC\ 53608}$ sample shows at least four contaminant bands.* ***B: Size exclusion peak of $GtfC_{100\text{-}23}$.*** *Chromatograph trace showing UV absorbance at 280nm as a function of elution time. GtfC used was defrosted from -80°C and run to remove damaged or aggregate protein before further crystallisation trials. The tail end of the peak appears more homogenous so was used for further work.*

Table 5.1 GtfC 100-23 apo 1:3 holo complex with UDP data collection and refinement statistics.

Wavelength	
Resolution range	50.98 - 2.508 (2.598 - 2.508)

Space group	P 1 21 1
Unit cell	72.406 70.6867 140.312 90 91.3742 90
Total reflections	339997 (31348)
Unique reflections	48851 (4837)
Multiplicity	7.0 (6.5)
Completeness (%)	99.37 (98.08)
Mean I/sigma(I)	4.72 (0.86)
Wilson B-factor	47.90
R-merge	0.2221 (1.604)
R-meas	0.2406 (1.746)
R-pim	0.09156 (0.6819)
CC1/2	0.978 (0.819)
CC*	0.994 (0.949)
Reflections used in refinement	48603 (4794)
Reflections used for R-free	2410 (255)
R-work	0.2179 (0.4089)
R-free	0.2732 (0.4395)
CC(work)	0.936 (0.747)
CC(free)	0.933 (0.619)
Number of non-hydrogen atoms	11391
macromolecules	10627
ligands	75
solvent	689
Protein residues	1322
RMS(bonds)	0.010
RMS(angles)	1.17
Ramachandran favored (%)	93.48
Ramachandran allowed (%)	4.98

Ramachandran outliers (%)	1.53
Rotamer outliers (%)	0.00
Clashscore	16.93
Average B-factor	68.33
macromolecules	68.64
ligands	74.20
solvent	63.00
Number of TLS groups	1

Statistics for the highest-resolution shell are shown in parentheses.

The lower resolution UDP-free $GtfC_{100-23}$ complex also underwent refinement before being used as a molecular replacement model for the above structure. Given the lower resolution, this model was not deemed of enough import to refine further, however, as it will be discussed later the statistics for this model are also provided:

Table 5.2 GtfC 100-23 apo structure data collection and refinement statistics.

Wavelength	
Resolution range	100.3 - 2.816 (2.917 - 2.816)
Space group	P 21 21 21
Unit cell	70.1772 140.759 142.96 90 90 90
Total reflections	448319 (43996)
Unique reflections	35019 (3417)
Multiplicity	12.8 (12.9)
Completeness (%)	99.59 (98.81)
Mean I/sigma(I)	5.50 (1.15)
Wilson B-factor	62.61
R-merge	0.423 (4.954)
R-meas	0.4408 (5.16)
R-pim	0.123 (1.432)
CC1/2	0.99 (0.613)
CC*	0.998 (0.872)

Reflections used in refinement	34899 (3400)
Reflections used for R-free	1658 (167)
R-work	0.3062 (0.4034)
R-free	0.3702 (0.4350)
CC(work)	0.898 (0.793)
CC(free)	0.877 (0.739)
Number of non-hydrogen atoms	10568
macromolecules	10568
Protein residues	1308
RMS(bonds)	0.017
RMS(angles)	2.39
Ramachandran favored (%)	85.84
Ramachandran allowed (%)	13.08
Ramachandran outliers (%)	1.08
Rotamer outliers (%)	0.35
Clashscore	37.89
Average B-factor	76.08
macromolecules	76.08

Statistics for the highest-resolution shell are shown in parentheses.

5.2 The fold of $GtfC_{100\text{-}23}$

Of the four molecules of $GtfC_{100\text{-}23}$ within the asymmetric unit only three bind UDP (Chains A, B and C). The electron density of chains A and B are of higher quality than chains C and D, most likely a result of stabilisation caused by crystal contacts with the content of adjacent asymmetric units. For the purposes of this description, only chains A and D, representing the UDP-bound and unbounds forms, respectively, will be referred to. $GtfC_{100\text{-}23}$ adopts a canonical GT-B fold characteristic of the Glycosyltransferase Family 2[119]. This fold includes two Rossman-like domains with an N-terminal acceptor substrate binding domain and a C-terminal donor-substrate binding domain (see Figure 56 A). Where present, UDP is in

complex with the C-terminal domain, confirming that this acts as the

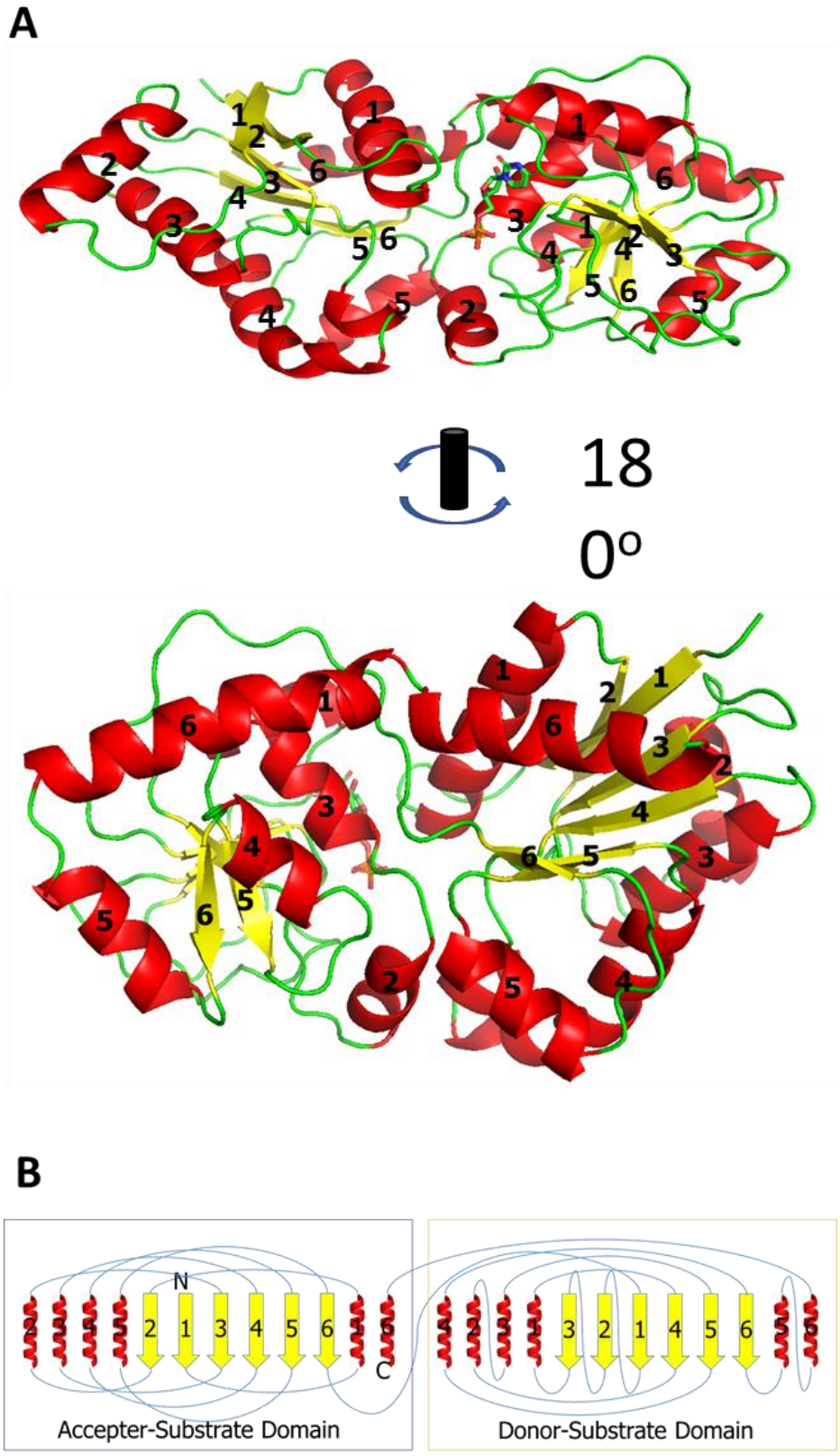

Figure 56: Structural Arrangement of GtfC100-23 ***A: Panoramic view of GtfC$_{100-23}$.*** *Ribbon diagram showing one monomer of GtfC coloured by secondary structure (α-helices in red, β-sheets in yellow, loops in green). Secondary structural elements are labelled. UDP is shown as sticks. The second structure represents a 180^0 rotation around the y-axis. B:* ***Topology diagram of GtfC$_{100-23}$.*** *Secondary structural elements are mapped with α-helices in red, β-sheets in yellow and loops in blue.*

donor-sugar binding site. The active site of the enzyme exists in the cleft between the two domains.

The N-terminal domain consists of six parallel β-sheets sandwiched between four heavily solvent exposed α-helices (α2-5) and two additional solvent-exposed α-helices at the interface between domains. The domain is composed of Leu1-Gln151 in a βα arrangement and the 6th α-helix provided by the C-terminal region of Gly318-Tyr334 (see Figure 56 B). The β-sheets are arranged in a β213456 sequence which shows some divergence from the classical β321456 pattern[128]. The N and C-terminal domains connect through a long solvent-exposed loop composed of Leu150-Val171 and through the α6 helix.

The C-terminal domain consists of six parallel β-strands adopting a canonical β321456 arrangement. The β1456 sheets are sandwiched between six α-helices, of which five are positioned around the interface between domains while α5 takes up a solvent exposed position at the outer face of the molecule. The remaining β32 are solvent exposed on one side while the other is protected by several loops. UDP binds at the interface of β1, β2, α1 and α3.

5.3 Identifying binding sites through homology analysis

The DALI server[313] results for the structure of $GtfC_{100\text{-}23}$ identifies two extremely close matches, the molecular replacement search model (Gtf3, the glycosyltransferase 3 from *Streptococcus parasanguinis,* PDB: 3QKW) and GtfC from *Streptococcus agalactiae* (PDB: 4W6Q)[325]. Both matches have Z-scores of 45.2 with RMSD of 1.2 Å over 329 Cα and sequence identities or 42 and 40% respectively (see Figure 57 A). The next highest hit is the structure of the N-acetlyglucosamyltransferase MshA from *Corynebacterium glutamicum* (PDB: 3C4V) with Z-score of 22.0, RMSD of 3.3 Å over 305 Cα and sequence identity of 13%, marking a significant drop in relatedness.

Superposition of $GtfC_{100\text{-}23}$ with Gtf3 (PDB: 3QKW)[274], identifies residues K180, W214 and K251 to be equivalent to R179, Y211 and K246 within the UDP binding site of Gtf3 (see Figure 57 B). These lysine residues are proposed to mediate the negative charge of the phosphate and their mutation produces inactive enzyme[274]. W214 stacks above the uracil ring of UDP and the equivalent Y211 was found to be necessary for efficient substrate binding. In addition to these, examination of our structure suggests that UDP binding may be mediated by L220 at the uridine ring and T16 at the phosphate. The authors also identified a region within the α6 of the acceptor-substrate domain of the protein that equates to residues Y319, F320, K323 and K325 which was involved in protein oligomerisation (see Figure 57 C), mutations of which reduced enzymatic activity[274]. Potential individually identified catalytic residues include E101, Y243, N249 and H250 based upon the likely position of the UDP sugar (see Figure 58 A).

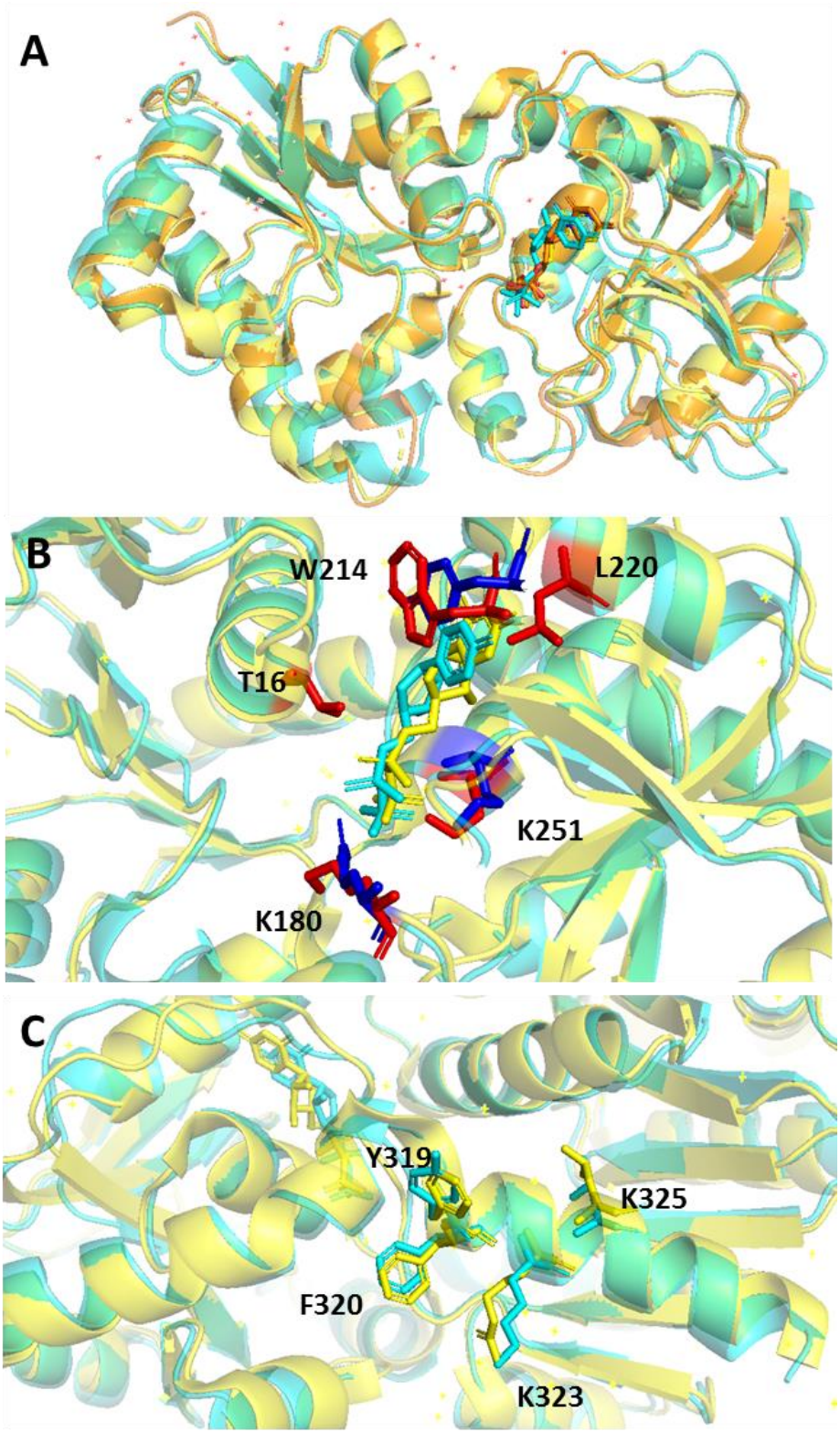

Figure 57: Homology Analysis of GtfC100-23 ***A: Superposition of Gtf3 and GtfC$_{S.\ agalactiae}$ with GtfC$_{100-23}$.*** *Ribbon diagram showing one monomer of GtfC$_{100-23}$ (cyan) superimposed with Gtf3 (Yellow, PDB: 3QKW) and GtfC$_{S.\ agalactiae}$ (orange, PDB: 4W6Q). UDP is shown as sticks.* ***B: UDP binding residues from GtfC$_{100-23}$.*** *Ribbon diagram of GtfC$_{100-23}$ (cyan) superimposed with Gtf3 (Yellow, PDB: 3QKW). The positions of the three UDP binding residues K180, W214 and K251 (red) and their equivalent residues in Gtf3 are shown (blue). Putative UDP binding residues are also displayed (red, no equivalent residue)* ***C: Oligomerisation-related residues of GtfC$_{100-23}$.*** *Ribbon diagram of GtfC$_{100-23}$ (cyan) superimposed with Gtf3 (Yellow, PDB: 3QKW). UDP and oligomerisation residues shown as sticks.*

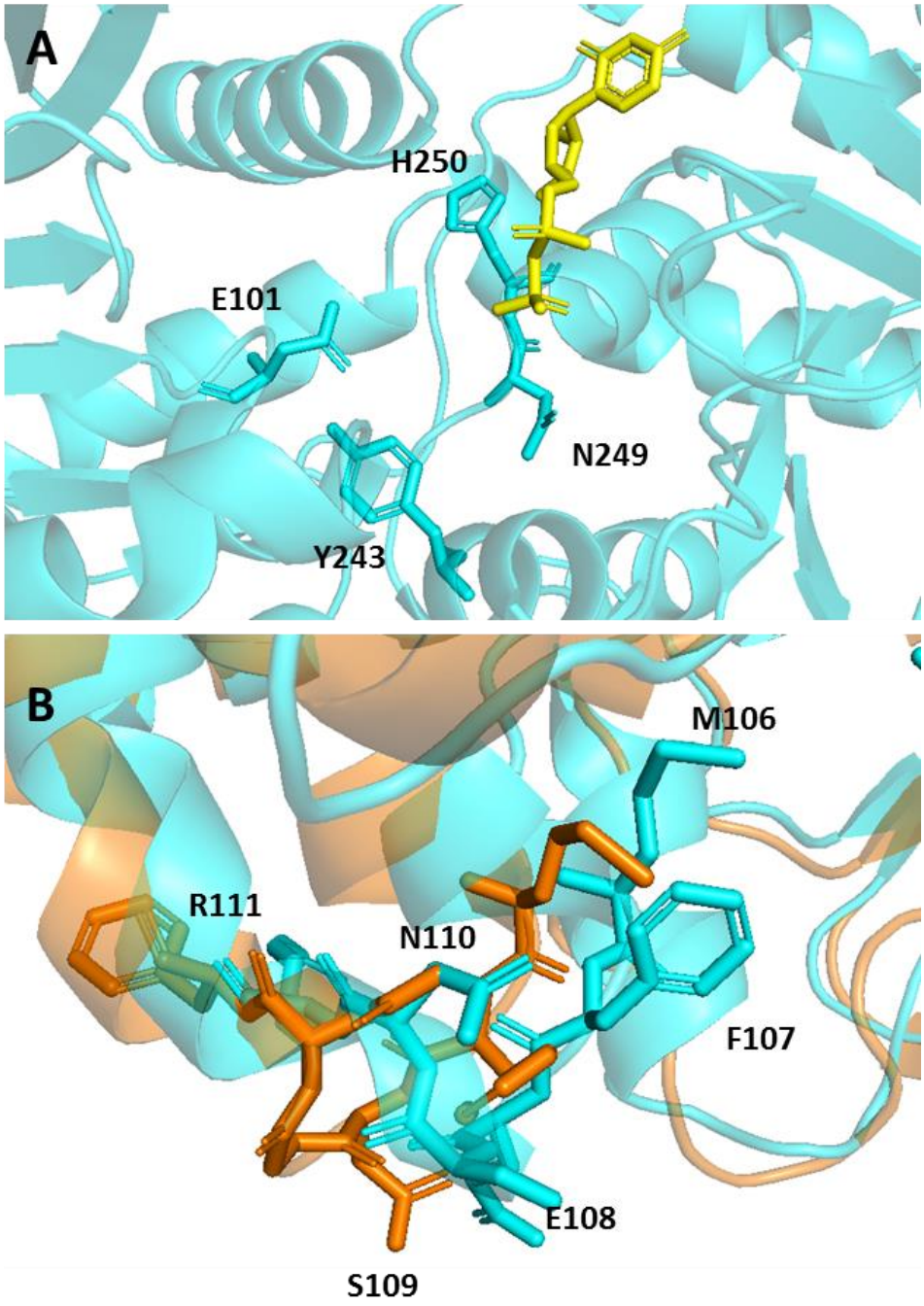

Figure 58: Further Homology Analysis of GtfC$_{100\text{-}23}$. ***A: Potential catalytic residues of GtfC$_{100\text{-}23}$.*** *Ribbon diagram showing GtfC (cyan), UDP (yellow, sticks) and four potential catalytic residues (cyan, sticks).* ***B: Superposition of the acceptor-substrate binding loops of GtfC$_{S.\ agalactiae}$ with GtfC$_{100\text{-}23}$.*** *Ribbon diagram showing one monomer of GtfC$_{100\text{-}23}$ (cyan) superimposed with GtfC$_{S.\ agalactiae}$ (orange, PDB: 4W6Q). Residues of the acceptor-substrate binding loop are shown (sticks).*

GtfC$_{S.\ agalactiae}$ has been found to contain an important loop in positions 106-111 that has been identified as necessary for enzymatic activity and is a proposed acceptor-substrate binding site; mutation of the same loop in Gtf3 was found to prevent activity as well[325]. This site is similar within the 100-23 structure (see Figure 58 B), with MFESNR replacing MFDGNF at the same residue numbers and position, however, in 100-23 this region maintains the helical structure of N-α4 and is drawn tighter into the active

cleft, suggesting that it may be less flexible. This is further supported by the partial disorder of this region within the apo structure of GtfC$_{100-23}$, as this site was completely disordered in GtfC$_{\textit{S. agalactiae}}$.

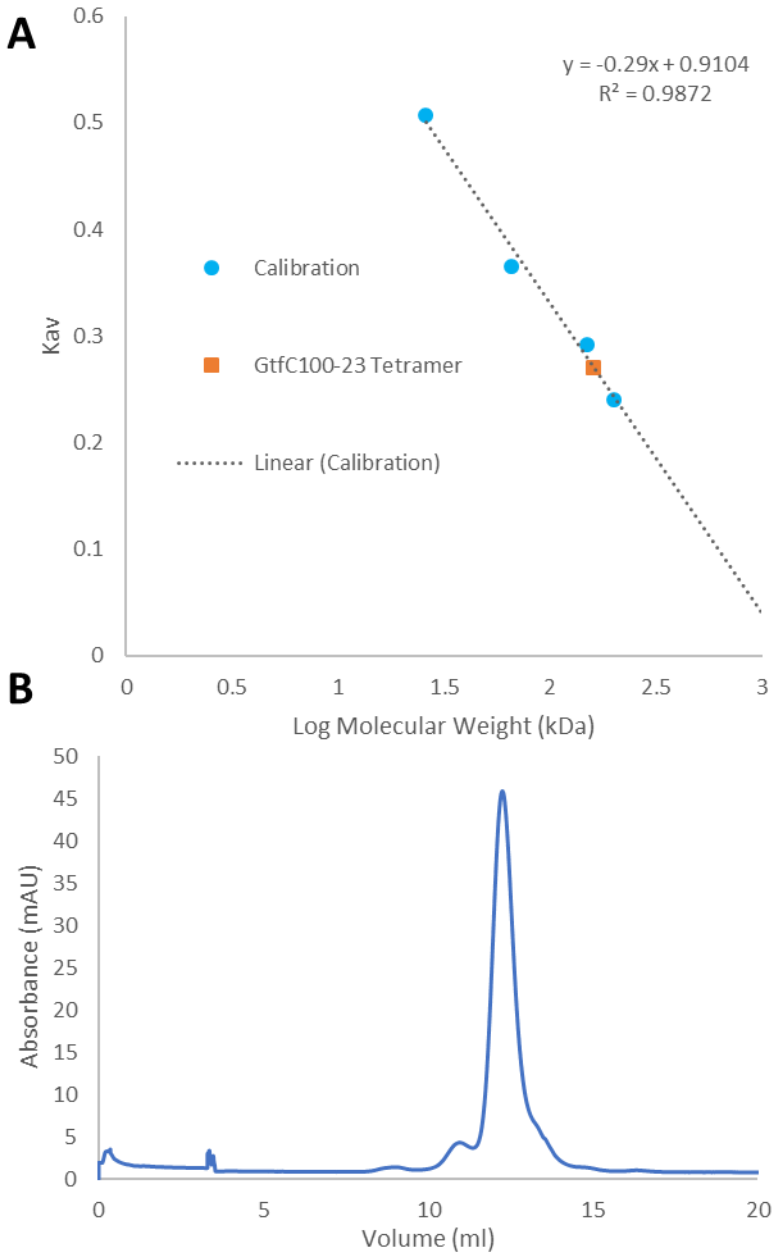

Figure 59: Determination of the biological unit of GtfC$_{100-23}$. ***A: Calibration of Superdex 200 Increase 10/30 GL column and fit of proposed oligomerisation states to calibration.*** *X-axis shows the log$_{10}$ of the molecular weight of the standards (calibration) and the proposed complex of GtfC. The y-axis shows Kav, as calculated in chapter 5.2.3 and represents the point at which a peak eluted as a fraction of the volume of the column.* ***B: Gel filtration of GtfC$_{100-23}$.*** *Trace shows absorbance at 280nm as a function of elution volume. The protein elutes at 12.24 ml which corresponds to 4.54ml after the void volume.*

5.4 The biological unit of GtfC$_{100-23}$ is a tetramer

The asymmetric unit of GtfC$_{100-23}$ contains four molecules of the protein. As a region of interest has been identified that may be related to oligomerisation of the protein and efficient catalysis, the α6 of the

acceptor-substrate domain, it was decided to determine the oligomerisation state of the protein using analytical size exclusion chromatography. $GtfC_{100-23}$ eluted 4.54 mL after the void volume which places the molecular weight of the complex at around 160 kDa, equivalent to a tetramer (see Figure 59). Analysis of the structure using ePISA also identified the tetramer as the biological unit[326].

5.5 Identifying specificity determining residues of $GtfC_{100-23}$

The amino acid sequence identity of $GtfC_{100-23}$ and $GtfC_{ATCC\ 53608}$ is 97%, which corresponds to a difference of 10 residues (see Figure 60 A). Of these residues, three are found within the SRRP-binding domain (M1, S40 and N331) and are thus unlikely to determine the specificity of the donor-binding domain. Of the remaining seven residues, P166, N191 and I203 are positioned on the outer face of the protein and so are unlikely to be directly involved, although the change of a proline to leucine is likely to change the positioning of the C-β1 strand which contributes to UDP binding through K180. Y157 is positioned towards the top of the active site but is facing away from it; while it may contribute to efficient substrate diffusion it is unlikely to be involved in binding.

The remaining three residues are likely to be donor-substrate determinants. W240 is aligned such that it would likely hinder the binding of a sugar substrate larger than UDP-glucose, such as UDP-N-acetylglucosamine. In the homologous enzyme, this residue is the much shorter cysteine, giving credence to this interpretation. The two other residues, L174 and S238, are potentially important as the corresponding F174 is significantly larger and therefore likely to distort the C-β1-β2 loop further into the active groove while P238 would introduce a sharper turn to helix C-α2 that would adjust the positioning of C240.

5.6 UDP binding promotes a domain shift in $GtfC_{100-23}$

Due to the poor quality of the electron density of the apo enzyme within the apo 1:3 holo, the UDP-free structure was used for comparison with the holo enzyme. Superposition of these structures reveals that UDP binding does not alter the quaternary structure of the tetramer (see Figure 61 A), which consists of two pairs of back-to-back dimers with outward-facing active sites. Closer inspection of the monomers reveals no change in the N-terminal acceptor-binding domain (RMSD 0.00 over 150 residues) but a substantial shift in the C-terminal donor-binding domain (RMSD 0.762 over 276 residues) in response to UDP binding. This includes a prominent shortening of β3-2-1 (see Figure 61 B) and mild lengthening of α2. The largest change can be seen in the β2-3 loop which shifts up to 4.6 Å closer to the α1-β2 loop (see Figure 61 C). The overall conformational shift brings the two domains closer upon UDP binding, representing an incomplete step from the open toward the closed conformation of the enzyme.

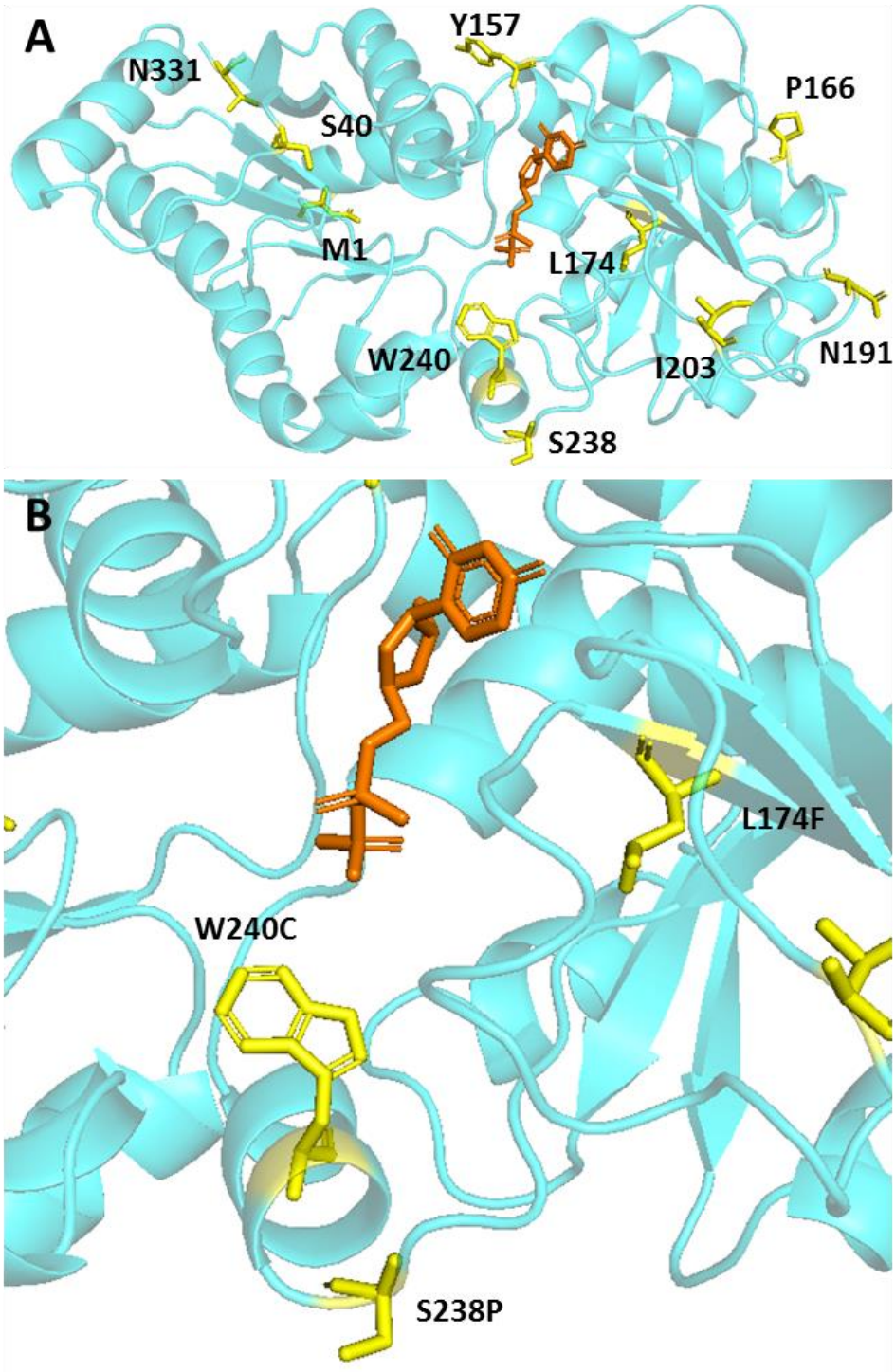

Figure 60: Homology Analysis of GtfC$_{100\text{-}23}$ against the sequence of GtfC$_{ATCC\ 53608}$. ***A: The structure of GtfC$_{100\text{-}23}$ highlighting unconserved residues with GtfC$_{ATCC\ 53608}$.*** *Ribbon diagram showing GtfC (cyan), UDP (orange, sticks) and non-conserved residues (yellow, sticks).* ***B: Putative donor-substrate determinants of GtfC$_{100\text{-}23}$.*** *As A.*

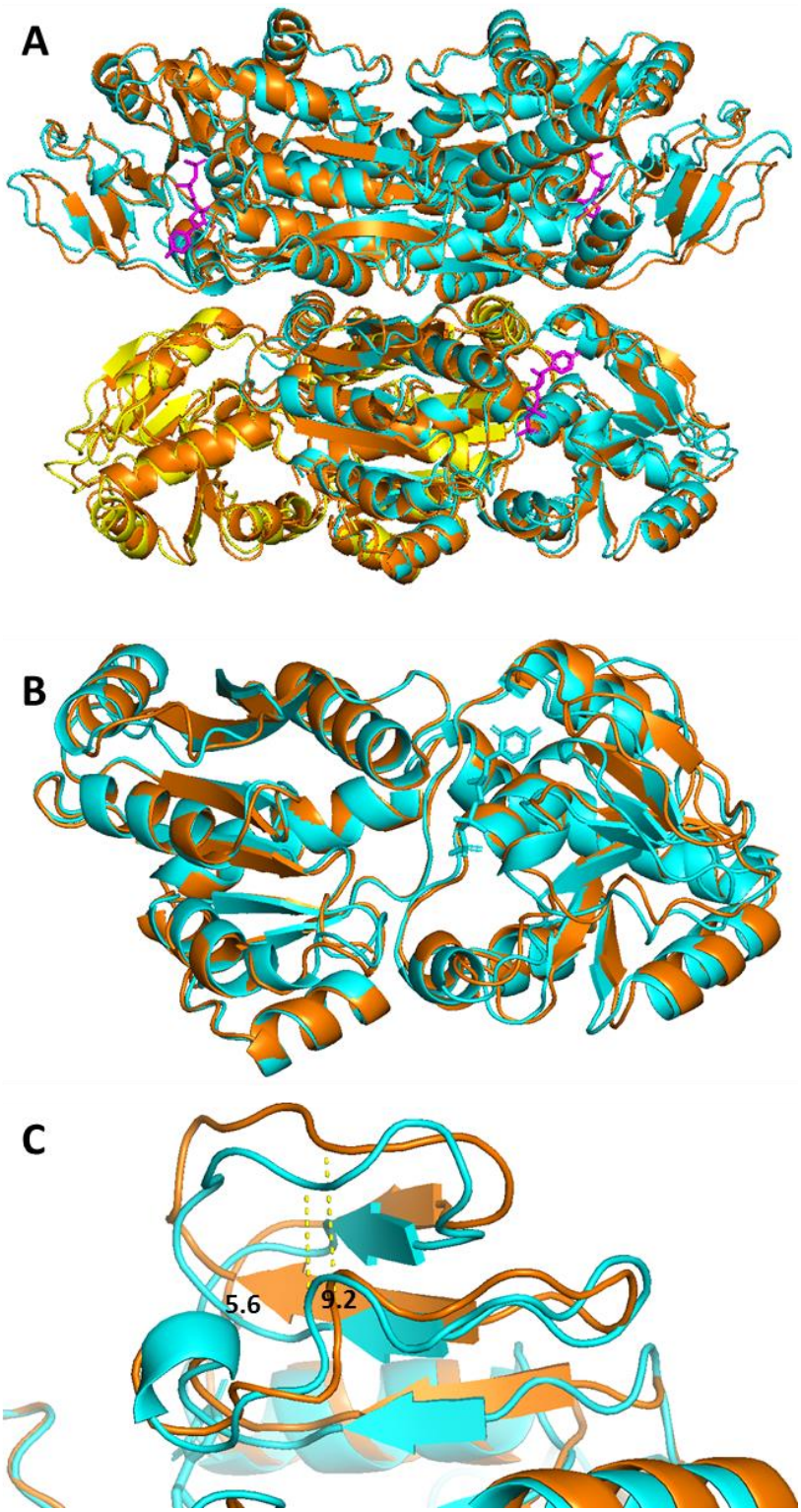

Figure 61: Superposition of apo GtfC$_{100-23}$ against apo 1:3 holo GtfC$_{100-23}$. ***A: Biological Assemblies of GtfC$_{100-23}$ and apo 1:3 holo GtfC$_{100-23}$.*** *Ribbon diagram showing fully apo GtfC (orange), holo GtfC (cyan), apo from mixed species GtfC (yellow) and UDP (magenta, sticks). The quaternary structure remains unchanged between structures, with four outward-facing active sites* ***B: Monomers of apo and holo GtfC.*** *Ribbon diagram showing apo (orange) and holo (cyan) monomers of GtfC$_{100-23}$ alongside UDP (sticks). Superposition was of the N-terminal 150 amino acids of the acceptor-substrate domain, which is mostly unchanged. The c-terminal donor-substrate domain shows a RMSD shift of 0.762 Å.* ***C: The α1-β2-β2-3 loops.*** *See B for key. Distances shown between V184 and P205 in Å.*

Chapter 6 - Discussion

The original stated aims of the Pel project were to clone, express and purify the seven genes/proteins of the *pel* operon from *P. aeruginosa* for crystallisation studies to determine functional assignments of those proteins that would improve our understanding of the biology of biofilm formation and provide detailed molecular models of these proteins for the purposes of drug discovery. At the start of this project, the structure of the cytoplasmic domain of PelD had already been solved[112,113] but structures of the remaining proteins, and of full-length PelD, had not been determined. These proteins were successfully cloned, or synthesised, for expression in *E.coli* but only PelA was found to express as a soluble protein, that was subsequently crystallised, the structure determined and characterised using aSEC.

The structure of PelA contains two domains, an N-terminal glycoside hydrolase domain of unusual $\alpha_7\beta_8$ configuration preceding a glutamine amido-transferase-like domain containing a retarded active site. The structure of this first domain has previously been published[327] shortly before the completion of this thesis. This publication identifies the N1 domain as an endo retaining α-1,4-N-acetylgalactosaminidase, in line with previous expectations[119]. Their model adopts a near-identical conformation to our own protein with and RMSD of 0.9 over 227 residues. Notably, in their structure the 104-128 gap that we observe in our structure forms a small βααβ loop (loop3) at the open face of the enzyme that supporting their conclusion that this represents a flexible region of the enzyme. Our structure equally reveals that the 297-310 region of the protein does not form the predicted supporting β-strand. The putative active site residue Glu-218 has also been shown to be catalytically important[327], although the roles of Asp160 and Asn199 remain to be investigated. This could be achieved relatively easily through a combination of alanine mutagenesis and the now well established biofilm disruption assay[71,150,327,328].

The structure of the second domain of PelA remains novel, as does its role in oligomerisation of the protein; a trait which, in itself, appears not to have previously been identified for PelA. This represents the first strong evidence that retarded GATase domains have a functional role beyond a generic increase of protein stability, which will serve to inform future functional assignment of these domains. Despite this discovery, the functional role of the GATase domain of PelA would benefit from further study as further roles in regulation and/or substrate binding remain both feasible and unexplored.

One possibility is that the N2 domain holds a regulatory function that is dependent upon interactions with another pel proteins. It has previously been found that PelB is able to inhibit the activity of the N1 domain in PelA 46-948 but is unable to do so for the N1 domain alone[150]. This suggests that PelB must interact with a region of the protein C-terminal to the N1 domain leading to inhibition. That a higher order oligomer of the N1

domain is necessary for such a direct interaction can be discounted as we have confirmed that the N1 domain in isolation is monomeric.

This hypothesis would be relatively easy to test as both PelA 46-507 and PelB 47-880 have been found (this study) and reported[150] to be soluble, respectively. This means that whether an interaction takes place could be determined by analytical size exclusion chromatography and further expanded by upon the aforementioned techniques. Should an interaction be found, the biological function of this activity could be assessed using a biofilm disruption assay[71,72,150,327]. As PelA 46-507 has been shown to separate into monomeric and dimeric fractions during aSEC, such assays would also allow us to differentiate between whether recognition of PelA is driven by a binding site on the N2 domain or is the result of recognition of the PelA trimer.

Some of the preliminary work to test this hypothesis has been completed as the full soluble domain of PelB (47-880) and the TPR section (PelB 351-588) identified as interacting with PelA[150] have been cloned into *pEHisTEV* using the more rapid sequence and ligation independent cloning method (Chapter 2.2.18). Additionally, sequence analysis of the PelA protein, and the point at cleavage point identified by mass spectroscopy, suggested that a thermolysin-type enzymatic activity might be responsible for degradation of the protein. To provide a suitable PelA 46-948 control for this experiment, the same method was used to clone a construct with a C-terminal His tag, mirroring the PelA constructed used in the original interaction study[150], and to mutate the predicted cleavage site residues (522-523) to glycine to prevent degradation. The remaining work of expression, purification and assay will need to be performed by a new set of hands.

A second possible role for PelA N2 is that it may be involved in pel binding. A homology model based on superposition of the structures of the N1 and N2 domains of PelA with the corresponding domains of the structure of the A4-β galactosidase from *Thermus thermophilus* yields an intriguing trimeric structure containing a channel of roughly 30 Å diameter – certainly large enough to accommodate a pel polysaccharide and which presents the active site of the N1 TIM domain towards the channel interior. The GATase enzymes from which the N2 fold originates are involved in binding of the amino acid glutamine and the transfer of the ammonia group to from this substrate to a second molecule. The degenerate active site would prevent transfer in this way but may allow recognition of the amine group of the galactosamine and glucosamine sugars that make up pel and therefore offer a means of differentiation between N-acetylated and de-N-acetylated sugars. It has been extensively shown the N1 domain is active in isolation from the remainder of PelA[71,72,150,327] and thus that N2 is not necessary for this activity, however, the activity of the de-N-acetylase domain of PelA has only been measured using protein containing the N2 domain[149] and so could be reliant upon it. The C-terminus of PelA without the N2 domain has also been consistently found to be insoluble (this study), suggesting that

the N2 domain may be necessary for stability or correct folding of the C-terminus.

Testing whether PelA N2 is involved in Pel binding is technically challenging as the Pel polysaccharide is insoluble, which prevents the use of traditional techniques such as surface plasmon resonance or isothermal titration calorimetry from functioning to measure binding kinetics. It might be possible to determine whether PelA N2 is able to bind galactosamine, N-acetyl-galactosamine, glucosamine or N-acetyl glucosamine using these methods, however, such an experiment would not be able to rule out these activities on a negative result.

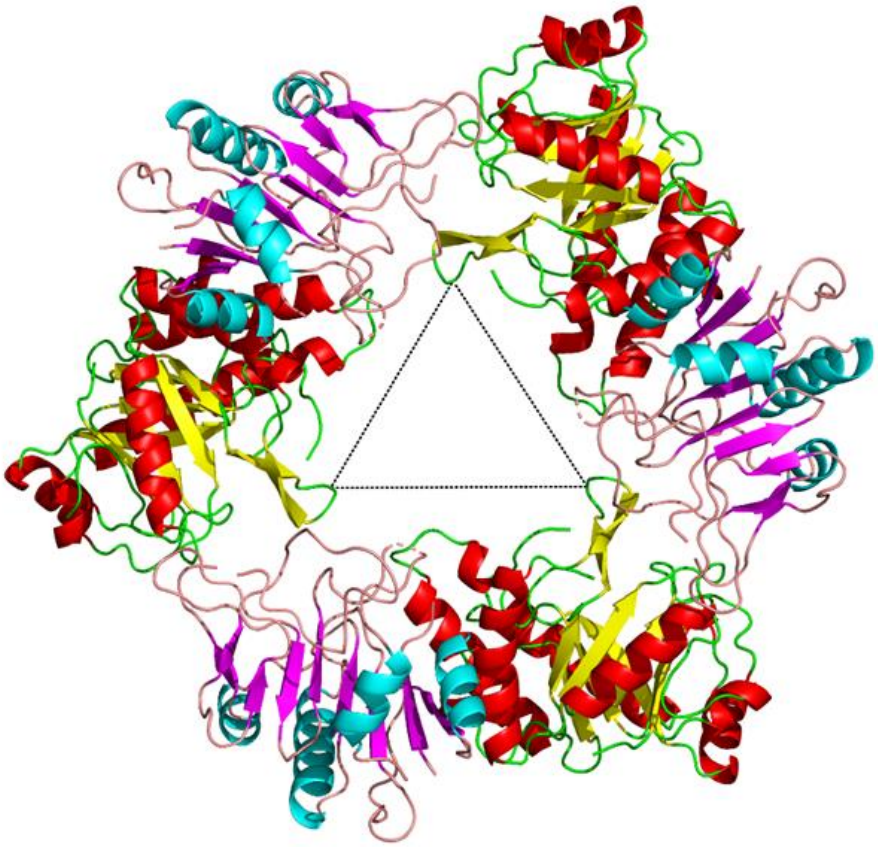

Figure 62: Molecular model of a putative PelA N1N2 timer based on the crystal structure of an A4-β galactosidase from Thermus thermophilus (PDB: 1KWG), a trimeric glycoside hydrolase. Domains are coloured according to secondary structure (N1 domains: α-helix red, β-strand yellow and N2 domains: α-helix cyan, β-strand magenta). Dotted lines are approximately 30Å in length and indicate the 3-fold symmetry of complex.

The results of size exclusion chromatographic analysis of PelA N1N2 presented in Chapter 3 suggest an oligomer is formed in solution. However, the order of the complex is unclear as its apparent molecular weight was found to lie intermediate between a dimer and trimer. No structures are available of homologues forming a dimeric biological unit. To test whether a trimeric structure could be a plausible arrangement for PelA N1N2 a

molecular model was built based on the known structure of an A4-β galactosidase from *Thermus thermophilus* (PDB: 1KWG), a trimeric glycoside hydrolase consisting of an N-terminal TIM domain followed by a homologous GAT-like domain[317]. Superposition of the PelA N1 and N2 domains with the corresponding domains of the A4-β galactosidase using Dali gave Z-scores of 9.6 and 6.0, respectively. This process produced a model (see Figure 62) where trimer assembly is predicted to involve intermolecular contacts involving both the N1 and N2 domains of PelA. This trimeric arrangement generates a channel of diameter approximately 30 Å with the catalytic residues of the N1 domain TIM barrels pointing towards the channel interior. This is consistent with the 32 Å diameter lumen of the PelC oligomer[152] through which mature pel must migrate to reach the pore of the OM transporter PelB, lending further plausibility to this model.

Together, this data regarding PelA suggests that the protein forms a trimeric complex channel in which the proof-reading domain N1 is placed at either the entrance or exit of the channel where it actively hydrolases passing Pel polysaccharide. The de-*N*-acetylase domains of the trimer are positioned at the opposing end of the channel based upon the C-terminus of the N2 domain, therefore any Pel that reaches the processing domain must pass by the N1 domain either upon its entrance or exit to the channel, depending upon orientation, and is therefore broken down. The N1 domain has been shown here to be incapable of oligomerisation alone but has been shown elsewhere to be active as a monomer[71,150,327], suggesting that oligomerisation of the protein is related to another function such as regulation of the N1 domain or the activity of the de-*N*-acetylase domain. Given that both of these functions are essential for Pel production, a clearer view on how this occurs would be invaluable to drug design in this area, both due to the potential to hinder these interactions and to better understand the accessible surfaces of the protein complex.

The structure presented in this work is of monomeric PelA N1-N2, as the potential multimeric forms of the protein during purification were assumed to be aggregate rather than a complexed protein. Given that these complexes can be isolated during purification, this opens an immediate avenue to attempting to crystallise the complex of the protein and gain definitive information regarding the residues responsible for oligomerisation. This would also allow ePISA[326] analysis of the complex to help determine the biological unit of the complex, as current estimates do not definitively support a single integer. Alternatively, should the complex not be amenable to crystallisation, the issue of biological unit could be addressed using analytical ultracentrifugation.

Despite advances in our understanding of the N-terminus of PelA, the C-terminal region remains a relative mystery, structurally speaking. The third domain of PelA is predicted to be a de-*N*-acetylase, and has been shown to facilitate that function[149], preceding a predicted β-jellyroll of unknown function. This latter fold is most commonly associated with viral capsids and may, speculatively, form a second oligomerisation site at the tail end of the protein. Even more speculatively, the broad nature of the β-sheets of

the oligomerised jellyroll domain combined with the need to maintain a 30 Å lumen channel for mature pel to pass through can be envisaged as an ideal shape to form a "suction cup" that would cover the lumen of oligomerised PelC and allow pel to be fed directly from PelA to the OM components of the system. Significantly more information would be needed to support this conclusion, however, subcellular fractionation of PelA does associate it with the OM as well as the periplasmic fractions[149] and the soluble domain of PelB does form important regulatory contacts with PelA[150].

Due to the unexpected size of the PelA 46-948 oligomer, an accurate measure of the biological unit of the enzyme with an intact C-terminus could not be made using the available aSEC columns. This illustrates the need to perform a more robust investigation using an alternative technique; namely analytical ultracentrifugation. Unfortunately, the degradation of the protein significantly increases the convolution of assessing the results using this method, peaks will appear for both PelA 46-948 and PelA 46-523 as well as combinations of the two. Initial experiments suggest that the oligomerisation state of PelA 46-948 is potentially octameric (see appendix), implying that the C-terminus of PelA contributes to oligomerisation, however, a more stable construct of PelA 46-948, such as is being made for the interaction studies with PelB, would be needed to draw firm conclusions.

A complete structure of PelA 46-948 remains an attractive prospect but has thus far not been feasible in this study due to contamination from the degraded protein. The knowledge that the protein forms oligomers also poses an interesting challenge as this was previously assumed to be aggregate. Should new PelA 46-948 construct prove to produce stable protein then further experimentation could be attempted in this field. Attempts to purify the monomer could follow the existing protocol, while attempts to purify the oligomer are likely to prove to be more challenging due to the need to separate it from protein aggregate; the separation of a 16/600 HiPrep Superdex 200PG column (GE Healthcare) would be insufficient to these needs. Alternative methods would be to centrifuge the protein sample, to hopefully pellet the aggregate, and/or to pass the sample through a 0.2 µm scale filter, to try to catch aggregate, prior to sample application to the column. A less orthodox approach would be to additionally run the sample through multiple gel filtration columns in series to provide sufficient volume for sample resolution.

While nothing has been published to the effect, it is highly likely that attempts to crystallise the PelA monomer have occurred elsewhere unsuccessfully. One method which may improve the chances of a successful crystal under these circumstances would be to incubate the protein with a small amount of protease in the crystallisation experiment; the protease could hopefully break down small, accessible (and more likely to be disordered) regions of the protein and reduce the size of the unit cell to improve the chances of a crystal forming[329,330]. Alternatively, cryo-electron microscopy may be a more suitable alternative to crystallography,

given that an oligomeric 105 kDa protein is likely to produce poor resolution crystals and that improvements to cryo EM have vastly narrowed the resolution gap between the two techniques[241]. This technique would also allow the identification of the oligomerisation interfaces of the protein and could be used to structurally characterised the interaction with PelB.

The prospects of advancing our understanding of other members of the Pel system are a little bleaker than for PelA. PelB proved difficult to clone and was not expressed successfully in *E.coli* from a synthesized construct. This is consistent with the work of Lindsay, *et al*[150,152]whose work studying interactions of PelB with PelA and with PelC relied upon truncated constructs of the protein rather than the whole protein. Given that both studies utilised crystallographic methods, including providing a structure of a TPR region of the protein, it is likely that the independent regions of PelB have been exposed to extensive crystallisation studies and are either not amenable to crystallisation or in the process of awaiting publication. While PelB remains an attractive potential drug target, it seems that without a specific advantage in the field that this project would be best explored using a different methodology, such as high resolution cryo-EM, and by expressing the protein as stoichiometric truncations rather than as a whole protein.

The structure of a homologue of PelC has been solved[152], notably, before the initiation of this project. This publication has also noted attempts to crystallise the protein from *P. aeruginosa* were unsuccessful, in multiple forms, likewise suggesting that further work here would be of limited value. The value of pursing PelD as a structural target is similarly limited by the existing structures[112,113], however, the function of PelD represents a knowledge gap in how interactions of the cytosolic domain of the protein with other components of the Pel system is achieved. Current models inconclusively speculate as to whether PelF, PelG or a combination of the two interact with PelD upon activation by cyclic-di-GMP. In this situation further work with PelD, especially the cytosolic domain which has been shown to be soluble and open to expression, would be justified in the event of successful purification of either PelF or PelG to support interaction protein: protein interaction studies to determine binding partners using co-immunoprecipitation[331] as a preferred method, due to the complications posed by membrane proteins for methods such as isothermal titration calorimetry[332] or surface plasmon resonance[333], which could then follow to provide more data.

This study has found that the soluble domain of PelE is viable during expression and can likely be purified (see Figure 33, section 3.8). Very little is known about this protein beyond a predicted role in protein: protein interactions and inner membrane localisation based upon its sequence. While current models[117] place the soluble domain of PelE in the periplasmic locale, this has not yet been confirmed experimentally and it remains possible that this domain is cytosol-facing[334]. In fact, cytosolic facing PelE would provide an attractive model for determining localisation of the IM components of the system with cytosolic PelF, which would provide greater

potential for co-operative regulation of multiple system components via PelD. Co-immunoprecipitation would be an ideal method for this purpose, although an approach more accessible to existing work would be to express and purify the available PelE soluble domain and utilise the N-terminal, and therefore membrane facing, His tag to bring the protein to an IMAC column and to then expose the column to cell lysate from *Pseudomonas aeruginosa* PA14 growing as Pel positive biofilms. This would capture PelE's binding partners and, in combination with mass spectroscopy, allow their identification and confirm the orientation of PelE in addition. Once appropriate binding partners are confirmed, this would allow the soluble domain of PelE to be used in co-crystallisation experiments, which may benefit the stability of partner proteins. A second method of interrogating this hypothesis would be to perform aSEC analysis of co-elution for the soluble domain of PelE in the presence and absence of PelA 46-948, its most likely periplasmic binding target. Should PelE be found to be cytosolic facing, rather than periplasmic, then this may go some way to explain the difficulties in expressing PelF.

This study has found that PelF can be expressed in *E.coli*, but appears within inclusion bodies and extraction using gently denaturing techniques and detergents has revealed that the protein does not behave as a correctly folded protein. PelF has previously been reported as being expressed as a soluble protein[123] using an N-terminal decahistidine tag in pBBR1-MCS-5 and with *Pseudomonas aeruginosa* PA01 as an expression vector. A noticeable difference in protocols is that the previous study did not report inducing overexpression of the protein before harvesting cells, which would not be viable for crystallographic experiments. It is possible that the ability of this system to produce stable PelF is a result of the presence of cytoplasmic chaperones not shared with *E.coli*, in which case further attempts to purify PelF would benefit from expression using a *Pseudomonas* species. While *P. aeruginosa* is a less than ideal expression vector *P. fluorescens* has had more widely used success with recombinant protein expression[335–337] and represents a platform that might benefit the expression of the Pel system proteins in general. Alternatively, experiments assessing protein: protein interactions between PelF and PelD, PelE or PelG would not require the same level of protein as crystallographic studies and may be viable based upon the method previously used to produce soluble PelF.

Additionally, a significant knowledge gap persists in the state of the pel polysaccharide as it is synthesised in the cytoplasm, with our only knowledge coming from mass spectroscopy of the mature polymer[10] which contradicts earlier conclusions that UDP-glucose is the principle substrate of the enzyme[123]. The earlier study focused upon only detection of UDP-glucose alone and therefore did not investigate the potential of UDP-galactose, and N-acetyl variants of those sugars, as substrates. The role of PelD and cyclic di-GMP were also not explored in this investigation. At present, the most likely scenario for the synthesis of Pel is that UDP-GalNAc and UDP-GluNAc are the typical substrates of PelF and that the only

modification to occur to the polymer is de-*N*-acetylation by PelA but more robust biochemical assays of PelF's glycosyltransferase activity, such as used in Chapter 4, would provide much needed direct evidence identifying the substrates of the enzyme, whether it's activity relates to PelD and, via mass spectroscopy, the length of the immature pel polysaccharides fed to PelG. This would represent a significant contribution to understanding the biosynthetic pathways the Pel system relies upon.

The IM protein PelG has not been successfully expressed in this study, but neither have avenues for improving expression been explored. The IM compartment is the smallest compartment available for expressions of proteins in *E.coli*, and thus the expression of IM proteins suffers from a lower maximum potential yield and heightened selection pressure towards non-producing cells than other types of protein, factors which were not considered during expression trials. The first of these issues can be address by increasing culture volume and by utilising a more sensitive technique for protein detection, such as Western blotting. The second issue can be addressed by reducing the level of stress the cells are subjected to; significantly lower levels of inducer, or even an absence of inducer, have been shown to be beneficial to protein yield[338]. Other alternatives include the use of a *Pseudomonas* species for expression, to assure the presence of appropriate chaperones and limit any divergences in Sec machinery, and co-expression of the protein alongside the IM proteins PelD and PelE, as interactions between them may be required to form a stable protein complex.

This study has identified the key binding residues for UDP in the glycosyltransferase WaaB in *S. enterica*, in addition to the hinge region that mediates the conformational change between open and closed forms of the enzyme. This presents a significant advance in our understanding of the structure of the enzyme and identifies numerous vital points within that structure where drug binding has the potential to undermine the activity of the enzyme. The Waa series of enzymes are attractive targets for drug development as undermining the process of OM biogenesis has the potential to either kill bacteria outwrite or to cause those bacteria to become vulnerable to antibiotics that would generally only be effective against Gram positive bacteria[339]. Additionally, this work has established a suitable assay for determining enzymatic activity for WaaB, which is a necessary component to determining the efficiency of targets during the drug development process, and suitable crystallisation conditions for both the open and closed conformations of the enzyme, which will guide inhibitor-binding co-crystallisation studies to determine the mode of action for drug targets.

A dynamic interaction between F13 and W243 has been identified which is likely to be necessary for the conformational shift from open to closed; the importance of W243 has already been shown but the effects of an F13A mutant has yet to be investigated and might reflect a mutation of the acceptor-substrate domain that could affect the hydrolytic activity of the donor-substrate domain. Should this be the case then drugs obfuscating

this region might inhibit enzyme activity by reducing conformational flexibility and provide another structural weak point for drug design. The experiments necessary to assess this possibility follow the workflow outlined in Chapter 4.

A potential membrane-anchor region has also been identified at Phe61-76Tyr on the N-terminal domain has also been identified by comparison with WaaA[323] and WaaG[202]. It has been suggested that this area of the protein may cause the conformational shift between open and closed conformations in response to membrane association. Analysis using PyMol suggests that this is one of the regions of the protein to change the least in adopting the mutant's closed conformation. As a large change in conformation would be expected at a regulatory region such as was proposed, this suggests that this mechanism is not conserved in WaaB. It would be interesting to know whether this is indicative of a divergence from other Waa proteins through an alternative mode of membrane interaction or by eschewing membrane association altogether, subcellular fraction with the protein should be informative in this regard with a cytoplasmic localisation suggesting a relatively late timing for glycosylation of LPS by WaaB in the sequence of construction by glycosyltransferases. Regardless, this indicates that attempts to exploit this region in drug design are unlikely to yield a positive outcome.

GtfC from the rat-derived strain 100-23 and from the pig-derived strain ATCC 53608 catalyse the addition of a glucose or N-acetylglucosamine moiety to SSRP, respectively, despite sharing a sequence identity of 97%[232]. Such a similarity between glycosyltransferases with different sugar-donors has been observed before and been found to be informative over the nature of the donor-substrate binding site of glycosyltransferase (GT) family 6, which adopts and GT-A fold [119,127,233,248]. GtfC was predicted to belong to either the GT-3 family by sequence[248], or to a smaller subfamily of SRRP GTs with structural similarity to the GT-4, 5 and 20 families[274]. Regardless, these proteins represent the first known pair of functionally diverse but sequentially similar enzymes predicted to adopt a GT-B family fold..

The crystal structures of $GtfC_{100\text{-}23}$ (in both the fully apo and mixed apo/UDP-bound forms) but not $GtfC_{ATCC\ 53608}$, were successfully determined in this study. As predicted, the protein adopts a GT-B fold containing two facing Rossman-like domains with the catalytic site occupying the cleft in between. Like related enzymes[274], $GtfC_{100\text{-}23}$ is a tetramer. Through examination of the structure and comparison with related enzymes, several potentially important features of the protein have been identified. UDP binding appears to be mediated by Leu224 and Trp214 and Leu220 at the uridine ring and Thr16, Arg179 and Lys251 at the phosphate. Phe319, Phe320, Lys323 and Leu325 are may be involved in oligomerisation of the protein. Potential catalytic residues include Glu101, Tyr243, Asn249 and His250 based upon the likely position of the UDP sugar.

Additionally, by comparison with the sequence of $GtfC_{ATCC\ 53608}$ and modelling the positions of UDP-glucose and UDP-N-acetylglucosamine within our structure we have identified three potential residues that are likely to determine the donor-substrate specificity of $GtfC_{100\text{-}23}$. The most apparent, Phe240, is positioned on C-α3 such that it is orientated directly towards the UDP-sugar. In the homologue, this residue is occupied by a cysteine. The role of this residue appears to be in determining the allowed length of the sugar moiety binding to the donor-substrate domain. A second residue of potential importance is Ser238 near the N-terminus of C-α2 becomes proline in $GtfC_{ATCC\ 53608}$, this has the potential to redirect the helix to bring it closer to the acceptor-binding domain and provide more space for the binding of UDP-N-acetylglucosamine. The final residue of interest is Leu174, which occupies the C-terminal-adjacent position of C-β1 and becomes the bulky phenylalanine in the homologue. This is likely to push the C-β1β2 loop further into the active cleft and redirect the UDP-sugar in line with the relative shift of C-α2 caused by Pro238 towards Cys240.

These residues were submitted to our collaborators at the Quadrum Institute for alanine mutagenesis and functional glycosyltransferase assays. It is our hope that these assays will reveal the mutations necessary to convert the substrate specificity of $GtfC_{100\text{-}23}$ to match that of $GtfC_{ATCC\ 53608}$ and in so doing provide valuable information to assist the structure-based engineering of glycosyltransferases of the GT-B fold.

Bibliography

1. Longitude Prize | Enter the prize. Available at: https://longitudeprize.org/. (Accessed: 18th August 2020)

2. Randremanana, R. *et al.* Epidemiological characteristics of an urban plague epidemic in Madagascar, August–November, 2017: an outbreak report. *Lancet Infect. Dis.* **19**, 537–545 (2019).

3. WHO | Plague – Madagascar. Available at: https://www.who.int/csr/don/15-november-2017-plague-madagascar/en/. (Accessed: 18th August 2020)

4. Manyi-Loh, C., Mamphweli, S., Meyer, E. & Okoh, A. Antibiotic use in agriculture and its consequential resistance in environmental sources: Potential public health implications. *Molecules* **23**, (2018).

5. Costerton, J. W., Stewart, P. S. & Greenberg, E. P. Bacterial biofilms: A common cause of persistent infections. *Science* **284**, 1318–1322 (1999).

6. Nickel, J. C., Ruseska, I., Wright, J. B. & Costerton, J. W. Tobramycin resistance of Pseudomonas aeruginosa cells growing as a biofilm on urinary catheter material. *Antimicrob. Agents Chemother.* **27**, 619–624 (1985).

7. Singh, P. K. *et al.* Quorum-sensing signals indicate that cystic fibrosis lungs are infected with bacterial biofilms. *Nature* **407**, 762–764 (2000).

8. Tseng, B. S. *et al.* The extracellular matrix protects Pseudomonas aeruginosa biofilms by limiting the penetration of tobramycin. *Environ. Microbiol.* **15**, 2865–2878 (2013).

9. Hall, C. W. & Mah, T. F. Molecular mechanisms of biofilm-based antibiotic resistance and tolerance in pathogenic bacteria. *FEMS Microbiology Reviews* **41**, 276–301 (2017).

10. Jennings, L. K. *et al.* Pel is a cationic exopolysaccharide that cross-links extracellular DNA in the Pseudomonas aeruginosa biofilm matrix. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 11353–8 (2015).

11. Bagge, N. *et al.* Dynamics and Spatial Distribution of β-Lactamase Expression in Pseudomonas aeruginosa Biofilms. *Antimicrob. Agents Chemother.* **48**, 1168–1174 (2004).

12. McPhee, J. B. *et al.* Contribution of the PhoP-PhoQ and PmrA-PmrB two-component regulatory systems to Mg2+-induced gene regulation in Pseudomonas aeruginosa. *J. Bacteriol.* **188**, 3995–4006 (2006).

13. Wilton, M., Charron-Mazenod, L., Moore, R. & Lewenza, S. Extracellular DNA acidifies biofilms and induces aminoglycoside resistance in Pseudomonas aeruginosa. *Antimicrob. Agents Chemother.* **60**, 544–553 (2016).

14. Lewenza, S. Extracellular DNA-induced antimicrobial peptide resistance mechanisms in Pseudomonas aeruginosa. *Frontiers in*

Microbiology **4**, (2013).

15. Johnson, L., Mulcahy, H., Kanevets, U., Shi, Y. & Lewenza, S. Surface-localized spermidine protects the Pseudomonas aeruginosa: Outer membrane from antibiotic treatment and oxidative stress. *J. Bacteriol.* **194**, 813–826 (2012).

16. Bae, J., Oh, E. & Jeon, B. Enhanced transmission of antibiotic resistance in campylobacter jejuni biofilms by natural transformation. *Antimicrob. Agents Chemother.* **58**, 7573–7575 (2014).

17. Evans, D. J., Allison, D. G., Brown, M. R. W. & Gilbert, P. Susceptibility of pseudomonas aeruginosa and escherichia coli biofilms towards ciprofloxacin: Effect of specific growth rate. *J. Antimicrob. Chemother.* **27**, 177–184 (1991).

18. Stewart, P. S. & Franklin, M. J. Physiological heterogeneity in biofilms. *Nature Reviews Microbiology* **6**, 199–210 (2008).

19. Werner, E. *et al.* Stratified growth in Pseudomonas aeruginosa biofilms. *Appl. Environ. Microbiol.* **70**, 6188–6196 (2004).

20. Borriello, G. *et al.* Oxygen limitation contributes to antibiotic tolerance of Pseudomonas aeruginosa in biofilms. *Antimicrob. Agents Chemother.* **48**, 2659–2664 (2004).

21. Schaible, B., Taylor, C. T. & Schaffer, K. Hypoxia increases antibiotic resistance in Pseudomonas aeruginosa through altering the composition of multidrug efflux pumps. *Antimicrob. Agents Chemother.* **56**, 2114–2118 (2012).

22. Tata, M. *et al.* RNAseq based transcriptional profiling of Pseudomonas aeruginosa PA14 after shortand long-term anoxic cultivation in synthetic cystic fibrosis sputum medium. *PLoS One* **11**, (2016).

23. Jensen, P. *et al.* Formation of hydroxyl radicals contributes to the bactericidal activity of ciprofloxacin against Pseudomonas aeruginosa biofilms. *Pathog. Dis.* **70**, 440–443 (2014).

24. Lewis, K. Multidrug tolerance of biofilms and persister cells. *Current Topics in Microbiology and Immunology* **322**, 107–131 (2008).

25. Spoering, A. L. & Lewis, K. Biofilms and planktonic cells of Pseudomonas aeruginosa have similar resistance to killing by antimicrobials. *J. Bacteriol.* **183**, 6746–6751 (2001).

26. Mah, T. F. *et al.* A genetic basis for Pseudomonas aeruginosa biofilm antibiotic resistance. *Nature* **426**, 306–310 (2003).

27. Sadovskaya, I. *et al.* High-level antibiotic resistance in Pseudomonas aeruginosa biofilm: The ndvB gene is involved in the production of highly glycerol-phosphorylated β(1→3)-glucans, which bind aminoglycosides. *Glycobiology* **20**, 895–904 (2010).

28. Beaudoin, T., Zhang, L., Hinz, A. J., Parr, C. J. & Mah, T. F. The biofilm-specific antibiotic resistance gene ndvB is important for expression of ethanol oxidation genes in Pseudomonas aeruginosa

biofilms. *J. Bacteriol.* **194**, 3128–3136 (2012).

29. Liao, J. & Sauer, K. The MerR-like transcriptional regulator BrlR contributes to Pseudomonas aeruginosa biofilm tolerance. *J. Bacteriol.* **194**, 4823–4836 (2012).

30. Liao, J., Schurr, M. J. & Sauera, K. The merR-like regulator brlR confers biofilm tolerance by activating multidrug efflux pumps in Pseudomonas aeruginosa biofilms. *J. Bacteriol.* **195**, 3352–3363 (2013).

31. Chambers, J. R. & Sauer, K. The MerR-like regulator BrlR impairs Pseudomonas aeruginosa biofilm tolerance to colistin by repressing PhoPQ. *J. Bacteriol.* **195**, 4678–4688 (2013).

32. Poudyal, B. & Sauer, K. The ABC of biofilm drug tolerance: the MerR-like regulator brlr is an activator of ABC transport systems, with PA1874-77 contributing to the tolerance of Pseudomonas aeruginosa biofilms to tobramycin. *Antimicrob. Agents Chemother.* **62**, (2018).

33. Gupta, K., Liao, J., Petrova, O. E., Cherny, K. E. & Sauer, K. Elevated levels of the second messenger c-di-GMP contribute to antimicrobial resistance of pseudomonas aeruginosa. *Mol. Microbiol.* **92**, 488–506 (2014).

34. Chambers, J. R., Liao, J., Schurr, M. J. & Sauer, K. BrlR from pseudomonas aeruginosa is a c-di-GMP-responsive transcription factor. *Mol. Microbiol.* **92**, 471–487 (2014).

35. Sukhodolets, M. V., Cabrera, J. E., Zhi, H. & Ding Jun Jin. RapA, a bacterial homolog of SWI2/SNF2, stimulates RNA polymerase recycling in transcription. *Genes Dev.* **15**, 3330–3341 (2001).

36. Lynch, S. V. *et al.* Role of the rapA gene in controlling antibiotic resistance of Escherichia coli biofilms. *Antimicrob. Agents Chemother.* **51**, 3650–3658 (2007).

37. Li, X. Z., Plésiat, P. & Nikaido, H. The challenge of efflux-mediated antibiotic resistance in Gram-negative bacteria. *Clin. Microbiol. Rev.* **28**, 337–418 (2015).

38. Issa, K. H. B., Phan, G. & Broutin, I. Functional mechanism of the efflux pumps transcription regulators from Pseudomonas aeruginosa based on 3D structures. *Frontiers in Molecular Biosciences* **5**, 57 (2018).

39. Mousa, J. J. & Bruner, S. D. Structural and mechanistic diversity of multidrug transporters. *Natural Product Reports* **33**, 1255–1267 (2016).

40. Poole, K. Pseudomonas aeruginosa: Resistance to the max. *Front. Microbiol.* **2**, (2011).

41. Zhang, L. & Mah, T. F. Involvement of a novel efflux system in biofilm-specific resistance to antibiotics. *J. Bacteriol.* **190**, 4447–4452 (2008).

42. Poole, K. Stress responses as determinants of antimicrobial

resistance in Pseudomonas aeruginosa: Multidrug efflux and more. *Canadian Journal of Microbiology* **60**, 783–791 (2014).

43. Mandsberg, L. F. *et al.* Antibiotic resistance in Pseudomonas aeruginosa strains with increased mutation frequency due to inactivation of the DNA oxidative repair system. *Antimicrob. Agents Chemother.* **53**, 2483–2491 (2009).

44. Driffield, K., Miller, K., Bostock, J. M., O'neill, A. J. & Chopra, I. Increased mutability of Pseudomonas aeruginosa in biofilms. *J. Antimicrob. Chemother.* **61**, 1053–1056 (2008).

45. Boucher, J. C., Yu, H., Mudd, M. H. & Deretic, V. Mucoid Pseudomonas aeruginosa in cystic fibrosis: Characterization of muc mutations in clinical isolates and analysis of clearance in a mouse model of respiratory infection. *Infect. Immun.* **65**, 3838–3846 (1997).

46. Häußler, S., Tümmler, B., Weißbrodt, H., Rohde, M. & Steinmetz, I. Small-colony variants of Pseudomonas aeruginosa in cystic fibrosis. *Clin. Infect. Dis.* **29**, 621–625 (1999).

47. Déziel, E., Comeau, Y. & Villemur, R. Initiation of biofilm formation by Pseudomonas aeruginosa 57RP correlates with emergence of hyperpiliated and highly adherent phenotypic variants deficient in swimming, swarming, and twitching motilities. *J. Bacteriol.* **183**, 1195–1204 (2001).

48. Häußler, S. *et al.* Highly adherent small-colony variants of Pseudomonas aeruginosa in cystic fibrosis lung infection. *J. Med. Microbiol.* **52**, 295–301 (2003).

49. Meissner, A. *et al.* Pseudomonas aeruginosa cupA-encoded fimbriae expression is regulated by a GGDEF and EAL domain-dependent modulation of the intracellular level of cyclic diguanylate. *Environ. Microbiol.* **9**, 2475–2485 (2007).

50. Drenkard, E. & Ausubel, F. M. Pseudomonas biofilm formation and antibiotic resistance are linked to phenotypic variation. *Nature* **416**, 740–743 (2002).

51. Phippen, C. W. *et al.* Formation and dimerization of the phosphodiesterase active site of the *Pseudomonas aeruginosa* MorA, a bi-functional c-di-GMP regulator. *FEBS Lett.* **588**, 4631–4636 (2014).

52. Davies, J. A. *et al.* The GacS sensor kinase controls phenotypic reversion of small colony variants isolated from biofilms of Pseudomonas aeruginosa PA14. *FEMS Microbiol. Ecol.* **59**, 32–46 (2007).

53. Nadal Jimenez, P. *et al.* The Multiple Signaling Systems Regulating Virulence in Pseudomonas aeruginosa. *Microbiol. Mol. Biol. Rev.* **76**, 46–65 (2012).

54. Donlan, R. M. & Costerton, J. W. Biofilms: Survival Mechanisms of Clinically Relevant Microorganisms. *Clin. Microbiol. Rev.* **15**, 167

(2002).

55. Buhl, M., Peter, S. & Willmann, M. Prevalence and risk factors associated with colonization and infection of extensively drug-resistant Pseudomonas aeruginosa: A systematic review. *Expert Review of Anti-Infective Therapy* **13**, 1159–1170 (2015).

56. Gonçalves-de-Albuquerque, C. F. *et al.* Possible mechanisms of Pseudomonas aeruginosa-associated lung disease. *International Journal of Medical Microbiology* **306**, 20–28 (2016).

57. O'callaghan, R., Caballero, A., Tang, A. & Bierdeman, M. Pseudomonas aeruginosa keratitis: Protease iv and pasp as corneal virulence mediators. *Microorganisms* **7**, (2019).

58. Horvat, R. T. & Parmely, M. J. Pseudomonas aeruginosa alkaline protease degrades human gamma interferon and inhibits its bioactivity. *Infect. Immun.* **56**, 2925–2932 (1988).

59. Parmely, M., Gale, A., Clabaugh, M., Horvat, R. & Zhou, W. W. Proteolytic inactivation of cytokines by Pseudomonas aeruginosa. *Infect. Immun.* **58**, 3009–3014 (1990).

60. Vessillier, S., Delolme, F., Bernillon, J., Saulnier, J. & Wallach, J. Hydrolysis of glycine-containing elastin pentapeptides by LasA, a metalloelastase from Pseudomonas aeruginosa. *Eur. J. Biochem.* **268**, 1049–1057 (2001).

61. Kharazmi, A., Eriksen, H. O., DÖRing, G., Goldstein, W. & HØIby, N. EFFECT OF PSEUDOMONAS AERUGINOSA PROTEASES ON HUMAN LEUKOCYTE PHAGOCYTOSIS AND BACTERICIDAL ACTIVITY. *Acta Pathol. Microbiol. Scand. Ser. C Immunol.* **94 C**, 175–179 (1986).

62. Pedersen, B. K. & Kharazmi, A. Inhibition of human natural killer cell activity by Pseudomonas aeruginosa alkaline protease and elastase. *Infect. Immun.* **55**, 986–989 (1987).

63. Heck, L. W., Morihara, K. & Abrahamson, D. R. Degradation of soluble laminin and depletion of tissue-associated basement membrane laminin by Pseudomonas aeruginosa elastase and alkaline protease. *Infect. Immun.* **54**, 149–153 (1986).

64. Shibuya, Y. *et al.* Pseudomonas aeruginosa alkaline proteinase might share a biological function with plasmin. *Biochim. Biophys. Acta (BBA)/Protein Struct. Mol.* **1077**, 316–324 (1991).

65. Toder, D. S., Gambello, M. J. & Iglewski, B. H. Pseudomonas aeruginosa LasA: a second elastase under the transcriptional control of lasR. *Mol. Microbiol.* **5**, 2003–2010 (1991).

66. Guyot, N. *et al.* Functional study of elafin cleaved by Pseudomonas aeruginosa metalloproteinases. *Biol. Chem.* **391**, 705–716 (2010).

67. Tümmler, B. & Klockgether, J. Recent advances in understanding Pseudomonas aeruginosa as a pathogen. *F1000Research* **6**, (2017).

68. Ciofu, O., Rojo-Molinero, E., Macià, M. D. & Oliver, A. Antibiotic treatment of biofilm infections. *APMIS* **125**, 304–319 (2017).

69. World Health Organisation. WHO publishes list of bacteria for which new antibiotics are urgently needed. Available at: https://www.who.int/en/news-room/detail/27-02-2017-who-publishes-list-of-bacteria-for-which-new-antibiotics-are-urgently-needed. (Accessed: 12th July 2019)

70. Goltermann, L. & Tolker-Nielsen, T. Importance of the Exopolysaccharide Matrix in Antimicrobial Tolerance of Pseudomonas aeruginosa Aggregates. *Antimicrob. Agents Chemother.* **61**, (2017).

71. Baker, P. *et al.* Exopolysaccharide biosynthetic glycoside hydrolases can be utilized to disrupt and prevent Pseudomonas aeruginosa biofilms. *Sci. Adv.* **2**, e1501632 (2016).

72. Pestrak, M. J. *et al.* Treatment with the *Pseudomonas aeruginosa* glycoside hydrolase PslG combats wound infection by improving antibiotic efficacy and host innate immune activity. *Antimicrob. Agents Chemother.* (2019). doi:10.1128/AAC.00234-19

73. Olsen, I. Biofilm-specific antibiotic tolerance and resistance. *Eur. J. Clin. Microbiol. Infect. Dis.* **34**, 877–886 (2015).

74. Walters, M. C., Roe, F., Bugnicourt, A., Franklin, M. J. & Stewart, P. S. Contributions of antibiotic penetration, oxygen limitation, and low metabolic activity to tolerance of Pseudomonas aeruginosa biofilms to ciprofloxacin and tobramycin. *Antimicrob. Agents Chemother.* **47**, 317–23 (2003).

75. Valentini, M. & Filloux, A. Biofilms and Cyclic di-GMP (c-di-GMP) signaling: Lessons from Pseudomonas aeruginosa and other bacteria. *Journal of Biological Chemistry* **291**, 12547–12555 (2016).

76. Basu Roy, A. & Sauer, K. Diguanylate cyclase NicD-based signalling mechanism of nutrient-induced dispersion by Pseudomonas aeruginosa. *Mol. Microbiol.* **94**, 771–793 (2014).

77. Schmidt, A. J., Ryjenkov, D. A. & Gomelsky, M. The ubiquitous protein domain EAL is a cyclic diguanylate-specific phosphodiesterase: Enzymatically active and inactive EAL domains. *J. Bacteriol.* **187**, 4774–4781 (2005).

78. Ryjenkov, D. A., Tarutina, M., Moskvin, O. V. & Gomelsky, M. Cyclic diguanylate is a ubiquitous signaling molecule in bacteria: Insights into biochemistry of the GGDEF protein domain. *J. Bacteriol.* **187**, 1792–1798 (2005).

79. Christen, M. *et al.* Asymmetrical distribution of the second messenger c-di-GMP upon bacterial cell division. *Science (80-.).* **328**, 1295–1297 (2010).

80. Merritt, J. H. *et al.* Specific control of Pseudomonas aeruginosa surface-associated behaviors by two c-di-GMP diguanylate cyclases. *MBio* **1**, (2010).

81. Hay, I. D., Remminghorst, U. & Rehm, B. H. A. MucR, a novel membrane-associated regulator of alginate biosynthesis in Pseudomonas aeruginosa. *Appl. Environ. Microbiol.* **75**, 1110–1120

(2009).

82. Moradali, M. F., Ghods, S. & Rehm, B. H. A. Activation Mechanism and Cellular Localization of Membrane-Anchored Alginate Polymerase in Pseudomonas aeruginosa. *Appl. Environ. Microbiol.* **83**, (2017).

83. Lee, V. T. *et al.* A cyclic-di-GMP receptor required for bacterial exopolysaccharide production. *Mol. Microbiol.* **65**, 1474–1484 (2007).

84. Merighi, M., Lee, V. T., Hyodo, M., Hayakawa, Y. & Lory, S. The second messenger bis-(3′-5′)-cyclic-GMP and its PilZ domain-containing receptor Alg44 are required for alginate biosynthesis in Pseudomonas aeruginosa. *Mol. Microbiol.* **65**, 876–895 (2007).

85. Baraquet, C., Murakami, K., Parsek, M. R. & Harwood, C. S. The FleQ protein from Pseudomonas aeruginosa functions as both a repressor and an activator to control gene expression from the Pel operon promoter in response to c-di-GMP. *Nucleic Acids Res.* **40**, 7207–7218 (2012).

86. Mikkelsen, H., Sivaneson, M. & Filloux, A. Key two-component regulatory systems that control biofilm formation in Pseudomonas aeruginosa. *Environmental Microbiology* **13**, 1666–1681 (2011).

87. Mancl, J. M., Ray, W. K., Helm, R. F. & Schubot, F. D. Helix Cracking Regulates the Critical Interaction between RetS and GacS in Pseudomonas aeruginosa. *Structure* **27**, 785–793 (2019).

88. Vincent, F. *et al.* Distinct oligomeric forms of the Pseudomonas aeruginosa RetS sensor domain modulate accessibility to the ligand binding site. *Environ. Microbiol.* **12**, 1775–1786 (2009).

89. Valentini, M., Laventie, B.-J., Moscoso, J., Jenal, U. & Filloux, A. The Diguanylate Cyclase HsbD Intersects with the HptB Regulatory Cascade to Control Pseudomonas aeruginosa Biofilm and Motility. *PLOS Genet.* **12**, e1006354 (2016).

90. Petrova, O. E. & Sauer, K. The novel two-component regulatory system BfiSR regulates biofilm development by controlling the small RNA rsmZ through CafA. *J. Bacteriol.* **192**, 5275–5288 (2010).

91. Schniederberend, M. *et al.* Modulation of flagellar rotation in surfaceattached bacteria: A pathway for rapid surface-sensing after flagellar attachment. *PLoS Pathog.* **15**, (2019).

92. Dasgupta, N., Ferrell, E. P., Kanack, K. J., West, S. E. H. & Ramphal, R. fleQ, the gene encoding the major flagellar regulator of Pseudomonas aeruginosa, is σ70 dependent and is downregulated by Vfr, a homolog of Escherichia coli cyclic AMP receptor protein. *J. Bacteriol.* **184**, 5240–5250 (2002).

93. West, S. E. H., Sample, A. K. & Runyen-Janecky, L. J. The vfr gene product, required for Pseudomonas aeruginosa exotoxin A and protease production, belongs to the cyclic AMP receptor protein family. *J. Bacteriol.* **176**, 7532–7542 (1994).

94. Beatson, S. A., Whitchurch, C. B., Sargent, J. L., Levesque, R. C. & Mattick, J. S. Differential regulation of twitching motility and elastase production by Vfr in Pseudomonas aeruginosa. *J. Bacteriol.* **184**, 3605–3613 (2002).

95. Wolfgang, M. C., Lee, V. T., Gilmore, M. E. & Lory, S. Coordinate regulation of bacterial virulence genes by a novel adenylate cyclase-dependent signaling pathway. *Dev. Cell* **4**, 253–263 (2003).

96. Fulcher, N. B., Holliday, P. M., Klem, E., Cann, M. J. & Wolfgang, M. C. The Pseudomonas aeruginosa Chp chemosensory system regulates intracellular cAMP levels by modulating adenylate cyclase activity. *Mol. Microbiol.* **76**, 889–904 (2010).

97. Persat, A., Inclan, Y. F., Engel, J. N., Stone, H. A. & Gitai, Z. Type IV pili mechanochemically regulate virulence factors in Pseudomonas aeruginosa. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 7563–7568 (2015).

98. Diaz, M. R., King, J. M. & Yahr, T. L. Intrinsic and extrinsic regulation of type III secretion gene expression in Pseudomonas aeruginosa. *Front. Microbiol.* **2**, (2011).

99. Almblad, H. *et al.* The cyclic AMP-Vfr signaling pathway in Pseudomonas aeruginosa is inhibited by cyclic Di-GMP. *J. Bacteriol.* **197**, 2190–2200 (2015).

100. Colvin, K. M. *et al.* The Pel and Psl polysaccharides provide Pseudomonas aeruginosa structural redundancy within the biofilm matrix. *Environ. Microbiol.* **14**, 1913–1928 (2012).

101. Zhao, K. *et al.* Psl trails guide exploration and microcolony formation in Pseudomonas aeruginosa biofilms. *Nature* **497**, 388–391 (2013).

102. Whitchurch, C. B., Tolker-Nielsen, T., Ragas, P. C. & Mattick, J. S. Extracellular DNA required for bacterial biofilm formation. *Science (80-.).* **295**, 1487 (2002).

103. Allesen-Holm, M. *et al.* A characterization of DNA release in Pseudomonas aeruginosa cultures and biofilms. *Mol. Microbiol.* **59**, 1114–1128 (2006).

104. Swearingen, M. C. *et al.* A novel technique using potassium permanganate and reflectance confocal microscopy to image biofilm extracellular polymeric matrix reveals non eDNA networks in *Pseudomonas aeruginosa* biofilms. *Pathog. Dis.* **74**, ftv104 (2015).

105. Irie, Y. *et al.* Self-produced exopolysaccharide is a signal that stimulates biofilm formation in Pseudomonas aeruginosa. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 20632–20636 (2012).

106. Barraud, N. *et al.* Nitric oxide signaling in Pseudomonas aeruginosa biofilms mediates phosphodiesterase activity, decreased cyclic di-GMP levels, and enhanced dispersal. *J. Bacteriol.* **191**, 7333–7342 (2009).

107. Sauer, K. *et al.* Characterization of nutrient-induced dispersion in Pseudomonas aeruginosa PAO1 biofilm. *J. Bacteriol.* **186**, 7312–

7326 (2004).

108. Roy, A. B., Petrova, O. E. & Sauer, K. The phosphodiesterase DipA (PA5017) is essential for Pseudomonas aeruginosa biofilm dispersion. *J. Bacteriol.* **194**, 2904–2915 (2012).

109. Morgan, R., Kohn, S., Hwang, S. H., Hassett, D. J. & Sauer, K. BdlA, a chemotaxis regulator essential for biofilm dispersion in Pseudomonas aeruginosa. *J. Bacteriol.* **188**, 7335–7343 (2006).

110. Petrova, O. E. & Sauer, K. Dispersion by Pseudomonas aeruginosa requires an unusual posttranslational modification of BdlA. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 16690–16695 (2012).

111. Ha, D.-G. & O'Toole, G. A. c-di-GMP and its Effects on Biofilm Formation and Dispersion: a Pseudomonas Aeruginosa Review. *Microbiol. Spectr.* **3**, MB-0003-2014 (2015).

112. Li, Z., Chen, J.-H., Hao, Y. & Nair, S. K. Structures of the PelD cyclic diguanylate effector involved in pellicle formation in Pseudomonas aeruginosa PAO1. *J. Biol. Chem.* **287**, 30191–204 (2012).

113. Whitney, J. C. *et al.* Structure of the Cytoplasmic Region of PelD, a Degenerate Diguanylate Cyclase Receptor That Regulates Exopolysaccharide Production in Pseudomonas aeruginosa. *J. Biol. Chem.* **287**, 23582–23593 (2012).

114. Hickman, J. W., Tifrea, D. F. & Harwood, C. S. A chemosensory system that regulates biofilm formation through modulation of cyclic diguanylate levels. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 14422–7 (2005).

115. Friedman, L. & Kolter, R. Genes involved in matrix formation in Pseudomonas aeruginosa PA14 biofilms. *Mol. Microbiol.* **51**, 675–690 (2003).

116. Madsen, J. S. *et al.* Facultative control of matrix production optimizes competitive fitness in Pseudomonas aeruginosa PA14 biofilm models. *Appl. Environ. Microbiol.* **81**, 8414–26 (2015).

117. Franklin, M. J., Nivens, D. E., Weadge, J. T. & Howell, P. L. Biosynthesis of the Pseudomonas aeruginosa Extracellular Polysaccharides, Alginate, Pel, and Psl. *Front. Microbiol.* **2**, 167 (2011).

118. Riley, L. M. *et al.* Structural and functional characterization of Pseudomonas aeruginosa AlgX: role of AlgX in alginate acetylation. *J. Biol. Chem.* **288**, 22299–314 (2013).

119. Lombard, V., Golaconda Ramulu, H., Drula, E., Coutinho, P. M. & Henrissat, B. The carbohydrate-active enzymes database (CAZy) in 2013. *Nucleic Acids Res.* **42**, D490–D495 (2014).

120. Ma, L. *et al.* Synthesis of multiple Pseudomonas aeruginosa biofilm matrix exopolysaccharides is post-transcriptionally regulated. *Environ. Microbiol.* **14**, 1995–2005 (2012).

121. Mulcahy, H., Charron-Mazenod, L. & Lewenza, S. *Pseudomonas*

aeruginosa produces an extracellular deoxyribonuclease that is required for utilization of DNA as a nutrient source. *Environ. Microbiol.* **12**, 1621–1629 (2010).

122. Devaraj, A., Justice, S. S., Bakaletz, L. O. & Goodman, S. D. DNABII proteins play a central role in UPEC biofilm structure. *Mol. Microbiol.* **96**, 1119–1135 (2015).

123. Ghafoor, A., Jordens, Z. & Rehm, B. H. A. Role of PelF in pel polysaccharide biosynthesis in Pseudomonas aeruginosa. *Appl. Environ. Microbiol.* **79**, 2968–78 (2013).

124. Stingele, F. *et al.* Introduction of the exopolysaccharide gene cluster from Streptococcus thermophilus Sfi6 into Lactococcus lactis MG1363: production and characterization of an altered polysaccharide. *Mol. Microbiol.* **32**, 1287–1295 (1999).

125. Kelley, L. A., Mezulis, S., Yates, C. M., Wass, M. N. & Sternberg, M. J. E. The Phyre2 web portal for protein modeling, prediction and analysis. *Nat. Protoc.* **10**, 845–858 (2015).

126. Zheng, Y., Anderson, S., Zhang, Y. & Garavito, R. M. The structure of sucrose synthase-1 from Arabidopsis thaliana and its functional implications. *J. Biol. Chem.* **286**, 36108–18 (2011).

127. Yamamoto, F., Clausen, H., White, T., Marken, J. & Hakomori, S. Molecular genetic basis of the histo-blood group ABO system. *Nature* **345**, 229–233 (1990).

128. Breton, C., Šnajdrová, L., Jeanneau, C., Koča, J. & Imberty, A. Structures and mechanisms of glycosyltransferases. *Glycobiology* **16**, 29R-37R (2006).

129. Hardin, S. C., Duncan, K. A. & Huber, S. C. Determination of Structural Requirements and Probable Regulatory Effectors for Membrane Association of Maize Sucrose Synthase 1. *Plant Physiol.* **141**, 1106–1119 (2006).

130. Kuk, A. C. Y., Mashalidis, E. H. & Lee, S.-Y. Crystal structure of the MOP flippase MurJ in an inward-facing conformation. *Nat. Struct. Mol. Biol.* **24**, 171–176 (2017).

131. Zheng, S. *et al.* Structure and mutagenic analysis of the lipid II flippase MurJ from Escherichia coli. *Proc. Natl. Acad. Sci. U. S. A.* **115**, 6709–6714 (2018).

132. Tanaka, Y., Iwaki, S. & Tsukazaki, T. Crystal Structure of a Plant Multidrug and Toxic Compound Extrusion Family Protein. *Structure* **25**, 1455-1460.e2 (2017).

133. Miyauchi, H. *et al.* Structural basis for xenobiotic extrusion by eukaryotic MATE transporter. *Nat. Commun.* **8**, 1633 (2017).

134. Kusakizako, T. *et al.* Structural Basis of H+-Dependent Conformational Change in a Bacterial MATE Transporter. *Structure* **27**, 293-301.e3 (2019).

135. Omote, H., Hiasa, M., Matsumoto, T., Otsuka, M. & Moriyama, Y.

The MATE proteins as fundamental transporters of metabolic and xenobiotic organic cations. *Trends Pharmacol. Sci.* **27**, 587–593 (2006).

136. Hvorup, R. N. *et al.* The multidrug/oligosaccharidyl-lipid/polysaccharide (MOP) exporter superfamily. *Eur. J. Biochem.* **270**, 799–813 (2003).

137. Quevillon, E. *et al.* InterProScan: protein domains identifier. *Nucleic Acids Res.* **33**, W116–W120 (2005).

138. Shahrizal, M. *et al.* Structural Basis for the Function of the β-Barrel Assembly-Enhancing Protease BepA. *J. Mol. Biol.* **431**, 625–635 (2019).

139. Baskin, J. M. *et al.* The leukodystrophy protein FAM126A (hyccin) regulates PtdIns(4)P synthesis at the plasma membrane. *Nat. Cell Biol.* **18**, 132–8 (2016).

140. Zeytuni, N. & Zarivach, R. Structural and Functional Discussion of the Tetra-Trico-Peptide Repeat, a Protein Interaction Module. *Structure* **20**, 397–405 (2012).

141. Peifer, M., Berg, S. & Reynolds, A. B. A repeating amino acid motif shared by proteins with diverse cellular roles. *Cell* **76**, 789–791 (1994).

142. Madhurantakam, C., Varadamsetty, G., Grütter, M. G., Plückthun, A. & Mittl, P. R. E. Structure-based optimization of designed Armadillo-repeat proteins. *Protein Sci.* **21**, 1015–28 (2012).

143. Tewari, R., Bailes, E., Bunting, K. A. & Coates, J. C. Armadillo-repeat protein functions: questions for little creatures. *Trends Cell Biol.* **20**, 470–481 (2010).

144. Weyer, F. A. *et al.* Structural basis of HypK regulating N-terminal acetylation by the NatA complex. *Nat. Commun.* **8**, 15726 (2017).

145. Liszczak, G. *et al.* Molecular basis for N-terminal acetylation by the heterodimeric NatA complex. *Nat. Struct. Mol. Biol.* **20**, 1098–105 (2013).

146. Lazarus, M. B., Nam, Y., Jiang, J., Sliz, P. & Walker, S. Structure of human O-GlcNAc transferase and its complex with a peptide substrate. *Nature* **469**, 564–7 (2011).

147. Zhang, Z., Kulkarni, K., Hanrahan, S. J., Thompson, A. J. & Barford, D. The APC/C subunit Cdc16/Cut9 is a contiguous tetratricopeptide repeat superhelix with a homo-dimer interface similar to Cdc27. *EMBO J.* **29**, 3733–44 (2010).

148. D'Andrea, L. D. & Regan, L. TPR proteins: the versatile helix. *Trends Biochem. Sci.* **28**, 655–662 (2003).

149. Colvin, K. M. *et al.* PelA deacetylase activity is required for Pel polysaccharide synthesis in Pseudomonas aeruginosa. *J. Bacteriol.* **195**, 2329–39 (2013).

150. Marmont, L. S. *et al.* PelA and PelB proteins form a modification and secretion complex essential for Pel polysaccharide-dependent biofilm formation in *Pseudomonas aeruginosa*. *J. Biol. Chem.* **292**, 19411–19422 (2017).

151. Wang, Y. *et al.* Structural Basis for Translocation of a Biofilm-supporting Exopolysaccharide across the Bacterial Outer Membrane. *J. Biol. Chem.* **291**, 10046–57 (2016).

152. Marmont, L. S. *et al.* Oligomeric lipoprotein PelC guides Pel polysaccharide export across the outer membrane of Pseudomonas aeruginosa. doi:10.1073/pnas.1613606114

153. Halbach, F., Reichelt, P., Rode, M. & Conti, E. The Yeast Ski Complex: Crystal Structure and RNA Channeling to the Exosome Complex. *Cell* **154**, 814–826 (2013).

154. Bertram, K. *et al.* Cryo-EM Structure of a Pre-catalytic Human Spliceosome Primed for Activation. *Cell* **170**, 701-713.e11 (2017).

155. Deng, P. *et al.* Transcriptional elongation factor Paf1 core complex adopts a spirally wrapped solenoidal topology. *Proc. Natl. Acad. Sci. U. S. A.* **115**, 9998–10003 (2018).

156. van den Berg, B., Black, P. N., Clemons, W. M. & Rapoport, T. A. Crystal structure of the long-chain fatty acid transporter FadL. *Science* **304**, 1506–9 (2004).

157. Hearn, E. M., Patel, D. R., Lepore, B. W., Indic, M. & van den Berg, B. Transmembrane passage of hydrophobic compounds through a protein channel wall. *Nature* **458**, 367–370 (2009).

158. Hearn, E. M., Patel, D. R. & van den Berg, B. Outer-membrane transport of aromatic hydrocarbons as a first step in biodegradation. *Proc. Natl. Acad. Sci. U. S. A.* **105**, 8601–6 (2008).

159. Yamashita, S. *et al.* Structural Insights into Ail-Mediated Adhesion in Yersinia pestis. *Structure* **19**, 1672–1682 (2011).

160. Amini, S., Goodarzi, H. & Tavazoie, S. Genetic dissection of an exogenously induced biofilm in laboratory and clinical isolates of E. coli. *PLoS Pathog.* **5**, e1000432 (2009).

161. Itoh, Y. *et al.* Roles of pgaABCD genes in synthesis, modification, and export of the Escherichia coli biofilm adhesin poly-beta-1,6-N-acetyl-D-glucosamine. *J. Bacteriol.* **190**, 3670–80 (2008).

162. Vasseur, P., Soscia, C., Voulhoux, R. & Filloux, A. PelC is a Pseudomonas aeruginosa outer membrane lipoprotein of the OMA family of proteins involved in exopolysaccharide transport. *Biochimie* **89**, 903–915 (2007).

163. Zgurskaya, H. I., López, C. A. & Gnanakaran, S. Permeability Barrier of Gram-Negative Cell Envelopes and Approaches to Bypass It. *ACS Infect. Dis.* **1**, 512–522 (2016).

164. Caroff, M. & Karibian, D. Structure of bacterial lipopolysaccharides. *Carbohydr. Res.* **338**, 2431–2447 (2003).

165. Gu, Y. *et al.* Lipopolysaccharide is inserted into the outer membrane through an intramembrane hole, a lumen gate, and the lateral opening of LptD. *Structure* **23**, 496–504 (2015).

166. Tamaki, S., Sato, T. & Matsuhashi, M. Role of lipopolysaccharides in antibiotic resistance and bacteriophage adsorption of Escherichia coli K-12. *J. Bacteriol.* **105**, 968–75 (1971).

167. Nikaido, H. Molecular basis of bacterial outer membrane permeability revisited. *Microbiol. Mol. Biol. Rev.* **67**, 593–656 (2003).

168. Ruiz, N., Kahne, D. & Silhavy, T. J. Transport of lipopolysaccharide across the cell envelope: the long road of discovery. *Nat. Rev. Microbiol.* **7**, 677–683 (2009).

169. Yuasa, R., Levinthal, M. & Nikaido, H. Biosynthesis of cell wall lipopolysaccharide in mutants of Salmonella. V. A mutant of Salmonella typhimurium defective in the synthesis of cytidine diphosphoabequose. *J. Bacteriol.* **100**, 433–44 (1969).

170. Majowicz, S. E. *et al.* The Global Burden of Nontyphoidal *Salmonella* Gastroenteritis. *Clin. Infect. Dis.* **50**, 882–889 (2010).

171. Jajere, S. M. A review of Salmonella enterica with particular focus on the pathogenicity and virulence factors, host specificity and antimicrobial resistance including multidrug resistance. *Vet. world* **12**, 504–521 (2019).

172. Bae, D., Cheng, C.-M. & Khan, A. A. Characterization of extended-spectrum β-lactamase (ESBL) producing non-typhoidal Salmonella (NTS) from imported food products. *Int. J. Food Microbiol.* **214**, 12–17 (2015).

173. Ren, D. *et al.* Phenotypes and antimicrobial resistance genes in *Salmonella* isolated from retail chicken and pork in Changchun, China. *J. Food Saf.* **37**, e12314 (2017).

174. Koebnik, R., Locher, K. P. & Van Gelder, P. Structure and function of bacterial outer membrane proteins: Barrels in a nutshell. *Molecular Microbiology* **37**, 239–253 (2000).

175. Bos, M. P. & Tommassen, J. Biogenesis of the Gram-negative bacterial outer membrane. *Current Opinion in Microbiology* **7**, 610–616 (2004).

176. De Keyzer, J., Van Der Does, C. & Driessen, A. J. M. The bacterial translocase: A dynamic protein channel complex. *Cellular and Molecular Life Sciences* **60**, 2034–2052 (2003).

177. Eppens, E. F., Nouwen, N. & Tommassen, J. Folding of a bacterial outer membrane protein during passage through the periplasm. *EMBO J.* **16**, 4295–4301 (1997).

178. Noinaj, N. *et al.* Structural insight into the biogenesis of β-barrel membrane proteins. *Nature* **501**, 385–390 (2013).

179. Iadanza, M. G. *et al.* Lateral opening in the intact β-barrel assembly machinery captured by cryo-EM. *Nat. Commun.* **7**, 1–12 (2016).

180. Han, L. *et al.* Structure of the BAM complex and its implications for biogenesis of outer-membrane proteins. *Nat. Struct. Mol. Biol.* **23**, 192–196 (2016).

181. Sankaran, K. & Wus, H. C. Lipid Modification of Bacterial Prolipoprotein: Transfer of Diacylglyceryl Moiety from Phosphatidylglycerol. *J. Biol. Chem.* **269**, 19701–19706 (1994).

182. Devs, I. K. & Ray, P. H. Rapid Assay and Purification of a Unique Signal Peptidase that Processes the Prolipoprotein from Escherichia coli B*. *J. Biol. Chem.* **8**, 11114–11120 (1984).

183. Gupta S D, Gan K, Schmid M B & Wu H C. Characterization of a temperature-sensitive mutant of Salmonella typhimurium defective in apolipoprotein N-acyltransferase. *J. Biol. Chem.* **268**, 16551–16556 (1993).

184. Terada, M., Kuroda, T., Matsuyama, S. & Tokuda, H. Lipoprotein Sorting Signals Evaluated as the LolA-dependent Release of Lipoproteins from the Cytoplasmic Membrane of *Escherichia coli*. *J. Biol. Chem.* **276**, 47690–47694 (2001).

185. Ito, Y., Kanamaru, K., Taniguchi, N., Miyamoto, S. & Tokuda, H. A novel ligand bound ABC transporter, LolCDE, provides insights into the molecular mechanisms underlying membrane detachment of bacterial lipoproteins. *Mol. Microbiol.* **62**, 1064–1075 (2006).

186. Missiakas, D., Betton, J. M. & Raina, S. New components of protein folding in extracytoplasmic compartments of escherichia coli SurA, FkpA and Skp/OmpH. *Mol. Microbiol.* **21**, 871–884 (1996).

187. Cronan, J. E. Bacterial Membrane Lipids: Where Do We Stand? *Annual Review of Microbiology* **57**, 203–224 (2003).

188. Huijbregts, R. P. H., De Kroon, A. I. P. M. & De Kruijff, B. Topology and transport of membrane lipids in bacteria. *Biochimica et Biophysica Acta - Reviews on Biomembranes* **1469**, 43–61 (2000).

189. May, K. L. & Silhavy, T. J. Making a membrane on the other side of the wall. *Biochimica et Biophysica Acta - Molecular and Cell Biology of Lipids* **1862**, 1386–1393 (2017).

190. Kol, M. A., Van Dalen, A., De Kroon, A. I. P. M. & De Kruijff, B. Translocation of Phospholipids Is Facilitated by a Subset of Membrane-spanning Proteins of the Bacterial Cytoplasmic Membrane. *J. Biol. Chem.* **278**, 24586–24593 (2003).

191. Dalebroux, Z. D. *et al.* Delivery of cardiolipins to the salmonella outer membrane is necessary for survival within host tissues and virulence. *Cell Host Microbe* **17**, 441–451 (2015).

192. Shrivastava, R., Jiang, X. & Chng, S. S. Outer membrane lipid homeostasis via retrograde phospholipid transport in Escherichia coli. *Mol. Microbiol.* **106**, 395–408 (2017).

193. Shrivastava, R. & Chng, S. S. Lipid trafficking across the Gram-negative cell envelope. *J. Biol. Chem.* **294**, 14175–14184 (2019).

194. Powers, M. J. & Stephen Trent, M. Intermembrane transport: Glycerophospholipid homeostasis of the Gram-negative cell envelope. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 17147–17155 (2019).

195. Okuda, S., Sherman, D. J., Silhavy, T. J., Ruiz, N. & Kahne, D. Lipopolysaccharide transport and assembly at the outer membrane: The PEZ model. *Nature Reviews Microbiology* **14**, 337–345 (2016).

196. Sherman, D. J. *et al.* Lipopolysaccharide is transported to the cell surface by a membrane-Tomembrane protein bridge. *Science (80-.).* **359**, 798–801 (2018).

197. Raetz, C. R. H. & Whitfield, C. Lipopolysaccharide endotoxins. *Annu. Rev. Biochem.* **71**, 635–700 (2002).

198. Whitfield, C. & Trent, M. S. Biosynthesis and Export of Bacterial Lipopolysaccharides. *Annu. Rev. Biochem.* **83**, 99–128 (2014).

199. Qian, J., Garrett, T. A. & Raetz, C. R. H. *In Vitro* Assembly of the Outer Core of the Lipopolysaccharide from *Escherichia coli* K-12 and *Salmonella typhimurium*. *Biochemistry* **53**, 1250–1262 (2014).

200. Amor, K. *et al.* Distribution of core oligosaccharide types in lipopolysaccharides from Escherichia coli. *Infect. Immun.* **68**, 1116–24 (2000).

201. Kaniuk, N. A., Monteiro, M. A., Parker, C. T. & Whitfield, C. Molecular diversity of the genetic loci responsible for lipopolysaccharide core oligosaccharide assembly within the genus Salmonella. *Mol. Microbiol.* **46**, 1305–1318 (2002).

202. Liebau, J., Pettersson, P., Szpryngiel, S. & Mäler, L. Membrane Interaction of the Glycosyltransferase WaaG. *Biophys. J.* **109**, 552–563 (2015).

203. Hoare, A. *et al.* The outer core lipopolysaccharide of Salmonella enterica serovar Typhi is required for bacterial entry into epithelial cells. *Infect. Immun.* **74**, 1555–1564 (2006).

204. Theriot, C. M., Bowman, A. A. & Young, V. B. Antibiotic-Induced Alterations of the Gut Microbiota Alter Secondary Bile Acid Production and Allow for Clostridium difficile Spore Germination and Outgrowth in the Large Intestine. *mSphere* **1**, (2016).

205. Kim, S., Covington, A. & Pamer, E. G. The intestinal microbiota: Antibiotics, colonization resistance, and enteric pathogens. *Immunological Reviews* **279**, 90–105 (2017).

206. Solomon, S. L. & Oliver, K. B. Antibiotic Resistance Threats in the United States: Stepping Back from the Brink. *Am Fam Physician* **89**, 938–941 (2014).

207. Shane, A. L. *et al.* 2017 Infectious Diseases Society of America Clinical Practice Guidelines for the Diagnosis and Management of Infectious Diarrhea. *Clinical Infectious Diseases* **65**, e45–e80 (2017).

208. Troeger, C. *et al.* Estimates of the global, regional, and national morbidity, mortality, and aetiologies of diarrhoea in 195 countries: a

systematic analysis for the Global Burden of Disease Study 2016. *Lancet Infect. Dis.* **18**, 1211–1228 (2018).

209. Woodworth, M. H., Hayden, M. K., Young, V. B. & Kwon, J. H. The Role of Fecal Microbiota Transplantation in Reducing Intestinal Colonization with Antibiotic-Resistant Organisms: The Current Landscape and Future Directions. *Open Forum Infectious Diseases* **6**, (2019).

210. Stalenhoef, J. E. *et al.* Fecal microbiota transfer for multidrug-resistant gram-negatives: A clinical success combined with microbiological failure. *Open Forum Infect. Dis.* **4**, (2017).

211. Theuretzbacher, U., Outterson, K., Engel, A. & Karlén, A. The global preclinical antibacterial pipeline. *Nat. Rev. Microbiol.* **18**, 275–282 (2020).

212. Oh, P. L. *et al.* Diversification of the gut symbiont lactobacillus reuteri as a result of host-driven evolution. *ISME J.* **4**, 377–387 (2010).

213. Walter, J., Britton, R. A. & Roos, S. Host-microbial symbiosis in the vertebrate gastrointestinal tract and the Lactobacillus reuteri paradigm. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 4645–4652 (2011).

214. Kechagia, M. *et al.* Health Benefits of Probiotics: A Review. *ISRN Nutr.* **2013**, 1–7 (2013).

215. Giraffa, G., Chanishvili, N. & Widyastuti, Y. Importance of lactobacilli in food and feed biotechnology. *Res. Microbiol.* **161**, 480–487 (2010).

216. Greifová, G. *et al.* Analysis of antimicrobial and immunomodulatory substances produced by heterofermentative Lactobacillus reuteri. *Folia Microbiol. (Praha).* **62**, 515–524 (2017).

217. Lebeer, S., Vanderleyden, J. & De Keersmaecker, S. C. J. Genes and molecules of lactobacilli supporting probiotic action. *Microbiol. Mol. Biol. Rev.* **72**, 728–64, Table of Contents (2008).

218. Abhisingha, M., Dumnil, J. & Pitaksutheepong, C. Selection of Potential Probiotic Lactobacillus with Inhibitory Activity Against Salmonella and Fecal Coliform Bacteria. *Probiotics Antimicrob. Proteins* **10**, 218–227 (2018).

219. Thomas, C. M. *et al.* Histamine derived from probiotic lactobacillus reuteri suppresses tnf via modulation of pka and erk signaling. *PLoS One* **7**, (2012).

220. Gao, C. *et al.* Histamine H2 receptor-mediated suppression of intestinal inflammation by probiotic lactobacillus reuteri. *MBio* **6**, (2015).

221. Chen, X. Y., Woodward, A., Zijlstra, R. T. & Gänzlea, M. G. Exopolysaccharides synthesized by Lactobacillus reuteri protect against enterotoxigenic Escherichia coli in piglets. *Appl. Environ. Microbiol.* **80**, 5752–5760 (2014).

222. Mu, Q., Kirby, J., Reilly, C. M. & Luo, X. M. Leaky gut as a danger signal for autoimmune diseases. *Frontiers in Immunology* **8**, (2017).

223. Rosenfeldt, V., Benfeldt, E., Valerius, N. H., Pærregaard, A. & Michaelsen, K. F. Effect of probiotics on gastrointestinal symptoms and small intestinal permeability in children with atopic dermatitis. *J. Pediatr.* **145**, 612–616 (2004).

224. Mu, Q., Tavella, V. J. & Luo, X. M. Role of Lactobacillus reuteri in human health and diseases. *Frontiers in Microbiology* **9**, (2018).

225. Salas-Jara, M., Ilabaca, A., Vega, M. & García, A. Biofilm Forming Lactobacillus: New Challenges for the Development of Probiotics. *Microorganisms* **4**, 35 (2016).

226. Roos, S. & Jonsson, H. A high-molecular-mass cell-surface protein from Lactobacillus reuteri 1063 adheres to mucus components. *Microbiology* **148**, 433–442 (2002).

227. Gunning, A. P. *et al.* Use of atomic force microscopy to study the multi-modular interaction of bacterial adhesins to mucins. *Int. J. Mol. Sci.* **17**, (2016).

228. Wegmann, U. *et al.* The pan-genome of Lactobacillus reuteri strains originating from the pig gastrointestinal tract. *BMC Genomics* **16**, (2015).

229. Duar, R. M. *et al.* Experimental evaluation of host adaptation of Lactobacillus reuteri to different vertebrate species. *Appl. Environ. Microbiol.* **83**, (2017).

230. Frese, S. A. *et al.* Molecular Characterization of Host-Specific Biofilm Formation in a Vertebrate Gut Symbiont. *PLoS Genet.* **9**, (2013).

231. Sequeira, S. *et al.* Structural basis for the role of serine-rich repeat proteins from Lactobacillus reuteri in gut microbe-host interactions. *Proc. Natl. Acad. Sci. U. S. A.* **115**, E2706–E2715 (2018).

232. Latousakis, D. *et al.* Serine-rich repeat protein adhesins from Lactobacillus reuteri display strain specific glycosylation profiles. *Glycobiology* **29**, 45–58 (2019).

233. Patenaude, S. I. *et al.* The structural basis for specificity in human abo(h) blood group biosynthesis. *Nat. Struct. Biol.* **9**, 685–690 (2002).

234. Williams, G. J., Zhang, C. & Thorson, J. S. Expanding the promiscuity of a natural-product glycosyltransferase by directed evolution. *Nat. Chem. Biol.* **3**, 657–662 (2007).

235. Gloster, T. M. Advances in understanding glycosyltransferases from a structural perspective. *Curr. Opin. Struct. Biol.* **28**, 131–141 (2014).

236. Liang, D.-M. *et al.* Glycosyltransferases: mechanisms and applications in natural product development. *Chem. Soc. Rev.* **44**, 8350–8374 (2015).

237. Ni, L. *et al.* Cytidine 5′-monophosphate (CMP)-induced structural

changes in a multifunctional sialyltransferase from Pasteurella multocida. *Biochemistry* **45**, 2139–2148 (2006).

238. Chen, C. I. *et al.* Structure of human POFUT2: Insights into thrombospondin type 1 repeat fold and O-fucosylation. *EMBO J.* **31**, 3183–3197 (2012).

239. Larivière, L., Gueguen-Chaignon, V. & Moréra, S. Crystal structures of the T4 phage β-glucosyltransferase and the D100A mutant in complex with UDP-glucose: Glucose binding and identification of the catalytic base for a direct displacement mechanism. *J. Mol. Biol.* **330**, 1077–1086 (2003).

240. Read, R. J. *et al.* A new generation of crystallographic validation tools for the Protein Data Bank. *Structure* **19**, 1395–1412 (2011).

241. Tan, Y. Z. *et al.* Sub-2 Å Ewald curvature corrected structure of an AAV2 capsid variant. *Nat. Commun.* **9**, (2018).

242. Weichenberger, C. X., Afonine, P. V., Kantardjieff, K. & Rupp, B. The solvent component of macromolecular crystals. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **71**, 1023–1038 (2015).

243. OPPF-UK: Research. Available at: https://www.oppf.rc-harwell.ac.uk/OPPF/protocols/cryoprotection.jsp. (Accessed: 11th November 2019)

244. Winter, G. & McAuley, K. E. Automated data collection for macromolecular crystallography. *Methods* **55**, 81–93 (2011).

245. Waterman, D. G. *et al.* Diffraction-geometry refinement in the DIALS framework. *Acta Crystallogr. Sect. D Struct. Biol.* **72**, 558–575 (2016).

246. Winter, G. Xia2: An expert system for macromolecular crystallography data reduction. *J. Appl. Crystallogr.* **43**, 186–190 (2010).

247. Vonrhein, C. *et al.* Data processing and analysis with the autoPROC toolbox. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **67**, 293–302 (2011).

248. Bateman, A. UniProt: A worldwide hub of protein knowledge. *Nucleic Acids Res.* **47**, D506–D515 (2019).

249. Winn, M. D. *et al.* Overview of the CCP4 suite and current developments. *Acta Crystallogr. D. Biol. Crystallogr.* **67**, 235–42 (2011).

250. Chen, Y. W. Solution solution: Using NMR models for molecular replacement. *Acta Crystallogr. - Sect. D Biol. Crystallogr.* **57**, 1457–1461 (2001).

251. Stein, N. CHAINSAW: A program for mutating pdb files used as templates in molecular replacement. *J. Appl. Crystallogr.* **41**, 641–643 (2008).

252. Bunkóczi, G. & Read, R. J. Improvement of molecular-replacement

models with Sculptor. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **67**, 303–312 (2011).

253. Matthews, B. W. Solvent content of protein crystals. *J. Mol. Biol.* **33**, 491–497 (1968).

254. Virtanen, J. J. & Zhang, Y. MR-REX: Molecular replacement by cooperative conformational search and occupancy optimization on low-accuracy protein models. *Acta Crystallogr. Sect. D Struct. Biol.* **74**, 606–620 (2018).

255. Vagin, A. & Teplyakov, A. Molecular replacement with MOLREP. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **66**, 22–25 (2010).

256. McCoy, A. J. Solving structures of protein complexes by molecular replacement with Phaser. in *Acta Crystallographica Section D: Biological Crystallography* **63**, 32–41 (Acta Crystallogr D Biol Crystallogr, 2006).

257. McCoy, A. J., Grosse-Kunstleve, R. W., Storoni, L. C. & Read, R. J. Likelihood-enhanced fast translation functions. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **61**, 458–464 (2005).

258. Storoni, L. C., McCoy, A. J. & Read, R. J. Likelihood-enhanced fast rotation functions. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **60**, 432–438 (2004).

259. Evans, P. & McCoy, A. An introduction to molecular replacement. in *Acta Crystallographica Section D: Biological Crystallography* **64**, 1–10 (International Union of Crystallography, 2007).

260. Emsley, P. & Cowtan, K. Coot: Model-building tools for molecular graphics. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **60**, 2126–2132 (2004).

261. Liebschner, D. *et al.* Macromolecular structure determination using X-rays, neutrons and electrons: recent developments in Phenix . *Acta Crystallogr. Sect. D Struct. Biol.* **75**, 861–877 (2019).

262. Evans, P. R. An introduction to data reduction: Space-group determination, scaling and intensity statistics. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **67**, 282–292 (2011).

263. Diederichs, K. & Karplus, P. A. Improved R-factors for diffraction data analysis in macromolecular crystallography. *Nat. Struct. Biol.* **4**, 269–275 (1997).

264. Weiss, M. S. Global indicators of X-ray data quality. *J. Appl. Crystallogr.* **34**, 130–135 (2001).

265. Brünger, A. T. Free R value: A novel statistical quantity for assessing the accuracy of crystal structures. *Nature* **355**, 472–475 (1992).

266. Karplus, P. A. & Diederichs, K. Linking crystallographic model and data quality. *Science (80-.).* **336**, 1030–1033 (2012).

267. Hintze, B. J., Lewis, S. M., Richardson, J. S. & Richardson, D. C. Molprobity's ultimate rotamer-library distributions for model

validation. *Proteins Struct. Funct. Bioinforma.* **84**, 1177–1189 (2016).

268. Gore, S. *et al.* Validation of Structures in the Protein Data Bank In Brief. *Struct. Des.* **25**, 1916–1927 (2017).

269. Lairson, L. L., Henrissat, B., Davies, G. J. & Withers, S. G. Glycosyltransferases: Structures, Functions, and Mechanisms. *Annu. Rev. Biochem.* **77**, 521–555 (2008).

270. Lizak, C., Gerber, S., Numao, S., Aebi, M. & Locher, K. P. X-ray structure of a bacterial oligosaccharyltransferase. *Nature* **474**, 350–356 (2011).

271. Charnock, S. J. & Davies, G. J. Structure of the nucleotide-diphospho-sugar transferase, SpsA from Bacillus subtilis, in native and nucleotide-complexed forms. *Biochemistry* **38**, 6380–6385 (1999).

272. Vrielink, A., Rüger, W., Driessen, H. P. & Freemont, P. S. Crystal structure of the DNA modifying enzyme beta-glucosyltransferase in the presence and absence of the substrate uridine diphosphoglucose. *EMBO J.* **13**, 3413–3422 (1994).

273. Flint, J. *et al.* Structural dissection and high-throughput screening of mannosylglycerate synthase. *Nat. Struct. Mol. Biol.* **12**, 608–614 (2005).

274. Zhu, F. *et al.* Structural and functional analysis of a new subfamily of glycosyltransferases required for glycosylation of serine-rich streptococcal adhesins. *J. Biol. Chem.* **286**, 27048–27057 (2011).

275. Moremen, K. W. & Haltiwanger, R. S. Emerging structural insights into glycosyltransferase-mediated synthesis of glycans. *Nature Chemical Biology* **15**, 853–864 (2019).

276. Vuong, T. V. & Wilson, D. B. Glycoside hydrolases: Catalytic base/nucleophile diversity. *Biotechnol. Bioeng.* **107**, 195–205 (2010).

277. Davies, G. J. & Sinnott, M. L. Sorting the diverse: The sequencebased classifications of carbohydrateactive enzymes. *Biochem. (Lond).* **30**, 26–32 (2008).

278. Sanz-Aparicio, J., Hermoso, J. A., Martínez-Ripoll, M., Lequerica, J. L. & Polaina, J. Crystal structure of β-glucosidase A from Bacillus polymyxa: Insights into the catalytic activity in family 1 glycosyl hydrolases. *J. Mol. Biol.* **275**, 491–502 (1998).

279. Gibson, R. P. *et al.* Molecular Basis for Trehalase Inhibition Revealed by the Structure of Trehalase in Complex with Potent Inhibitors. *Angew. Chemie Int. Ed.* **46**, 4115–4119 (2007).

280. Bartlett, G. J., Porter, C. T., Borkakoti, N. & Thornton, J. M. Analysis of catalytic residues in enzyme active sites. *J. Mol. Biol.* **324**, 105–121 (2002).

281. Guce, A. I. *et al.* Catalytic mechanism of human α-galactosidase. *J.*

Biol. Chem. **285**, 3625–3632 (2010).

282. Jordan, D. B., Li, X. L., Dunlap, C. A., Whitehead, T. R. & Cotta, M. A. Structure-function relationships of a catalytically efficient β-D-xylosidase. *Appl. Biochem. Biotechnol.* **141**, 51–76 (2007).

283. Karaveg, K. *et al.* Mechanism of class 1 (glycosylhydrolase family 47) α-mannosidases involved in N-glycan processing and endoplasmic reticulum quality control. *J. Biol. Chem.* **280**, 16197–16207 (2005).

284. Hinou, H., Kurogochi, M., Shimizu, H. & Nishimura, S. I. Characterization of Vibrio cholerae neuraminidase by a novel mechanism-based fluorescent labeling reagent. *Biochemistry* **44**, 11669–11675 (2005).

285. Chavas, L. M. G. *et al.* Crystal structure of the human cytosolic sialidase Neu2: Evidence for the dynamic nature of substrate recognition. *J. Biol. Chem.* **280**, 469–475 (2005).

286. André, G. *et al.* Computational and experimental studies of the catalytic mechanism of Thermobifida fusca cellulase Cel6A (E2). *Protein Eng.* **16**, 125–134 (2003).

287. Kapust, R. B. & Waugh, D. S. *Escherichia coli* maltose-binding protein is uncommonly effective at promoting the solubility of polypeptides to which it is fused. *Protein Sci.* **8**, 1668–1674 (1999).

288. Jones, D. T. Protein Secondary Structure Prediction Based on Position-specific Scoring Matrices. *J. Mol. Biol.* **272**, 195–202 (1999).

289. Cuff, J. A. & Barton, G. J. Application of multiple sequence alignment profiles to improve protein secondary structure prediction. *Proteins Struct. Funct. Genet.* **40**, 502–511 (2000).

290. Rost, B. Review: Protein Secondary Structure Prediction Continues to Rise. *J. Struct. Biol.* **134**, 204–218 (2001).

291. Lupas, A., Dyke, M. Van & Stock, J. Predicting coiled coils from protein sequences. *Science (80-.).* **252**, 1162–1164 (1991).

292. Tusnády, G. E. & Simon, I. Principles governing amino acid composition of integral membrane proteins: application to topology prediction. *J. Mol. Biol.* **283**, 489–506 (1998).

293. Nugent, T. & Jones, D. T. Transmembrane protein topology prediction using support vector machines. *BMC Bioinformatics* **10**, 159 (2009).

294. Käll, L., Krogh, A. & Sonnhammer, E. L. . A Combined Transmembrane Topology and Signal Peptide Prediction Method. *J. Mol. Biol.* **338**, 1027–1036 (2004).

295. Hiller, K., Grote, A., Scheer, M., Munch, R. & Jahn, D. PrediSi: prediction of signal peptides and their cleavage positions. *Nucleic Acids Res.* **32**, W375–W379 (2004).

296. Ward, J. J., Sodhi, J. S., McGuffin, L. J., Buxton, B. F. & Jones, D. T. Prediction and Functional Analysis of Native Disorder in Proteins from

the Three Kingdoms of Life. *J. Mol. Biol.* **337**, 635–645 (2004).

297. Dosztányi, Z., Csizmók, V., Tompa, P. & Simon, I. The Pairwise Energy Content Estimated from Amino Acid Composition Discriminates between Folded and Intrinsically Unstructured Proteins. *J. Mol. Biol.* **347**, 827–839 (2005).

298. Shevchenko, A., Tomas, H., Havlis, J., Olsen, J. V & Mann, M. In-gel digestion for mass spectrometric characterization of proteins and proteomes. *Nat. Protoc.* **1**, 2856–60 (2006).

299. Perkins, D. N., Pappin, D. J. C., Creasy, D. M. & Cottrell, J. S. Probability-based protein identification by searching sequence databases using mass spectrometry data. in *Electrophoresis* **20**, 3551–3567 (Wiley-VCH Verlag, 1999).

300. Reece-Hoyes, J. S. & Walhout, A. J. M. Gateway recombinational cloning. *Cold Spring Harbor Protocols* **2018**, 1–6 (2018).

301. Li, M. Z. & Elledge, S. J. Harnessing homologous recombination in vitro to generate recombinant DNA via SLIC. *Nat. Methods* **4**, 251–256 (2007).

302. Martinez-Fleites, C. *et al.* Insights into the Synthesis of Lipopolysaccharide and Antibiotics through the Structures of Two Retaining Glycosyltransferases from Family GT4. *Chem. Biol.* **13**, 1143–1152 (2006).

303. Liu, H. & Naismith, J. H. An efficient one-step site-directed deletion, insertion, single and multiple-site plasmid mutagenesis protocol. *BMC Biotechnol.* **8**, 91 (2008).

304. Zhang, Z. Structural and functional studies on Gram-negative bacteria lipid transport and modification proteins. (Univiersity of East Anglia, 2017).

305. Guzman, L.-M., Belin, D., Carson, M. J. & Beckwith, J. Tight Regulation, Modulation, and High-Level Expression by Vectors Containing the Arabinose P BAD Promoter. *J. Bacteriol.* **177**, 4121–4130 (1995).

306. Leibly, D. J. *et al.* Stabilizing Additives Added during Cell Lysis Aid in the Solubilization of Recombinant Proteins. *PLoS One* **7**, e52482 (2012).

307. Singh, A., Upadhyay, V., Upadhyay, A. K., Singh, S. M. & Panda, A. K. Protein recovery from inclusion bodies of Escherichia coli using mild solubilization process. *Microbial Cell Factories* **14**, (2015).

308. Wierenga, R. K. The TIM-barrel fold: A versatile framework for efficient enzymes. *FEBS Letters* **492**, 193–198 (2001).

309. Mancia, F. & Evans, P. R. Conformational changes on substrate binding to methylmalonyl CoA mutase and new insights into the free radical mechanism. *Structure* **6**, 711–720 (1998).

310. McCarter, J. D. & Stephen Withers, G. Mechanisms of enzymatic glycoside hydrolysis. *Curr. Opin. Struct. Biol.* **4**, 885–892 (1994).

311. Adamson, C. *et al.* Structural Snapshots for Mechanism-Based Inactivation of a Glycoside Hydrolase by Cyclopropyl Carbasugars. *Angew. Chemie - Int. Ed.* **55**, 14978–14982 (2016).

312. Tamura, J. I., Hasegawa, K., Kadowaki, K., Igarashi, Y. & Kodama, T. Molecular cloning and sequence analysis of the gene encoding an endo α-1,4 polygalactosaminidase of Pseudomonas sp. 881. *J. Ferment. Bioeng.* **80**, 305–310 (1995).

313. Holm, L. & Sander, C. Dali: a network tool for protein structure comparison. *Trends Biochem. Sci.* **20**, 478–480 (1995).

314. Holm, L., Kääriäinen, S., Rosenström, P. & Schenkel, A. Searching protein structure databases with DaliLite v.3. *Bioinformatics* **24**, 2780–2781 (2008).

315. Schwarzenbacher, R. *et al.* Crystal structure of a putative glutamine amido transferase (TM1158) from Thermotoga maritima at 1.7 Å resolution. *Proteins Struct. Funct. Bioinforma.* **54**, 801–805 (2004).

316. Hidaka, M. *et al.* The crystal structure of galacto-N-biose/lacto-N-biose I phosphorylase. A large deformation of a tim barrel scaffold. *J. Biol. Chem.* **284**, 7273–7283 (2009).

317. Hidaka, M. *et al.* Trimeric crystal structure of the glycoside hydrolase family 42 β-galactosidase from Thermus thermophilus A4 and the structure of its complex with galactose. *J. Mol. Biol.* **322**, 79–91 (2002).

318. Solomon, H. V. *et al.* Crystallization and preliminary crystallographic analysis of GanB, a GH42 intracellular β-galactosidase from Geobacillus stearothermophilus. *Acta Crystallogr. Sect. F Struct. Biol. Cryst. Commun.* **69**, 1114–1119 (2013).

319. Waugh, D. S. An overview of enzymatic reagents for the removal of affinity tags. *Protein Expression and Purification* **80**, 283–293 (2011).

320. Royer, C. J. & Cook, P. D. A structural and functional analysis of the glycosyltransferase BshA from Staphylococcus aureus: Insights into the reaction mechanism and regulation of bacillithiol production. *Protein Sci.* **28**, 1083–1094 (2019).

321. Wang, X.-P. *et al.* Structural and enzymatic analyses of a glucosyltransferase Alr3699/HepE involved in Anabaena heterocyst envelop polysaccharide biosynthesis. *Glycobiology* **26**, 520–31 (2016).

322. Ramírez, A. S. *et al.* Structural basis of the molecular ruler mechanism of a bacterial glycosyltransferase. *Nat. Commun.* **9**, 445 (2018).

323. Schmidt, H. *et al.* Structural and mechanistic analysis of the membrane-embedded glycosyltransferase WaaA required for lipopolysaccharide synthesis. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 6253–6258 (2012).

324. Qi, G., Lee, R. & Hayward, S. A comprehensive and non-redundant database of protein domain movements. *Bioinformatics* **21**, 2832–2838 (2005).

325. Zhu, F., Zhang, H. & Wu, H. A conserved domain is crucial for acceptor substrate binding in a family of glucosyltransferases. *J. Bacteriol.* **197**, 510–7 (2015).

326. Krissinel, E. & Henrick, K. Inference of Macromolecular Assemblies from Crystalline State. *J. Mol. Biol.* **372**, 774–797 (2007).

327. Le Mauff, F. *et al.* Molecular mechanism of Aspergillus fumigatus biofilm disruption by fungal and bacterial glycoside hydrolases. *J. Biol. Chem.* **294**, 10760–10772 (2019).

328. Snarr, B. D. *et al.* Microbial glycoside hydrolases as antibiofilm agents with cross-kingdom activity. *Proc. Natl. Acad. Sci. U. S. A.* **114**, 7124–7129 (2017).

329. Wernimont, A. & Edwards, A. In Situ Proteolysis to Generate Crystals for Structure Determination: An Update. *PLoS One* **4**, e5094 (2009).

330. Dong, A. *et al.* In situ proteolysis for protein crystallization and structure determination. *Nat. Methods* **4**, 1019–1021 (2007).

331. Identification of associated proteins by coimmunoprecipitation. *Nat. Methods* **2**, 475–476 (2005).

332. Rajarathnam, K. & Rösgen, J. Isothermal titration calorimetry of membrane proteins - Progress and challenges. *Biochimica et Biophysica Acta - Biomembranes* **1838**, 69–77 (2014).

333. Patching, S. G. Surface plasmon resonance spectroscopy for characterisation of membrane protein-ligand interactions and its potential for drug discovery. *Biochimica et Biophysica Acta - Biomembranes* **1838**, 43–55 (2014).

334. Lee, C. H., Chou, C. C., Hsu, M. F. & Wang, A. H. J. Determining the N-terminal orientations of recombinant transmembrane proteins in the Escherichia coli plasma membrane. *Sci. Rep.* **5**, 1–11 (2015).

335. Huang, K. xue, Badger, M., Haney, K. & Evans, S. L. Large scale production of Bacillus thuringiensis PS149B1 insecticidal proteins Cry34Ab1 and Cry35Ab1 from Pseudomonas fluorescens. *Protein Expr. Purif.* **53**, 325–330 (2007).

336. Retallack, D. M., Jin, H. & Chew, L. Reliable protein production in a Pseudomonas fluorescens expression system. *Protein Expression and Purification* **81**, 157–165 (2012).

337. Dong, C., Kotzsch, A., Dorward, M., van Pée, K.-H. & Naismith, J. H. Crystallization and X-ray diffraction of a halogenating enzyme, tryptophan 7-halogenase, from *Pseudomonas fluorescens*. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **60**, 1438–1440 (2004).

338. Zhang, Z. *et al.* High-level production of membrane proteins in E. coli BL21(DE3) by omitting the inducer IPTG. *Microb. Cell Fact.* **14**, 142 (2015).

339. Muheim, C. *et al.* Identification of a fragment-based scaffold that inhibits the glycosyltransferase WaaG from Escherichia Coli. *Antibiotics* **5**, (2016).

340. Spraggon, G. *et al.* The structures of anthranilate synthase of Serratia marcescens crystallized in the presence of (i) its substrates, chorismate and glutamine, and a product glutamate, and (ii) its end-product inhibitor, L-tryptophan. *Proc. Natl. Acad. Sci. U. S. A.* **98**, 6021–6026 (2001).

341. Petrova, T. *et al.* Novel hexamerization motif is discovered in a conserved cytoplasmic protein from Salmonella typhimurium. *J. Struct. Funct. Genomics* **8**, 19–25 (2007).

342. Taschner, M. *et al.* Intraflagellar transport proteins 172, 80, 57, 54, 38, and 20 form a stable tubulin-binding IFT -B2 complex . *EMBO J.* **35**, 773–790 (2016).

343. Viborg, A. H. *et al.* A β1-6/β1-3 galactosidase from <scp> *B* </scp> *ifidobacterium animalis* subsp. *lactis* <scp>B</scp> l-04 gives insight into sub-specificities of β-galactoside catabolism within <scp> *B* </scp> *ifidobacterium*. *Mol. Microbiol.* **94**, 1024–1040 (2014).

344. Morero, N. R. *et al.* HemR is an OmpR/PhoB-like response regulator from *L eptospira* , which simultaneously effects transcriptional activation and repression of key haem metabolism genes. *Mol. Microbiol.* **94**, 340–352 (2014).

345. Herrou, J., Foreman, R., Fiebig, A. & Crosson, S. A structural model of anti-anti-σ inhibition by a two-component receiver domain: The PhyR stress response regulatormmi. *Mol. Microbiol.* **78**, 290–304 (2010).

346. Mechaly, A. E. *et al.* Structural Coupling between Autokinase and Phosphotransferase Reactions in a Bacterial Histidine Kinase. *Structure* **25**, 939-944.e3 (2017).

347. Glanville, D. G. *et al.* RitR is an archetype for a novel family of redox sensors in the streptococci that has evolved from two-component response regulators and is required for pneumococcal colonization. *PLoS Pathog.* **14**, (2018).

348. Thomä, N. H., Evans, P. R. & Leadlay, P. F. Protection of Radical Intermediates at the Active Site of Adenosylcobalamin-Dependent Methylmalonyl-CoA Mutase † , ‡. *Biochemistry* **39**, 9213–9221 (2000).

349. McNamara, D. E. *et al.* Structure of dihydromethanopterin reductase, a cubic protein cage for redox transfer. *J. Biol. Chem.* **289**, 8852–8864 (2014).

Appendix

Table A1: Primers

Ligation Dependent Cloning Primers

Name	Sequence
PA_PelA_F_1-948	ATA TCC ATG GGT ATGCGGTTCAGCAAGAAAGGAATCG
PA_PelA_R_1-948	ATAT AAGCTT TTA GTG GTG GTG GTG ATG ATG ATG ATG GCGGCAGACGAGTTGGCC
PA_PelA_F_288-948	ATA TCC ATG GGT ACCCCGGCGCTGGACTAC
PA_PelA_R_288-948	ATAT AAGCTT TTA GCGGCAGACGAGTTGGCC
PA_PelA_46-948_F	AAGTTA CC ATG GCG GGGCCGTCCAGC
PA_PelA_46-948_R	AACGAT AAGCTT TCA GCGGCAGACGAGTTGGCG
PA_PelB_1-1193_F4	AATTTA CCCGGG ATG GCCAACTCGTCTGCCGCTGA
PA_PelB_1-1193_R4	AGGTTC GTCGAC TTA GTGGTGGTGGTGATGATGATGATGATG AGCACGCGCTTCGGCCGC
PA_PelC_F_1-172	ATA TCC ATG GGT ATGCAATCCATCCGCTGCCT
PA_PelC_R_1-172	ATAT GTC GAC CTA GTG ATG GTG ATG GTG ATG GTG ATG CACTCGAGCCGCAGGTC
PA_PelD_F	ATAT T CTA GAC ATGTCCGCGCACAAGGATTTCAC
PA_PelD_R	ATAT AAG CTT TTA GTG GTG GTG GTG ATG ATG ATG ATG AATAGCCACTTGCTGATCATTCAGGC
PA_PelE_F_1-329	ATA TCC ATG GCT ATGATCAGCAAGTGGCTATTTAGCG
PA_PelE_R_1-329	ATAT AAG CTT TTA GTG GTG GTG GTG ATG ATG ATG ATG TGTCCAGTATCTCGCCAGGG
PA_PelF_F_1-506	ATAT GGA TCC ATGACCGAACACACCGCTCCGA
PA_PelF_R_1-506	ATAT GTC GAC CTA TCATGCAATCTCCGTGGCTTCGC

PA_PelF_F2_1-506	ATA T CC ATG GGT ATGACCGAACACACCGCTCCG
PA_PelG_F	ATA TCC ATG GCCGGCATCGGCTTC
PA_PelG_R	ATAT GTC GAC TTA GTG GTG GTG GTG ATG ATG ATG ATG GCGATTGAGCATGAAGGTCTCGTAC

Gateway Cloning Primers

Name	Sequence
Adapter2_F	GGG GAC AAG TTT GTA CAA AAA AGC AGG CT TCCTGGAAGTTCTGTT
Adapter2_R	GGG GAC CAC TTT GTA CAA GAA AGC TGG GTT
3Cprotease_PelA523_F	CTG GAA GTT CTG TTT CAG GGC CCG ATG GCC ACC GTG CAC ATC GAT GGC
3C_protease_PelA544_F	CTG GAA GTT CTG TTT CAG GGC CCG ATG GCC GGT CAG CAG GTG CTG GAG
3C_protease_PelA507_F	CTG GAA GTT CTG TTT CAG GGC CCG ATG CCG CTG CCG AGT CCC GAC
attB2_PelA920_R	CAA GAA AGC TGG GTT TTA TTT GCC GGC CGC ACT TAG CCG
attB2_PelA948_R	CAA GAA AGC TGG GTT TTA GCG GCA GAC GAG TTG GCC ATC
3C_protease_PelA46_F	CTG GAA GTT CTG TTT CAG GGC CCG GGG CCG TCC AGC GTG G
3C_protease_PelA262_F	CTG GAA GTT CTG TTT CAG GGC CCG ATG CTG CCG CCG GAG CGG CGC
attB2_PelA262_R	CAA GAA AGC TGG GTT TTA CAG GTA GTC GAT GGC GAC GAT
aatB2_PelA303_R	CAA GAA AGC TGG GTT TTA CGG TTG CAC CTC GAC GTC
attB2_PelA507_R	CAA GAA AGC TGG GTT TTA CAC AAG ACG CAA GGC CTT GCG

Sequence and Ligation Independent Cloning (SLIC) Primers

Name	Sequence
SLIC_F_PelB47	TTTTCAGGGC GATGGCGAGCCGGATGC
SLIC_R_PelB880	CAGCCATGGT TTA CTGCGGGGTACGTTCCA

SLIC_R_PelB47_pHisTEV	GCTCGCCATC GCCCTGAAAATACAGGTT
SLIC_F_PelB880_pHisTEV	C CCG CAG TAA ACC ATG GCT GAT ATC GGA
SLIC_F_PelB351	TTTTCAGGGC GCG GAT GAT CCG GCG CTG
SLIC_R_PelB588	CAGCCATGGT TTA CTCCGGCAGACCTTCCGC
SLIC_R_PelB351_pHisTEV	G ATC ATC CGC GCCCTGAAAATACAGGTT
SLIC_F_PelB588_pHisTEV	G CCG GAG TAA ACC ATG GCT GAT ATC GGA
SLIC_F_PelA46_pHisToverhang	TATACATATG GGG CCG TCC AGC GTG GC
SLIC_R_PelA_SDM	TGTGCACGGACCACCCGCGCCGGCCATTCTCGG
SLIC_F PelA_ SDM	CCGAGAATGGCCGGCGCGGGTGGTCCGTGCACA
SLIC_R_PelA946_His6	G ATG GTG ATG GCG GCA GAC GAG TTG GCC ATC
SLIC_R_pHisTEV_PelA46Overhang	T GGA CGG CCC CATATGTATATCTCCTTCTTAAAGTT
SLIC_F_pHisTEV_His6Overhang	CAT CAC CAT CAC CAT CAC TAA GAT TAC GAT

Site Directed Mutagenesis Primers

Name	Sequence
Sen_WaaB_V186A_F	TTCTCTATGCGGGGCGTCTTAAATTTGAAGGGC AGAAAAGAGTTAAAG
Sen_WaaB_V186A_R	GCCCCGCATAGAGAAATACAGCGGGTTTATCGC GC
Sen_WaaB_Q194A _F	GAAGGGGCGAAAAGAGTTAAAGATTTATTTGAT GGCTTAGCTCGTACG
Sen_WaaB_Q194A_R	CTCTTTTCGCCCCTTCAAATTTAAGACGCCCTAC ATAGAGAAATACAG
Sen_WaaB_K195A_F	AGGGCAGGCGAGAGTTAAAGATTTATTTGATGG CTTAGCTCGTACG
Sen_WaaB_K195A _R	CTCTCGCCTGCCCTTCAAATTTAAGACGCCCTAC ATAGAGAAATAC
Sen_WaaB_I216A_F	ATATTGCGGGTGATGGCTCAGATTTTGAAAAGT GCCAGGC

Sen_WaaB_I216A_R	CCATCACCCGCAATATGTAGCTGCCATTCCCCTGTC
Sen_WaaB_W243A_F	ATGGTGCGCAAAGCGCGCCGTGGCAAGTCGTAC
Sen_WaaB_W243A_R	GCTTTGCGCACCATACCAGATCACACGCTGCTCAATACC
Sen_WaaB_E268A_F	GCATTTGCGGGATTTCCTATGACCCTGCTGGAAGCAATG
Sen_WaaB_E268A_R	GAAATCCCGCAAATGCAGAGGTAAGTAGTAACGCGGTGAC
Sen_WaaB_T273A_F	CTATGGCGCTGCTGGAAGCAATGTCATATGGAATTCCGTG
Sen_WaaB_T273A_R	CAGCAGCGCCATAGGAAATCCCTCAAATGCAGAGGTAAG
Sen_WaaB_E276A_F	GCTGGCGGCAATGTCATATGGAATTCCGTGTATTAGTTCTGATTG
Sen_WaaB_E276A_R	CATTGCCGCCAGCAGGGTCATAGGAAATCCCTCAAATG
PLOU3_SEQ_F	GTCGATGAAGCCCTGAAAGAC
PLOU3_SEQ_R	GTTTTCCCAGTCACGACGTTG

Table A2: Optimised crystallisation conditions of PelA

Protein	Concentration (mg/mL)	Well solution	T (°C)
PelA 46-948 (pLou3)	29.5	0.1M Bis-Tris pH 5.5, 0.2M $MgCl_2.6H_2O$, 25% w/v PEG 3350	21
PelA 46-948 (pLou3)	9.5	0.1M Bis-Tris pH 6.5, 0.2M $MgCl_2.6H_2O$, 25% w/v PEG 3350	21
PelA 46-948 (pEHisTEV)	5.0	0.2M Calcium acetate hydrate pH7.5, 20% w/v PEG 3350	21
55 kDa fragment (pEHisTEV)	11.5	0.2M Sodium fluoride pH7.3, 20% w/v PEG 3350	21
55 kDa fragment (pEHisTEV)	11.5	0.2M Ammonium citrate dibasic pH 5.1, 20% w/v PEG 3350	21

55 kDa fragment (pEHisTEV)	11.5	0.2M Sodium phosphate dibasic dehydrate pH 9.1, 20% w/v PEG 3350	21
PelA46-948 (pEHisTEV)	7.0	0.2M Ammonium phosphate monobasic pH 8.0, 20% w/v PEG 3350	16
HisPelA46-948 (pEHisTEV)	10.0	0.2M Potassium formate pH 7.2, 20% w/v PEG 3350	16
PelA55 (pEHisTEV)	10.0	0.1M Bis-Tris Propane pH 6.5, 0.2M sodium fluoride, 20% w/v PEG 3350	16
PelA55 (pEHisTEV)	10.0	0.1M Bis-Tris Propane pH 7.5, 0.2M sodium fluoride, 20% w/v PEG 3350	16
PelA46-948 (pEHisTEV)	10.0	0.1M Tris pH 8.5, ammonium phosphate monobasic, 50% w/v 2-methyl-2,4-pentanediol	16
PelA46-948 (pEHisTEV)	10.0	0.1M Sodium cacodylate pH 6.5, 5% w/v PEG 8000, 40% w/v 2-methyl-2,4-pentanediol	16
PelA55 (pEHisTEV)	10.0	0.2M magnesium acetate tetrahydrate pH 7.5, 20% w/v PEG 3350	16
PelA 46-507	11.5	0.1M Tris pH 8.5, 0.2M Trimethylamine N-oxide, 20% w/v PEG 2000 MME	21
PelA 46-507	10.0	29 % w/v PEG 6000, 0.2 M Sodium citrate 3.5	16

Table A3: PCR Programs

Gateway Cloning: Adapter 1

1.	98°C	3:00 min
2.	98°C	0:20 min
3.	55-72°C	0:30 min
4.	72°C	1:00 min

5.	Go to 2.	27x
6.	72°C	10:00 min
7.	12°C	∞

Gateway Cloning: Adapter 2

1.	98°C	3:00 min
2.	98°C	0:20 min
3.	45°C	0:30 min
4.	72°C	0:45 min
5.	Go to 2.	5x
6.	98°C	0:20 min
7.	55°C	0:30 min
8.	72°C	0:45 min
9.	Go to 6.	15x
10.	72°C	10:00
11.	12°C	∞

Colony PCR

1.	95°C	2:00 mins
2.	95°C	0:30 mins
3.	50°C	0:30 mins
4.	72°C	1:30 mins
5.	Go to 2.	28x
6.	72°C	5:00 mins
7.	12°C	∞

Site Directed Mutagenesis

1.	94°C	7:00 min
2.	94°C	1:00 min
3.	66°C	1:00 min
4.	72°C	10:00 min
5.	Go to 2.	12x
6.	95°C	1:00 min
7.	61°C	1:00 min

8.	72°C	30:30 min
9.	12°C	∞

SLIC 1

1.	22°C	35:00 min
2.	70°C	10:00 min

SLIC 2

1.	75°C	10:00 min
2.	75°C	00:08 min
3.	Go to 2.	549x, decrease temperature by 0.01°C per repeat

Table A4: DALI server results for PelA 300 - 507

PDB	Z-Score	ID (%)	Function
4jqs-A	11.9	10	Hypothetical thua-like protein, *Bacteroides*, no publication.
1o1y-A	11.7	13	Glutamine amidotransferase, *Thermotoga*, transfers amide nitrogen of glutamine, C88 H173 E175 active site, α-β-α flavodoxin-like fold[315]
1i7q-b	11.1	11	TrpG of Anthranilate synthase, *Serratia*, glutamine hydrolysis in tryptophan biosynthesis, Class I glutamine amidotransferase domain, catalytic Cyc85, His170, Glu172[340]
2gk3-E	11.1	13	Stm3548, *Salmonella*, hexamer[341]
5fmr-C	10.7	13	IFT52N*, Chlamydomonas*, degenerate β-galactosidase/glutaminase domain, K130/R204 contribute to complex formation, K64/66/69 to tubulin binding[342]
3rht-A	10.7	13	Glutamine amidotransferase, *Planctomyces*, unpublished
1q7r-A	10.3	10	Predicted amidotransferase, *Geobacillus*, no publication

1t0b-G	10.1	14	ThuA-like protein, *Bacillus*, unpublished
4uni-A	10.1	12	B-galactosidase, *Bifidobacterium*, equivalent domain B contribute to trimerisation and stabilising the adjacent $(\beta/\alpha)_8$ TIM barrel of domain A[343]
4oif-a	10.1	13	GanB, GH42 β-galactosidase, *Geobacillus*, no proposed function for domain[318]
4q7e-A	6.6	20	HemR, *Leptospira*, two component signalling response regulator. $(\beta\alpha)_5$ fold with α4-β5-α5 forming a dimerization interface with DNA-binding winged helix-turn-helix motif[344]
3n0r-A	6.2	21	PhyR, *Caulobacter*, dimeric anti-σ factor binding protein activated by phosphorylation[345]
4uhj-C	5.8	22	CpxR, *E.coli*, response regulator transcription factor component of histidine kinase sensor[346]
5u8k-B	5.5	22	RitR, *Streptococcus*, dimeric two component transcription factor response regulator[347].
1e1c-A	5.2	20	Methylmalonyl-CoA mutase, *Propionibacterium*, catalysis of interconversion of 2-methylmalonyl-CoA and succinyl-CoA, similar area is involved in binging cobalamin (coenzyme) binding[348]
1ys6-B	5.2	20	PrrA, *Mycobacterium*, transcriptional response regulator (no paper)
3wis-A	5.1	20	DmrB, *Burkholderia*, 24-mer composed of homotrimers, binds and reduces a dihydromethanopterin ligand at each momonmer-monomer interface[349]

Table A5: Optimised crystallisation conditions of $GtfC_{100-23}$

Concentration (mg/mL)	Well solution	T (°C)	Crystal Form	Cryoprotectant of best diffracting crystal
7.5 - 15	0.1M Bis-Tris pH 7.5, 0.2M Potassium thiocyanate,	16	Plates	30% v/v glycerol

	20% w/v PEG 3350			
7.5 - 15	0.1M Sodium Acetate pH 4.9, 0.02M Calcium chloride 33% V/V Hexylene Glycol, 1mM UDP	16	Rods	No Cryo
10	0.15M Potassium Bromide, 33% w/v Polyethylene Glycol Mono-methyl Ether, pH 6.5, 1mM UDP	16	Bipyramidal	N/A
7.5 - 15	0.2M Potassium Thiocynate, 0.1M Bis-Tris Propane pH 7.5, 20% w/v Polyethylene Glycol 3350, 1mM UDP.	16	Plates	30% v/v glycerol

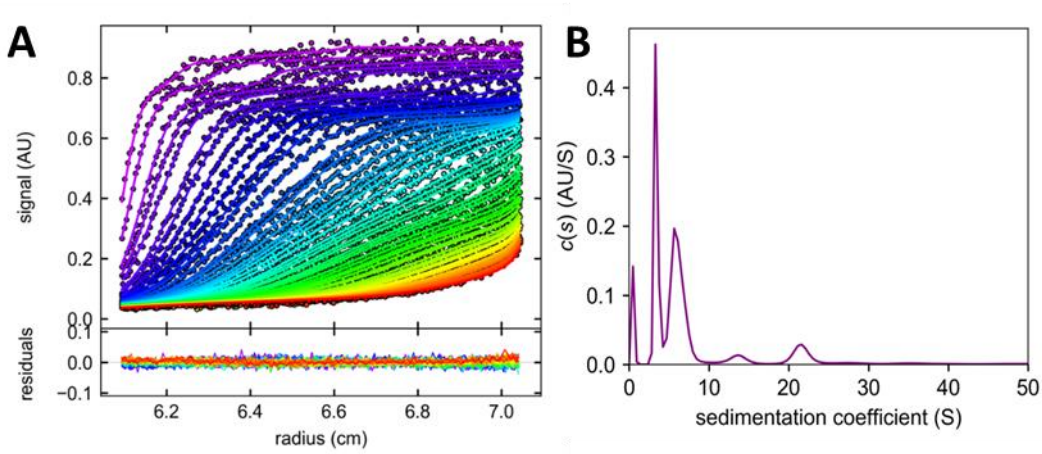

Figure 63: : Analytical ultracentrifugation of showing the oligomerisation of PelA. ***A.*** *: Lamm equation solutions fitted to absorbance data using the c(s) model for boundary migration during sedimentation velocity PelA 46-948 (lower). Residual absorption from the fitted data is shown below each panel.* ***B:*** *PelA 46-948 (lower). PelA 46-948 a profile of five species at sedimentation values of 0.5, 3.53 (MW: 46.8 kDa +/- 4.9kDa), 6.36 (MW: 99.8kDa +/-10kDa), 14.3 (MW: 383kDa +/-40kDa) and 22.5 S (MW: 759kDa +/- 40kDa). This indicates the presence of a monomer of PelA 46-948, as well as a tertramer and octamer.*