# Structural, functional and biophysical characterization of the secreted, β-helical adhesin EtpA of enterotoxigenic *Escherichia coli*

by

Clifford Manyo Ntui

(14334012)

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Thesis submitted in partial fulfilment for the degree *Philosophiae doctor* (PhD) at the Department of Biochemistry, Genetics and Microbiology, Natural and Agricultural Sciences Faculty, University of Pretoria

Supervisor: Prof Wolf-Dieter Schubert

# DECLARATION OF ORIGINALITY UNIVERSITY OF PRETORIA

The Department of Biochemistry, Genetics and Microbiology places great emphasis upon integrity and ethical conduct in the preparation of all written work submitted for academic evaluation. While academic staff teach you about referencing techniques and how to avoid plagiarism, you too have a responsibility in this regard. If you are at any stage uncertain as to what is required, you should speak to your lecturer before any written work is submitted.
You are guilty of plagiarism if you copy something from another author's work (e. g a book, an article or a website) without acknowledging the source and pass it off as your own. In effect you are stealing something that belongs to someone else. This is not only the case when you copy work word-for-word (verbatim), but also when you submit someone else's work in a slightly altered form (paraphrase) or use a line of argument without acknowledging it. You are not allowed to use work previously produced by another student. You are also not allowed to let anybody copy your work with the intention of passing if off as his/her work. Students who commit plagiarism will not be given any credit for plagiarized work. The matter may also be referred to the Disciplinary Committee (Students) for a ruling. Plagiarism is regarded as a serious contravention of the University's rules and can lead to expulsion from the University. The declaration which follows must accompany all written work submitted while you are a student of the Department of Biochemistry. No written work will be accepted unless the declaration has been completed and attached.

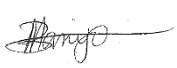
**Full name of student**: Clifford Manyo Ntui
**Student number**: 14334012
**Topic of work**: Structural, functional and biophysical characterization of the secreted, β-helical adhesin EtpA of Enterotoxigenic *Escherichia coli*

1. I understand what plagiarism is and am aware of the University's policy in this regard.

2. I declare that this dissertation is my own original work. Where other people's work has been used (either from a printed source, Internet or any other source), this has been properly acknowledged and referenced in accordance with departmental requirements.

3. I have not used work previously produced by another student or any other person to hand in as my own.

4. I have not allowed, and will not allow anyone to copy my work with the intention of passing it off as his or her own work.

**SIGNATURE**:  **DATE**: January 2024

# Acknowledgements

First and foremost, I would like to thank the creator of heaven and earth for HIS protection and guidance throughout my studies.

I would also like to express my sincere gratitude to the following people:

- My late grandparents (Zachariah Tabe Ntui, and Paulina Ebanga Ntui), and my late mother and brother (Lucia Ndip Ntui and Collins Nkongho Ntui) who taught and encouraged me to be the better person that I am today. Without their support and encouragement, possibly I would not have reached this far in life. I really appreciate their contribution in my life.
- My supervisor, Professor Wolf-Dieter Schubert for giving me the opportunity to undertake this project. Thank you for the contributions you made towards my academic journey. It was a privilege to be in the Structural Biology research group as I have learnt a lot in this field.
- Professor James M. Fleckenstein for providing some of the constructs used for this project.
- Dr Previn Naicker for assisting me with my CD spectroscopy experiments.
- Last but not the least, thanks to past colleagues (Mr Ade Adewumi, Mr Valentine Anye, Mr Mthwelanga Ndengane, Mr Lungelo Mandyoli, Mr Hildsley Noome, Mr Eugene Duvenage, Mr Kofi Boafoh, Mr Lenience Maphosa and Ms Clare Boswell) from the Structural Biology of Infectious Disease research group, University of Pretoria. Their presence in the lab and constructive criticism contributed to the success of this project.
- Special thanks to the Synchrotron techniques for African research and technology (START) for providing funding for this project.

# Summary

Enterotoxigenic *Escherichia coli* (ETEC) is a diarrhoeal pathogen associated with severe morbidity and mortality among young children in developing countries. At present, there is no vaccine for ETEC. One candidate vaccine antigen, EtpA, is a conserved secreted adhesin that presumably binds to the tips of the *E. coli* flagellum to link ETEC to host intestinal glycans. EtpA is exported through a Gram-negative, two-partner secretion (TPS) system (type Vb) comprised of the secreted EtpA passenger (TpsA) protein and EtpB (TpsB) a transporter integrated into the outer bacterial membrane. TpsA proteins share a conserved, N-terminal domain followed by an extensive C-terminal domain with divergent sequence repeats. Three constructs of EtpA were prepared and analysed respectively including residues 67 to 447 ($EtpA^{67-447}$), residues 1 to 606 ($EtpA^{1-606}$) and residues 67-930 ($EtpA^{67-930}$). The crystal structure of $EtpA^{67-447}$ solved at 1.76 Å resolution revealed a right-handed parallel β-helix with two extra-helical hairpins and an N-terminal β-strand cap. Analyses by circular dichroism spectroscopy confirmed the β-helical fold and indicated high resistance to chemical and thermal denaturation as well as rapid refolding. A theoretical model of full-length EtpA by AlphaFold largely concurs with the crystal structure adding a long β-helical C-terminal domain after an interdomain kink. We propose that robust folding of the TPS domain upon secretion provides a template to extend the N-terminal β-helix into the C-terminal domains of TpsA proteins. Unexpectedly, size exclusion chromatography and molecular pulldown assays with the N-terminal domain of EtpA failed to show any interaction with flagellin implying that other factors may be involved.



**Keywords:** Infectious Diseases, Structural Biology, CD Spectroscopy.

# Contents

# List of figures

# List of tables

# List of appendices

# Abbreviations

| | |
|---|---|
| A/E | Attaching and effacing (lesions) |
| ACE-2 | Angiotensin converting enzyme-2 |
| ADP | Adenosine diphosphate |
| AIEC | Adherent invasive E. coli |
| AmpR | Ampicillin resistant |
| APS | Ammonium persulfate |
| AR | Acid resistant |
| bp | Base pair |
| cAMP | Cyclic adenosine monophosphate |
| CD | Circular dichroism |
| CFA/I | Colonization factor antigen I |
| CFs | Colonization factors |
| CFTR | Cystic fibrosis transmembrane conductance regulator |
| CS | Coli surface antigen |
| CV | Column volume |
| Da | Dalton |
| $ddH_2O$ | Double distilled water |
| DLS | Dynamic light scattering |
| DNA | 2'-Deoxyribonucleic acid |
| dNTPs | Deoxynucleoside triphosphate |
| *E. coli* | *Escherichia coli* |
| EAEC | Enteroaggregative *E. coli* |
| EDTA | Ethylenediaminetetraacetic acid |
| EHEC | Enterohaemorrhagic *E. coli* |
| EIEC | Enteroinvasive *E. coli* |
| ELISA | Enzyme-linked immunosorbent assay |
| EPEC | Enteropathogenic *E. coli* |
| ETEC | Enterotoxigenic *E. coli* |
| FHA | Filamentous hemagglutinin |
| FliC | Flagellin |
| FPLC | Fast pressure liquid chromatography |
| GB3 | Globotriaosylceramide |
| GC-C | Guanylyl cyclase C |
| GI | Gastrointestinal |
| GS | Glutathione sepharose |
| GST | Glutathione S-transferase |
| $His_6$-tag | Hexa-histidine tag |
| HMW1-PP | High molecular weight 1 pro-piece |
| HRM-qPCR | High-resolution melting qPCR |

| | |
|---|---|
| IEC | Ion exchange chromatography |
| IPTG | Isopropyl β-D-thiogalactoside |
| $Kan^R$ | Kanamycin resistance |
| kb | Kilo base |
| LB | Lysogeny broth (medium) |
| LDDT | Local Distance Difference Test |
| LEE | Locus of enterocyte effacement |
| LT | Liable enterotoxins |
| MCS | Multiple cloning site |
| NAD | Nicotinamide adenine dinucleotide |
| Ni-NTA | Nickel nitrilotriacetic acid |
| OD600 | Optical density at a wavelength of 600 nm |
| PBS | Phosphate buffer saline |
| PCR | Polymerase chain reaction |
| PMSF | Phenylmethylsulfonyl fluoride |
| POTRA | Polypeptide transporter associated |
| qPCR | Quantitative PCR |
| RNA | Ribonucleic acid |
| rpm | Revolution per minute |
| rSAP | Shrimp alkaline phosphatase |
| SARS-CoV-2 | Severe acute respiratory syndrome coronavirus |
| scFv | Single-chain fragment variable |
| SDM | Site-directed mutagenesis |
| SDS-PAGE | Sodium dodecyl-sulphate polyacrylamide gel electrophoresis |
| SEC | Size exclusion chromatography |
| SP | Signal peptide |
| ST | Stable enterotoxins |
| STEAEC | Shiga toxin-producing enteroaggregative *E. coli* |
| Stx | Shiga toxins |
| TAE | Tris acetic acid-EDTA |
| TEMED | Tetramethylethylenediamine |
| TPS | Two-partner secretion |
| TPSS | Two-partner secretion system |

# 1 Introduction

Despite decades of medical advances and much improved lifestyles, pathogen-related diseases continue to pose a persistent threat to humanity. The infectious agents, be they viruses, bacteria, fungi, protozoa, or helminths, can trigger a host of responses allowing them to colonize and administer their infectious particles to their targets. Since its documentation in 1885 [1, 2], *Escherichia coli* (*E. coli*) has probably become the most broadly studied living species. Although *E. coli* is an essential intestinal commensal of mammals, some strains encode pathogenic systems that cause diseases in humans and animals [3, 4]. In humans, *E. coli* causes diverse enteric infections using virulence factors that target a wide range of cellular processes [5, 6]. Strains are assigned to a range of pathotypes based on their virulence profiles. Pathotypes include enteropathogenic (EPEC), enterohaemorrhagic (EHEC), enterotoxigenic (ETEC), enteroinvasive (EIEC), enteroaggregative (EAEC), adherent invasive (AIEC) and Shiga toxin-producing enteroaggregative *E. coli* (STEAEC) [6-8]. While this project will focus on ETEC, the major cause of infant diarrhoea in developing countries, other pathotypes will also be briefly introduced.

To infect humans, pathogenic enteric *E. coli* must not only endure the passage through the human gastrointestinal (GI) tract but must also achieve their pathogenic progression by complex and coordinated multistage tactics. These include adherence to the host intestine and toxin/virulence protein assembly [9, 10]. ETEC elicits diarrhoea through heat-stable (ST) and heat liable (LT) enterotoxins [11]. But before releasing toxins, the bacteria first colonize and attach to the mucus lining of the small intestine through colonization factors (CS) such as the colonization factor antigen I (CFA/I) [12] and non-fimbrial virulence factors such as EtpA [13]. In this regards, ETEC physiopathology, its diagnosis, the regulation of virulence genes by gastrointestinal cues and the pathogenesis-related mechanism of individual virulence factors will also be discussed.

## 1.1 Pathogenic *E. coli* associated with gastrointestinal illness

Pathogenic *E. coli* causing gastrointestinal illnesses are grouped into six pathotypes based on virulence mechanisms, infectious processes and damages triggered within the target cells.

### 1.1.1 Enterohaemorrhagic or Shiga toxin-producing *E. coli*

Enterohaemorrhagic, Shiga toxin-producing *E.* coli (EHEC/STEC) are foodborne bacterial pathogens characterised by the production of one or more Shiga cytotoxins [14], that cause symptoms ranging from mild diarrhoea to haemorrhagic colitis and haemolytic uremic syndrome [15]. The main Shiga toxins are encoded by the genes *stx1* and *stx2*, dispersed by lysogenic phages [16, 17]. Stx Shiga toxins are a subgroup of the catalytic AB5 protein toxins that bind the multi-organ host receptor Gb3. Receptor binding induces Stx internalization, where it inhibits host protein synthesis by removing an adenine from the 28S RNA of the 60S ribosomal subunit [18]. Intimin, a membrane-embedded virulence factor, facilitates tight bacterial adhesion to epithelial cells creating characteristic histopathological attaching and effacing (A/E) lesions characterised by intimate bacterial adhesion, the localised destruction of brush border microvilli, and significant cytoskeletal restructuring. A/E lesion formation is coordinated by a large pathogenicity island known as the locus of enterocyte effacement (LEE). LEE encodes intimin, the translocated intimin receptor (Tir), a type III secretion system (T3SS), and other secreted proteins. Intimin determines host-cell binding, A/E lesion formation, and colonisation of mucosal surfaces [19, 20].

### 1.1.2 Enteropathogenic *E. coli*

Enteropathogenic *E. coli* (EPEC), the first pathotype of *E. coli* to be described, resembles EHEC but lacks Shiga toxins [21]. Strains are classified as typical (tEPEC) and atypical (aEPEC) based on the presence of the EPEC adherence factor plasmid (pEAF). The plasmid encoded *bfp* gene encodes the bundle-forming pilus required for adherence [22]. EPEC induce lesions by translocating effector proteins into the host cell using a LEE-encoded T3SS. Other LEE genes encode outer membrane adhesion, translocator, and effector proteins, chaperones, the intimin receptor and other regulatory proteins [23, 24]. Non-LEE proteins also contribute to bacterial virulence by modulating the host inflammatory response and by disrupting the tight junctions and the cytoskeleton of the host cells [25, 26].

### 1.1.3 Enteroaggregative *E. coli*

Enteroaggregative *E. coli* (EAEC) causes persistent diarrhoea in children and adults. It adheres to host cells through aggregative adherence fimbriae (AAFs). Upon colonizing the intestinal mucosa, it secretes enterotoxins and cytotoxins to damage the mucosa. EAEC toxins include autotransporter mucinase Pic, “Shigella enterotoxin 1” (ShET1) common to most *Shigella*

*flexneri* 2a strains, and the "enteroaggregative *E. coli* ST" (EAST1) enterotoxin. ShET1 and EAST1 both contribute to watery diarrhoea. EAEC virulence factor production is regulated by AggR, a member of the AraC family of transcriptional regulators [27, 28].

### 1.1.4 Diffusely adherent *E. coli*

Diffusely adherent *E. coli* (DAEC) causes diarrhoea in children of twelve months and older. DAEC strains produce the fimbrial adhesin F1845 from the "Dr family" of adhesins. The adhesins all use the "decay accelerating factor" (DAF) as their host receptor. DAF, an extracellular, glycosylphosphatidylinositol-anchored protein, inhibits complement activation on host cell surfaces by dissociating C3 convertase enzymes [29, 30]. By binding and clustering the DAF receptor through fimbriae, DAEC strains induce the development of long cellular extensions that wrap around the adherent bacteria. [31, 32].

### 1.1.5 Enteroinvasive *E. coli*

Enteroinvasive *E. coli* (EIEC) cause human dysentery, especially in developing countries. EIEC strains behave like Shigella in their ability to invade gut epithelia and cause dysentery-like illness. EIEC do not produce enterotoxins. Instead, a large pINV plasmid encodes the genes required for invasion, survival, and diffusion of the bacteria within the host [33, 34].

### 1.1.6 Enterotoxigenic *E. coli*

Enterotoxigenic *E. coli* (ETEC) is among the top four causes of moderate to severe diarrhoea in African and South Asian children. It kills hundreds of thousands of children and naïve adults annually, especially in areas with poor sanitation and a lack of clean drinking water [35, 36]. Serogroups O6, O78, O8, O128 and O153 of the pathogen are the key cause of outbreaks [12]. ETEC is an economic burden to farmers and industry, as a key pathogen of broiler chickens, swine, cattle, and other farm animals [37].

## 1.2 ETEC Infections

### 1.2.1 Aetiology of ETEC infection

ETEC is acquired through the consumption of contaminated water or food with humans being the reservoir of infection. Inadequate sanitation and sewage facilities hence drive the contamination of surface or drinking water. The infective dose for ETEC is between $10^8$ to $10^{10}$ cells in adults, but much lower in infants [38-40]. The infection develops one to three days

post exposure and typically lasts three to four days. Most patients recuperate without hospitalization or antibiotics. Symptoms include fever, headache, nausea, vomiting, and muscle ache [40, 41].

ETEC infections are mostly self-limiting and are clinically difficult to distinguish from cholera. In developing countries, diarrhoea and malnutrition interlink to impact the health status of children. Treatment is currently mainly supportive, with oral rehydration preventing dehydration and loss of electrolytes. Antimicrobial drug use is problematic during ETEC infections since etiologic diagnoses take too long [42]. Fluoroquinolones effectively treat ETEC traveller's diarrhoea but are prescribed with caution to prevent antimicrobial resistance. Alternative prophylactic or therapeutic approaches to ETEC are being developed [43]. Existing vaccines are used with variable success in animal models and/or humans, but are not commercially available [44].

### 1.2.2 Virulence factors

After ingestion by humans, ETEC induces an infection by colonizing the intestinal mucosal surface and producing toxins. ETEC pathogenicity is, however, not limited to toxins but combines different virulence behaviours. In the small intestine, ETEC strains produce fimbrial and non-fimbrial adhesions that facilitate bacterial attachment and colonization. [45, 46]. To survive the acid barrier of the human stomach, *E. coli* strains have developed acid resistant (AR) systems including the glutamate-dependent AR system. [47].

### 1.2.3 Colonization factors

ETEC adheres to the intestinal epithelium using more than 30 antigenically and structurally diverse extracellular colonization factors (CF) [48]. While CF modulate the disease in humans, about 40 % of worldwide isolates lack detectable CF [40]. CF include fimbrial, fimbrillar and helical structures. They are assigned to three groups according to antigenic, genetic, and biochemical properties:

0) Coli surface antigens CS3, 6, 10 and 11 are not yet assigned to any CF group [48].

1) The colonization factor antigen I (CFA/I)-like group includes the original CF (CF/I), the rod-like pili of the class 5a fimbriae, as well as CS1, 2, 4, 14, 17 and 19 [49]. CFA/I assembly depends on the plasmid-encoded *cfaABCE* operon, with subunit CFaA ensuring pilus export across the periplasm and CFaB promoting interactions with mucosal glycosphingolipids and glycans of the small intestine [50, 51].

2) The coli surface antigen 5 (CS5)-like group comprises helical CS5 and 7. CS5 operons encode major and minor subunits, outer membrane ushers, two chaperones and pili length regulators [52].

3) Class 1b group includes CS12, 18, 20, 26, 27, 28 and 30. They are all related to the F6 adhesin of neonatal piglet specific ETEC and share its operon structure [48]. Mucous and cell surface glycoproteins of the small intestine appear to function as host receptors for CF [53].

Non-fimbrial virulence factors also contribute to ETEC virulence: Outer membrane protein Tia and autotransporter TibA moderate adhesion to and invasion of colonic epithelial cell line HTC-8 [54]. EtpA binds to the tips of ETEC flagella linking them to host cell receptors and mediating adherence to HCT-8 and Caco-2 epithelial cells [13]. LeoA is a protein with GTPase activity essential for LT secretion [55]. EatA, a member of the serine protease autotransporters of the Enterobacteriaceae family, and YghJ, a mucin-binding metalloprotease, degrade mucin to provide access to the intestinal cell membrane and promote bacterial adhesion [56].

### 1.2.4 Secretion of toxins

After adhering to the intestinal mucosa, ETEC strains produce enterotoxins that cause key clinical symptoms and are thus seen as the main ETEC virulence factors. Heat-stable toxin (ST) and heat-labile toxin (LT) are the main toxins associated with ETEC diarrhoea.

LT toxins structurally and functionally resemble cholera toxins [57]. They consist of five identical protein monomers (11.5 kDa) where the five B domains form a pentameric ring. The singleA subunit consists of an enzymatic A1 domain plus an extended A2 α-helix that insert into the centre of B subunit pentamer. Transported by outer membrane vesicles, LT irreversibly binds to the host monoganglioside receptor GM1 via its B subunit inducing toxin internalization and transport to the endoplasmic reticulum. Here, the $A_1$ domain is cleaved from the $A_2$ domain and released into the cytoplasm. The $A_1$ domain cleaves the cofactor $NAD^+$ to ADP-ribosylate stimulatory G proteins. The activated G proteins upregulate adenylate cyclase increasing intracellular cyclic adenosine monophosphate (cAMP) concentrations. Rising cAMP activates protein kinase A, which phosphorylates ion channels stimulating them to release $Cl^-$ and reduce $Na^+$ uptake. Associated massive fluid release into the intestinal lumen initiates watery diarrhoea. LT additionally regulates host cell function and confers a competitive advantage to ETEC for adherence to cultured intestinal epithelial cells.

ST toxins are cysteine-rich peptides that mimic the human hormone guanylin. The two ST variants, STa and STb, reversibly bind guanylyl cyclase C (GC-C) and sulphatide, respectively, to once again activate CFTR initiating diarrhoea. The ST toxins can occur alone or in combination with LT [58, 59].

### 1.2.5 Regulation of virulence by gastrointestinal cues

For bacteria to attain full pathogenicity, they must not only survive the human GI tract but also need to synchronise the production of virulence determinants with the local gut microenvironment. ETEC responds to various GI cues to modulate its virulence factor production.

#### 1.2.5.1 pH

Upon ingestion, pathogens encounter the host digestive tract. The pH drops from largely neutral values to highly acidic in the stomach, before returning to near-neutral values in the small intestine. ST release by ETEC appears pH-dependent [60], while LT release increases with a rise in alkalinity [61, 62]. The pH gradient in the GI tract thus allows ETEC to regulate LT toxin production and secretion to reach a maximum in the small intestine [63].

#### 1.2.5.2 Bile

Pathogens are extensively challenged by bile salts in the small intestine. Due to re-absorption, however, the bile concentration decreases between duodenum and colon. *In vitro*, 2 g/L bile prevents the binding of LT toxin to GM1, mainly due to arachidonic, linoleic and oleic unsaturated fatty acids [64]. In rabbit ileal loops, increasing amounts of linoleic acid prevented LT-mediated fluid accumulation. On the other hand, concentrations of 30 g/L led to genes *estA, eltA*, or *etpA* respectively encoding STa, LTa and EtpA, to be transcriptionally upregulated whereas genes *csoA* and *cstA* for the CS1 and CS3 colonization factors were downregulated [65]. Transcriptional responses to bile salts are strain-dependent, such that individual observations should be applied to the entire pathovar with care. A threshold of 1.5 g/L bile salts were required for surface presentation of CS5, CS7, CS8, CS12, CS14, CS17 and CS19 [66-68] while bile salts were not required to stimulate the production of CS1, CS2 and CS3 [68] . Varying bile acid concentrations along the human intestine thus clearly differentially modulate both the interaction of LT toxin with its receptors and the expression of ETEC colonization factors.

#### 1.2.5.3 Digestive enzymes

The impact of human digestive enzymes on the expression of pathogen virulence genes have only rarely been studied. In ETEC, *in vitro* assays revealed that trypsin secreted by the duodenal epithelial cells, increase LT production and its secretory activity [61, 69].

### 1.2.6 Diagnosis of ETEC infections

The diagnosis of ETEC associated diarrhoea is complicated by the diversity of *E. coli* virulence factors. ETEC identification requires detecting LT and/or ST in addition to virulence genes clyA, eatA, tia, tibC, leoA, or east-1 by complementary PCR [70]. Detection may involve specific, quantitative, or multiplex PCR assays with primers targeting *lt* and *st* genes [71]. ETEC was previously detected phenotypically using supernatants from single *E. coli* colonies and procedures such as the rabbit ileal loop test, the suckling mouse assay or studies of the cytopathic effect on CHO or Y1 adrenal cell monolayers, where the presence of LT in supernatants induced rounding of Y1 cells or elongation of CHO cells after 24 h incubation [72-74]. Radioimmunoassay and enzyme-linked immunosorbent assay (ELISA) for ST detection through binding to the GM1 receptor confirm the suckling mouse assay [75-77].

### 1.2.7 Vaccines challenges

Although ETEC strains were first discovered about five decades ago, no broadly protective vaccine has been developed. Reasons include the short peptide length of ST molecules, which have low intrinsic immunogenicity, and the plasmid-encoded fimbrial colonization factors' antigenic heterogeneity, the principal targets of ETEC vaccines [78, 79]. Despite these challenges, humans develop immunity after natural or experimental [39] ETEC infections suggesting that a vaccine is theoretically possible. Molecular and proteomic [80] research have proved that ETEC infection is more complex, involving both unique virulence factors and chromosomally encoded proteins. Antigens distinct from virulence factors may thus also need to be targeted for a highly protective vaccine. For ETEC vaccines to prevent the conveyance of toxins to receptors on epithelial cell, toxins would either need to be neutralized directly or their effective delivery prevented [81-84].

For ETEC to deliver toxins it must colonize and adhere to epithelial cells, facilitated by adhesins, flagella and other virulence factors. The protein EtpA has been identified as one protein among others that is critical for ETEC adherence to epithelial cells.

## 1.3 The two-partner secretion system

Pathogenic bacteria employ various strategies to infiltrate their hosts, cause tissue damage, and interfere with the immune system. Secreting proteins across phospholipid membranes is a crucial part of these strategies. Secreted proteins enhance adhesion to eukaryotic cells, or directly intoxicate target cells and impair their functioning to promote bacterial pathogenicity. Specialized protein secretion systems are used by bacterial pathogens to release virulence factors into host cells or the environment. Bacterial protein secretion machineries are broadly classified according to their architectures, roles, and selectivity. Some systems are exclusive to a few bacterial species or to one or more proteins, while others are common to multiple bacterial groups and/or transport multiple substrates.

At least seven distinct secretion systems have been identified in Gram-negative bacteria to secrete proteins across both the inner and outer cell membranes. They are known as type I to type VI and type VIII secretion systems denoted T1SS to T6SS and T8SS [85, 86]. Among these systems, Type V secretion systems are further subdivided into subtypes a to e (Figure 1-1). All subtypes consist of both a passenger protein/domain and an outer membrane transporter either produced as distinct proteins (types b and d) or fused into single polypeptides (types a, c, and e). The passenger proteins are mostly quite large. In the case of types a and b, passengers form β-helical structures characterised by extended sequence repeats frequently containing further internal repeats. The transporters are transmembrane β-barrel pores with 12 β-strands in types a, c, and e, and 16 β-strands in types b and d. Both secretion and folding of type V proteins through and beyond the outer membrane occur without energy sources such as hydrolysable nucleotides [87, 88]. In this thesis, subtype Vb secretion also known as “two-partner secretion system” (TPSS) is of particular relevance as it is used to secrete the protein EtpA.

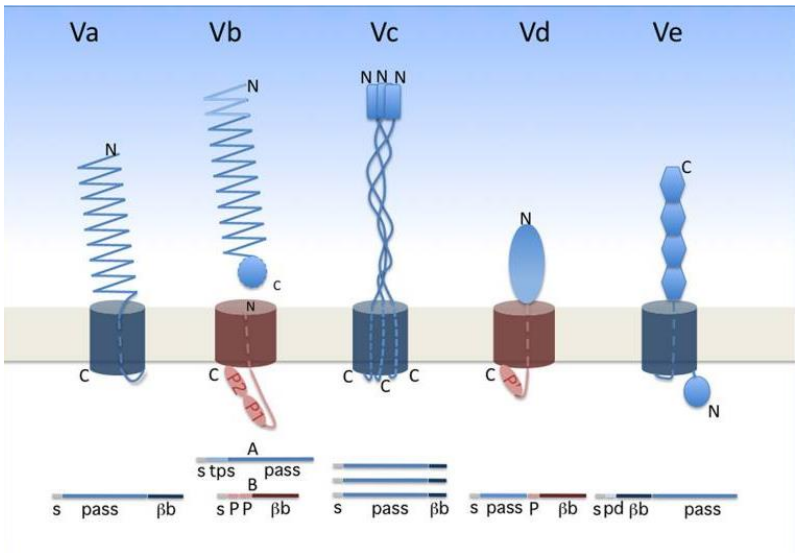



**Figure 1-1: Type V secretion system subtypes**

Mature proteins for each Type V secretion subsystem are depicted in the main figure summarized by a schematic line representation below. (Adapted from [89]).

### 1.3.1 TpsA proteins

In type Vb or two partner secretion (TPS) systems (TPSS), the passenger proteins and membrane transporters are generically referred to as TpsA and TpsB. TpsA proteins are large secreted exoproteins typically larger than 100 kDa. The nascent proteins contain an N-terminal signal sequence for Sec secretion across the inner membrane into the periplasm. Next in sequence is a TPS domain, which is recognised by the periplasmic domains of a dedicated transporter to initiate secretion across the outer membrane. TpsA proteins invariably adopt right-handed β-helical folds, with folding presumably being initiated by the TPS domains at the cell surface in parallel with their secretion. The proteins also characteristically share slow folding rates, high solubility once folded and a tendency to avoid aggregation when unfolded [90]. TpsA proteins typically contain multiple nested but imperfect sequence repeats resulting in partly repeated structural motifs in the extended β-helices. The structural analysis of TpsA proteins is complicated by their large size and poor stability *in vitro*. Because a cognate partner is required to secrete each TpsA and achieve native folding overexpression studies have proven difficult. Correspondingly, only a small number of TpsA crystal structures are currently available, mostly limited to N-terminal TPS domain or fragments thereof. They include four TPS domain structures respectively from *Bordetella pertussis* (protein FHA30), *Proteus mirabilis* (HpmA265), and *Hemophilus influenzae* (HMW1-PP and HxuA) [91-94]. The only full-length structure is that of the smallest

TpsA protein HxuA of 96 kDa [92]. The structures reveal TPS domains consisting of right-handed β-helices with few α-helical elements. TpsA structures occasionally also include structural motifs that extend out of the β-helix core that might be essential for inter-bacterial interactions [92, 94, 95].

### 1.3.2 TpsA protein functions

The TPSS is used to secrete a range of large protein virulence factors of Gram-negative human pathogens such as ETEC, *Haemophilus influenza*, and *Bordetella pertussis*. The TpsA proteins are cluster by function as cytolysins/hemolysins, adhesins, proteases or functioning in iron acquisition (Table 1-1).

**Table 1-1: Functions of typical TPSS proteins [89]**

| Subgroup | | Function | Reference |
|---|---|---|---|
| **CYTOLYSINS/HEMOLYSINS** | | | |
| Shla | *Serratia marcescens* | Cytolysin, hemolysin, pore-forming toxin, induces autophagy | [96-98] |
| HpmA | *Proteus mirabilis* | Cytolysin, hemolysin | [99, 100] |
| EthA | *Edwardsiella tarda* | Cytolysin, hemolysin, host cell adherence, internalization in fish | [101-103] |
| HhdA | *Haemophilus ducreyi* | Hemolysin | [104] |
| PhlA | *Photorhabdus luminescens* | Hemolysin | [105] |
| ExlA | *Pseudomonas aeruginosa* | Exolysin, pore-forming toxin, ruptures plasma membrane of human cells | [106, 107] |
| **PROTEASES** | | | |
| LepA | *Pseudomonas aeruginosa* | Induces inflammatory responses via human protease-activated receptors (PARs) | [108, 109] |
| **IRON ACQUISITION** | | | |
| HxuA | *Haemophilus influenzae* | Heme acquisition from hemopexin | [110, 111] |
| **ADHESINS** | | | |
| FHA | *Bordetella pertussis* | Adhesion to epithelial cells, biofilm formation, immunomodulation | [112-114] |
| HMW1/ HMW2 | *Haemophilus influenza* | Adhesion to epithelial cells | [115] |
| EtpA | *Escherichia coli* | Intestinal colonization, host cell adhesion by binding to the tip of flagella | [13] |
| EtpB transporter | *Escherichia coli* | Adhesion to epithelial cells | [116] |
| CdrA | *Pseudomonas aeruginosa* | Biofilm, binding to Ps1 exopolysaccharides | [117] |
| Ap58 (EnfA) transporter | *Escherichia coli* | Adhesion and hemagglutination activities | [118] |

The protein EtpA from ETEC appears to support intestinal colonization by binding to the tip of the flagella, while LspA1 and LspA2 proteins from *Haemophilus ducreyi* phosphorylate

eukaryotic Src tyrosine kinase to inhibit phagocytosis and hence support virulence [13, 89, 119, 120].

### 1.3.3 TpsB proteins

TpsB proteins are members of the Omp85/TpsB superfamily [121, 122]. TpsB proteins invariably translocate their TpsA partner proteins across the outer membranes while other members of the Omp85 family also insert proteins into outer membranes [123]. Crystal structures of Omp85/TpsB proteins show that they share C-terminal transmembrane β-barrel domains preceded by N-terminal polypeptide transporter associated (POTRA) domains (Figure 1-2) [124-128].

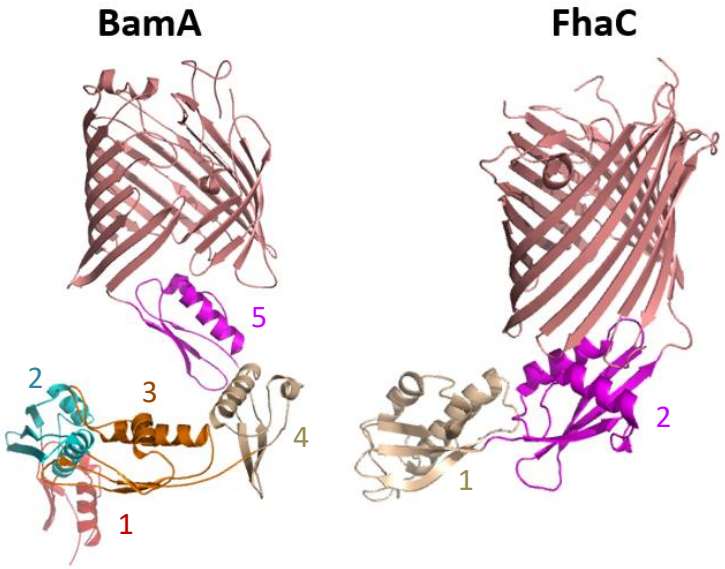



**Figure 1-2: Crystal structures of BamA from *N. gonorrhoeae* and FhaC from *B. pertussis***

Cartoon representations of Omp85 protein BamA (PDB code 4K3B) and TpsB protein FhaC (4QKY). The membrane-integral β-barrels constituting the translocation pore are shown in salmon. POTRA domains in different colours are numbered 1 to 5.

The only crystal structure of a TpsB transporter is that of FhaC from *B. pertussis* [127, 128], which translocates its partner TpsA FHA across the outer membrane [129]. Transporters from the Omp85 family like BamA integrate proteins into the outer membrane of Gram-negative bacteria [130-132]. In both groups, POTRA domains recognize their partners for transport across or integration into the outer membrane [95, 133-136].

### 1.3.4 Export of TpsA across both inner and outer membranes

TpsA proteins are secreted across cytoplasmic membrane by the Sec-system due to N-terminal signal peptides [137, 138]. In the periplasm their TPS domains are recognised by TpsB

POTRA domains as a prelude to their translocation across the outer membrane. Initially evidence appeared to indicate that the TPS domain remained bound to the POTRA domains until the remaining protein had translocated and folded in the extracellular space. More recent evidence, however, indicates that the TPS domain is translocated first and initiates protein folding [89, 139, 140]. Experimental evidence includes the fact that N-terminal fragments of TpsA proteins retaining the TPS domains fold much more efficiently and that the N-terminus of FHA constructs stalled during translocation is located at the cell surface [141, 142]. The TPS fold is particularly stable and facilitates TpsA folding *in vitro* [93, 143] further supporting the idea that TPS domains initiate folding in two-partner secretion.

### 1.3.5 The Two-partner secretion system operon and its presence in ETEC

The TPSS has been described in various pathogens including *B. pertussis* (FhaAB), *H. influenza* (HMW1ABC and HMW2ABC), *P. mirabilis* (HpmAB) and *S. marcescens* (Sh1AB). The genes for a particular TPSS mostly cluster in the same operon (Figure 1-3), with some exceptions [144]. In some TPSS, a single TpsB recognises and secrete multiple TpsA proteins though the specificity for their partners may vary [95, 145]. Also the order of genes may vary with the order being *Hmw1A* (TpsA), *Hmw1B* (TpsB) and *Hmw1C* for the HMW1ABC TPSS in *H. influenza*, where both proteins HMW1B and HMW1C are required to secrete the HMW1A passenger protein [146, 147]. In ETEC, the order is *etpB*, *etpA, etpC*, where the encoded EtpC glycosylates the passenger protein EtpA [116].

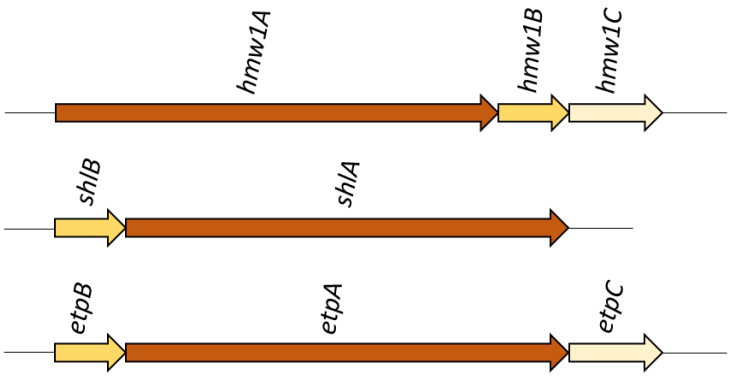



**Figure 1-3: Organization of typical TPSS operons**

A single TpsB recognises and secret multiple TpsA proteins. Genes for TpsB, TpsA and glycosylation are indicated in yellow, orange and white respectively.

## 1.4 EtpA

In bacterial pathogens, adherence precedes colonization and toxin delivery through specific ligand-receptor or less specific interactions [148]. The receptor is mostly either a protein or a glycoconjugate on the host cell surface. In Gram-negative bacteria, adherence factors include fimbrial, non-fimbrial and polysaccharide adhesion proteins [149]. The EtpA glycoprotein of ETEC is an adhesin that promotes adherence and intestinal colonization of ETEC by binding to the tips of flagella and interacting with host glycans [13, 150].

### 1.4.1 Molecular features

The ETEC protein EtpA consists of 1767 amino acid residues. It starts with an N-terminal signal peptide, followed by a TPS domain, a long C-terminal region of unknown function and a short tail fragment involved in flagellal anchoring. The extended C-terminal region encompasses four 228 residue sequence repeats (R1, R2, R3 and R4) (Figure 1-4) [151].

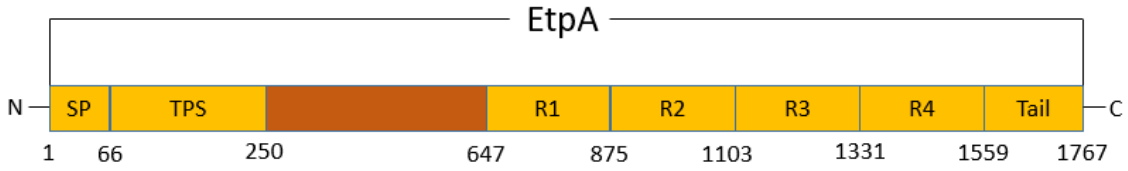



**Figure 1-4: Schematic domain structure of ETEC EtpA**

An N-terminal signal peptide (SP) signals the protein for Sec secretion across the inner membrane. The TPS domain ensures recognition by its cognate secretion partner across the outer membrane. Four consecutive 228 residue repeats are denoted R1, R2, R3 and R4. A C-terminal tail domain presumably binds EtpA to the flagellal tip. An uncharacterised central region between TPS domain and C-terminal repeats is shown in dark orange. Numbers mark the start and end of each domain.

### 1.4.2 Functions of EtpA protein in ETEC

Research on EtpA have identified several roles in ETEC pathogenicity. Most involve its interaction with host glycans or bacterial flagellin to facilitate host engagement prior to toxin delivery.

#### 1.4.2.1 EtpA interacts with glycan to facilitate ETEC toxin delivery.

Mucins are protective glycoproteins separating the intestinal lumen from the underlying epithelium. For bacteria to deliver toxins to host cells, they first need to overcome this protective layer. EtpA engages multiple mucin glycoproteins specifically through the glycan moieties to support bacterial attachment and toxin delivery [152]. Cells depleted of mucin

result in reduced ETEC adhesion and toxin delivery [152]. EtpA appears somewhat specific for blood group A glycans on intestinal epithelial cells [171].

#### 1.4.2.2 EtpA mediates adhesion between flagella and host cells.

Flagella are complex molecular structures that enable bacterial motility and ensure their survival. Flagella also participate in host cell adhesion and invasion [152]. The bacterial flagellum consists of a basal body, a hook, and a filament (Figure 1-5A) composed of as many as 20 000 flagellin (FliC) or related subunits (Figure 1-5B).

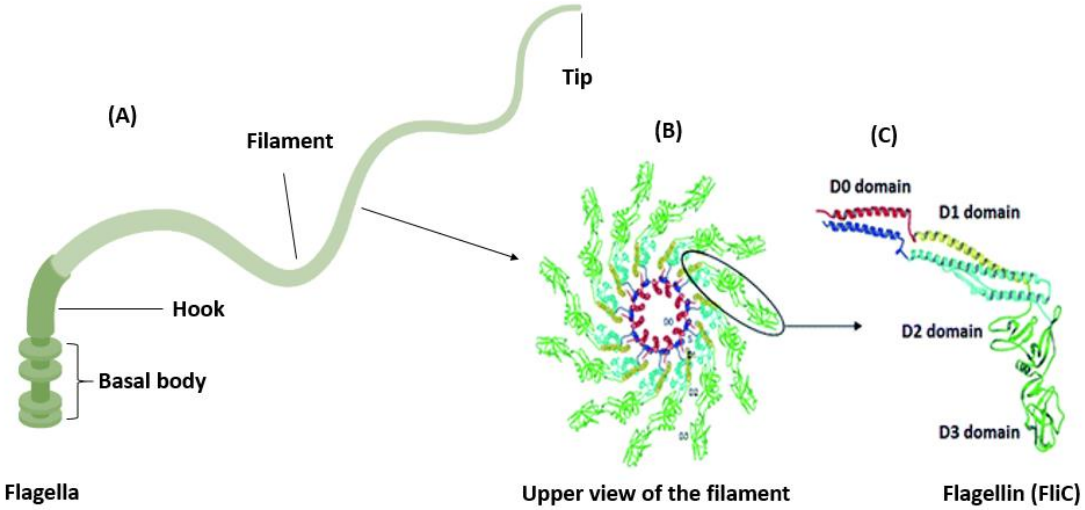



**Figure 1-5: Structural overview of a Gram-negative flagellum**

**(A)** Structural components of a flagellum as indicated. **(B)** Cross-section of the filament revealing individual flagellin subunits. **(C)** Structural domains of flagellin [153].

Structurally, each flagellin consists of highly conserved N- and C-terminal helical domains denoted D0 and D1 that interact to form the filament core (red/blue and cyan/yellow in Figure 1-5B and C). By contrast, the central domains D2 and D3 in terms of amino acid sequence create the outer surface of the filament, are highly variable in sequence between different bacteria, and determine the flagellin H antigen serotyping [154]. The flagellum participates in bacterial virulence by functioning as an adhesin, by facilitating bacterial motility or by regulating the production of other virulence factors [155, 156]. Pathogenic *Listeria monocytogenes*, *Vibrio anguillarum* and *Helicobacter pyroli*, for instance, require flagellum-mediated motility to successfully colonize and invade a host [157-159].

Apart from motility, ETEC flagella also down-regulate $NF_kB$-dependent signalling and hence the host innate immune response [160, 161] and promote attachment to intestinal epithelial cells to initiate biofilm formation [162, 163].

An N-terminal fragment of EtpA was experimentally found to be sufficient to mediate ETEC adhesion to host cells [13]. EtpA is thus an important virulence protein in ETEC. Previous studies on EtpA concentrated on its functional relevance. Here biophysical and structural techniques were used to further characterise EtpA with respect to production efficiency, protein stability, its overall fold, and to link its various functions to different structural regions. By investigating the structure of EtpA an attempt was made to characterise its interaction with its binding partners flagellin and glycans for ETEC adherence and toxin delivery as a possible prelude to designing novel microproteins useful for downline vaccine development.

# 2 Aim

The overarching aim of this thesis was to investigate the structural and biophysical properties of EtpA from ETEC as well as its interaction with flagellin.

## 2.1 Objectives

- Design truncated fragments of full-length EtpA for improved stability.
- Amplify and clone *etpA* gene fragments into suitable production vectors.
- Optimize production of EtpA fragments in *E. coli*.
- Develop purification strategies for EtpA fragments.
- Characterise EtpA fragments using biophysical techniques.
- Crystallize and refine crystal structures of EtpA fragments.
- Clone the *flic* gene and produce and purify ETEC flagellin (FliC).
- Investigate the interaction of various EtpA fragment with FliC.

# 3 Materials and methods

## 3.1 Materials

### 3.1.1 Chemicals

Unless otherwise stated, chemicals and reagents used in this project were of biological grade and were supplied by Biolabs, Cytiva, Inqaba Biotech, Invitrogen, Merck Millipore, Novagen, Promega, Qiagen, Roche, Stratagene and Thermo Scientific.

**Table 3-1: Enzymes and kits used in the methods outlined below**

| Category | Supplier |
|---|---|
| **Enzymes** | |
| *BamH1* | New England Biolabs, Ipswich, MA, USA |
| *Not1* | |
| *Nde1* | |
| *Xho1* | |
| T4 DNA ligase | |
| Shrimp alkaline phosphatase | |
| Phusion Hot Start II DNA Polymerase | |
| **Kits** | |
| GeneJET Plasmid Mini-Prep Kit | Thermo Fisher Scientific, Waltham, MA, USA |
| GeneJET gel extraction Kit | |
| **Others** | |
| Mucin | Sigma Aldrich, Burlington, MA, USA |
| Galactosamine | |
| Ni-NTA resins | Thermo Fisher Scientific, Waltham, MA, USA |
| Glutathione sepharose resins | Cytiva, Danaher, Washington D.C., USA |
| **Molecular weight standards** | |
| Smart Ladder DNA Marker (200-10000 bp) | Thermo Fisher Scientific, Waltham, MA, USA |
| Unstained Protein Ladder (14.4-116 kDa) | |
| Prestained Protein ladder All blue (10-170 kDa) | |
| Prestained Protein ladder All blue (11-250 kDa) | New England Biolabs, Ipswich, MA, USA |
| **Crystallization screens** | |
| The $(NH_4)_2SO_4$ Suite | Qiagen, Venlo, Netherlands |
| The Compas Suite | |
| The Cryo Suite | |
| Procomplex Suite | |
| The PACT Suite | |
| The PEGs Suite | |

### 3.1.2 Buffers

Unless stated otherwise, buffers used for this study are listed below.

**Table 3-2: Buffers and solutions as well as their composition**

| Buffers/Solutions | Components |
|---|---|
| TAE buffer | 40 mM Tris pH 7.5, 20 mM sodium acetate, 1 mM EDTA, pH adjusted to 8.2 with acetic acid |
| 10 × DNA loading buffer | 70 % (w/v) sucrose, 0.25 % (w/v) bromophenol blue, 0.1 M EDTA |
| Ni-NTA lysis buffer | 20 mM Tris pH 7.9, 0.5 M NaCl, 60 mM imidazole |
| Ni-NTA wash buffer | 20 mM Tris pH 7.9, 0.5 M NaCl, 20 mM imidazole |
| Ni-NTA elution buffer | 20 mM Tris pH 7.9, 0.5 M NaCl, 200 mM imidazole |
| SDS-PAGE lower buffer (4 ×) | 1.5 M Tris pH 8.8 |
| SDS-PAGE upper buffer (4 ×) | 0.5 M Tris pH 6.8 |
| SDS-PAGE running buffer | 25 mM Tris pH 7.9, 192 mM glycine, 0.1 % (w/v) SDS |
| SDS-PAGE sample buffer (2 ×) | 100 mL Tris pH 6.8, 4 % (v/v) SDS, 20 % (v/v) glycerol, 200 mM mercaptoethanol, 0.2 % (w/v) bromophenol blue |
| SDS-PAGE separating gel (15 %) | 7.6 mL 1.5 M Tris pH 6.8, 15 mL acrylamide 30 % (w/v), 300 µL 10 % (w/v) SDS, 10 mL dd$H_2O$, 40µL TEMED, 100 µL 2 % (w/v) APS |
| SDS-PAGE stacking gel (5 %) | 2.5 mL 0.5 M Tris pH 6.8, 1.5 mL acrylamide, 30 % (w/v) SDS, 5.9 mL dd$H_2O$, 15 µL TEMED, 25 µL 25 % (w/v) APS |
| SDS-PAGE staining solution | 0.25 % (w/v) Coomassie Brilliant Blue R-250, 30 % (v/v) ethanol, 10 % (v/v) acetic acid |
| SDS-PAGE destaining solution | 40 % (v/v) ethanol, 10 % (v/v) acetic acid |
| Buffer A for ion-exchange | 25 mM Tris pH 8.0, 25 mM NaCl |
| Buffer B for ion-exchange | 25 mM Tris pH 8.0, 1 M NaCl |

### 3.1.3 Equipment

**Table 3-3: Laboratory equipment**

| Equipment | Model | Supplier |
|---|---|---|
| Autoclave | J.S.D. 400 | HOSPI, Chatsworth, CA, USA |
| Centrifuges | Sorvall Lynx 6000 | Thermo Fisher Scientific, Waltham, MA, USA |
| | Megafuge 8R | |
| | 5417C | Eppendorf, Hamburg, Germany |
| Centrifugal filter units | Amicon Ultra-15 | Merck, Darmstadt, Germany |
| Chromatography columns | Gravity flow column | MoBiTech, Goettingen, Germany |
| | Mono Q 10/100 GL | Cytiva, Washington D.C., USA |
| | Superdex 200 10/300 GL | |
| Circular dichroism spectrometer | Chirascan | Applied Photophysics, Surrey, UK |
| Digital camera | DP11 | Olympus, Tokyo, Japan |
| Electrophoresis chamber | | Bio-Rad, Hercules, CA, USA |
| Electrophoresis power supply | Power Pac Basic | |
| FPLC system | ÄKTA 900, ÄKTA pure | Cytiva (Boston, MA, USA) |
| Laminar flow | Hera safe | Thermo Fisher Scientific, Waltham, MA, USA |

| | | |
|---|---|---|
| Microscopes | Leica DFC320 digital microscope | Keyence Corporation, Osaka, Japan |
| | Zeiss Discovery V8 microscope | Zeiss, Oberkochen, Germany |
| Mosquito nanolitre crystallisation robot | PK243 – 01A | TTP labtech, Cambridge, UK |
| pH meter | CRISON | Lasec, Johannesburg, South Africa |
| Protein concentrator | Vivaspin | Merck, Darmstadt, Germany |
| Roller incubator | Roller Mixer SRT6 | Stuart Scientific, Stone, UK |
| Shaking incubator | Multitron II | Infors, Bottmingen, Switzerland |
| Sonicator | Qsonica | Lasec (Johannesburg, South Africa) |
| Spectrophotometer | Nanodrop ND-1000 spectrophotometer | Thermo Fisher Scientific, Waltham, MA, USA |
| UV transilluminator | Safe view | Cleaver Scientific (Rugby, UK) |
| Water purification system | Milli - Q | Merck, Darmstadt, Germany |
| Zetasizer | Zetasizer lab | Malvern Panalytical, Malvern, UK |
| Thermal cycler | T100 | Bio-Rad, Hercules, CA, USA |
| Gel imager | Gel doc XR | Bio-Rad, Hercules, CA, USA |

## 3.2 General molecular biological techniques

### 3.2.1 Preparation of competent cells

All competent cells used for this study were prepared in-house. Frozen glycerol stock of bacterial cells was streaked out onto an antibiotic-free LB-medium plate as cells contain no plasmids. The plate was incubated overnight at 37°C. A single colony from the LB plate was transferred to a 20 mL starter culture and incubated overnight at 37°C in a shaker incubator. The 20 mL starter culture was transferred to 200 mL LB medium and incubated at 37°C in a shaker incubator until the $OD_{600}$ reached 0.4. The cells were placed on ice for 30 min, with occasional swirling to ensure even cooling. The cooled cells were transferred into ice-cold sterile centrifuge tubes and harvested by centrifugation at 3000 x *g* for 15 min at 4°C. The supernatant was discarded, and each pellet was gently resuspended in 10 mL ice-cold 100 mM $MgCl_2$. All suspensions were combined in one centrifuge tube and the cells were harvested at 2000 x *g* for 15 min at 4°C. The supernatant was discarded, and the pellet was gently resuspended in 40 mL ice-cold 100 mM $CaCl_2$. The suspension was kept on ice for 30 min. The cells were harvested by centrifugation at 2000 x *g* for 15 min at 4°C. The supernatant was discarded, and the pellet was gently resuspended in 20 mL ice-cold 85 mM $CaCl_2$, 15 % (v/v) glycerol. The suspension was transferred to a new sterile centrifuge tube. The cells were harvested by centrifugation at 1000 x *g* for 15 min at 4°C. The supernatant was discarded, and the pellet was resuspended in 3 mL ice-cold 85 mM $CaCl_2$, 15 % (v) glycerol.

Aliquots of 50 µL were made in ice cold sterile 1.5 mL centrifuge tubes, snap freeze with liquid nitrogen and stored at -80°C.

### 3.2.2 Agarose gel electrophoresis

All agarose gels used for DNA analysis for this study were in the range of 0.8 to 1 % (w/v). For a 1 % agarose gel, 1 g agarose was combined with 100 mL TAE buffer. The agarose was melted in a microwave until no gel particles were visible and allowed to cool to 50°C. Five µL SYBR Safe DNA stain were added, the mix was poured into an electrophoresis tray, and a comb inserted immediately. After polymerization at room temperature for ~30 min, the comb was removed, and the gel transferred to an electrophoresis tank with 700 mL TAE buffer. For DNA samples, 5 µL of DNA loading dye was mixed with 25 µL sample. Samples were loaded into wells alongside a DNA reference marker and separated at 100 V, 300 mA for 1 to 2 h. DNA was visualized and recorded using a Bio-Rad gel doc XR imaging system.

### 3.2.3 Sodium dodecyl sulphate-polyacrylamide gel electrophoresis (SDS-PAGE)

Manually cast SDS-PAGE between 8 and 10 % (w/v) prepared from SDS-PAGE components listed in Table 3-2 were used for all proteins visualization methods. A 25 µL of protein sample was mixed with 25 µL (2 x) SDS sample loading dye, heated at 95°C for 5 min to denature any proteins, and transferred to a well of an SDS-PAGE gel, in a tank filled with 1 x SDS running buffer (Table 3-2). Samples were separated at 40 mA for 20 to 30 min. The gel was stained in a Coomassie Brilliant Blue R-250 solution (Table 3-2) before destaining to remove excess stain and recorded using a Bio-Rad gel doc XR imaging system.

### 3.2.4 Protein concentration

Proteins were concentrated for downstream experiments by centrifugation using Vivaspin concentrators with a molecular weight cut-off size depending on the size of the protein. The concentration of proteins was determined by measuring the absorbance at 280 nm using a Nanodrop ND-1000 spectrophotometer.

## 3.3 Plasmids and DNA based techniques

### 3.3.1 Plasmids

**A range of plasmids were used to transform cells and ensure high-yield protein production.**

**Table 3-4: Plasmids used in experiments are described in more detail below.**

| Plasmid | Size (Kb) | Antibiotic selection | Tag | Position | Source |
|---|---|---|---|---|---|
| pET28a | 5.3 | $Kan^R$ | $His_6$ | N/C-terminal | Cytiva, Marlborough, MA, USA |
| pGEX-6P-2 | 4.9 | $Amp^R$ | GST | N-terminal | |
| pBAD/ $EtpBA^{1-606}$-Myc-His A | 5.9 | $Amp^R$ | $His_6$ | C-terminal | Prof James Fleckenstein |
| pGEX-6P-2-$EtpA^{67-447}$ | 6.0 | $Amp^R$ | GST | N-terminal | This work |
| pGEX-6P-2-$EtpA^{448-930}$ | 6.3 | $Amp^R$ | GST | N-terminal | Gene Universal, Newark, DE, USA |
| pGEX-6P-2-$EtpA^{67-930}$ | 7.4 | $Amp^R$ | GST | N-terminal | This work |
| pGEX-6P-2-ETEC-FliC | 6.6 | $Amp^R$ | GST | N-terminal | Gene Universal |
| pET28a-ETEC-FliC | 7.1 | $Kan^R$ | $His_6$ | N-terminal | This work |

#### 3.3.1.1 The pGEX-6P-2 plasmid

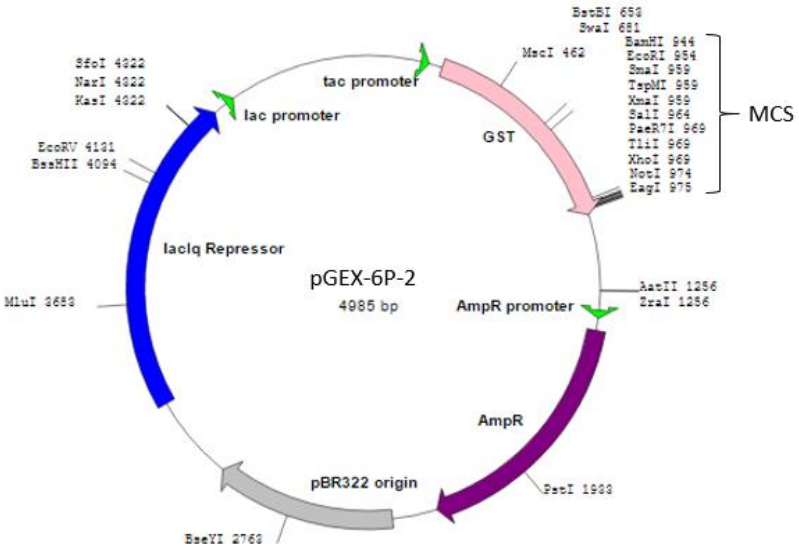



**Figure 3-1: Plasmid map for pGEX-6P-2**

The GST tag, the multiple cloning site (MCS), the tac promoter and the ampicillin resistance gene (AmpR) are all indicated.

Source: Addgene.

The plasmid pGEX-6P-2 is a bacterial expression vector that encodes an N-terminal glutathione S-transferase (GST, pink gene in Figure 3-1) for the production, purification and detection of GST-tagged fusion proteins in *E. coli*. The gene construct of interest may be inserted into the multicloning site (MCS). The vector includes an ampicillin resistance gene

for selection. Expression is under the control of the *tac* promoter induced by isopropyl β-D-thiogalactoside (IPTG). GST is a highly soluble, 26 kDa protein dimer with nanomolar affinity for reduced glutathione so that it rapidly binds to glutathione covalently linked to a sepharose matrix in glutathione sepharose (GS). GST fusion proteins bound to GS may be eluted using reduced glutathione buffers. The method therefore provides a one-step, non-denaturing purification of GST-fusion proteins. The GST tag may be removed by proteolytically cleaving linker between GST and the target protein with C3 protease.

#### 3.3.1.2 The pET28a plasmid

The plasmid pET28a (Figure 3-2) is a bacterial expression vector that encodes an N-terminal $His_6$-tag as well as an optional C-terminal $His_6$-tag for the production, purification and detection of $His_6$-tagged fusion proteins in *E. coli*. Expression is under the control of the T7 *lac* promoter induced by IPTG. A kanamycin resistance gene encoded by the vector allows for the selection of plasmid-bearing cells.

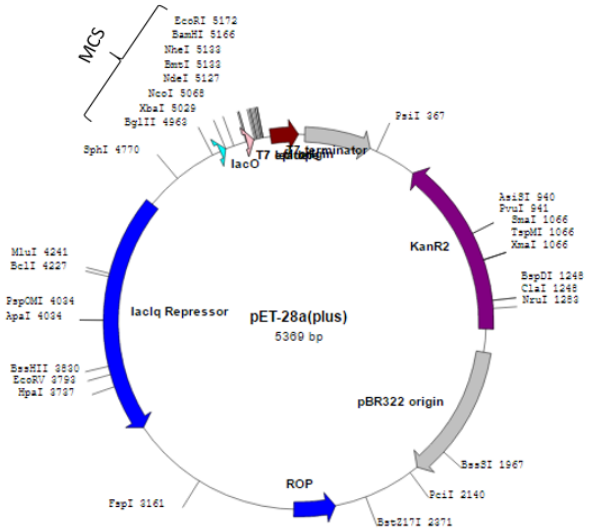



**Figure 3-2: Plasmid map for pET28a**

The T7 promoter, the MCS, and the kanamycin resistance gene (KanR2) are all indicated.

Source: Addgene

The high affinity of histidine for transition metal ions such as $Ni^{2+}$ and $Co^{2+}$ provides the basis for the protein purification in particular if the metal ions are linked to a matrix via chelating groups. $His_6$-tagged proteins may be eluted with high concentrations of imidazole, the functional group of histidine. If required, an N-terminal thrombin cleavage site allows the $His_6$-tag to be physically removed from the protein of interest.

### 3.3.1.3 The pBAD/Myc-His A plasmid

The plasmid pBAD/Myc-His A (Figure 3-3) is an expression vector derived from pBR322 for controlled, dose-dependent protein production in *E. coli*. Expression is controlled by the ARA promoter induced by L-arabinose and tightly regulated by AraC. A *Myc* epitope tag allows fusion proteins to be detected by anti-*Myc* antibodies. An ampicillin resistance gene provides for plasmid selection. A C-terminal $His_6$-tag provides a metal ion binding site for affinity purification of recombinant fusion proteins.

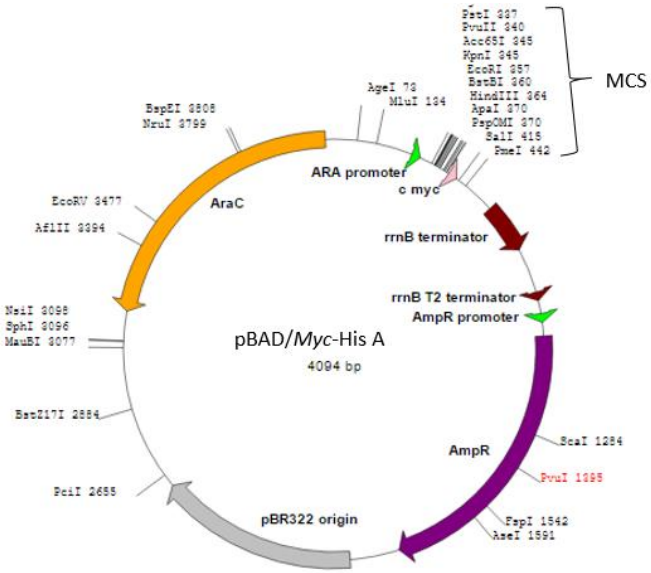



**Figure 3-3: Plasmid map for** pBAD/Myc-His A

The ARA promoter, the MCS and the AmpR gene are all indicated.

Source: Addgene

The pBAD/ *EtpBA*$^{1\text{-}606}$-*Myc*-His A construction to produce EtpA$^{1\text{-}606}$-Myc-His A was generated by inserting the joint *etpB* / *etpA*$^{1\text{-}606}$ encoding gene construct between the Xho1 and Hind111 restriction sites (provided by Prof James M. Fleckenstein, Washington University, St Louis, MO, USA). The *etpB* gene that encodes the transporter protein EtpB for EtpA is located upstream of the *EtpA*$^{1\text{-}606}$ gene. The joint production of both proteins allows the secretion of EtpA$^{1\text{-}606}$ into the extracellular environment.

## 3.3.2 Polymerase chain reaction amplification

Polymerase chain reaction (PCR) allows the specific and rapid amplification of selected DNA fragments for cloning, identification or quantification [164]. The process is typically controlled by temperature in a thermal cycler. During the denaturation step, a high temperature separates the complementary strands of the target DNA. For the annealing or hybridization

step the temperature is lowered to allow designed primers to bind to complementary regions of the target DNA. Finally, the temperature is raised again, to allow DNA polymerase to extend the annealed primers by adding nucleotides to the growing DNA strand [165]. The three steps are cyclically repeated doubling the DNA copy number for every cycle. Resulting products may be visualized by intercalation dyes such as SYBR safe or by labelling either PCR primers or terminating nucleotides with fluorescent dyes [166]. Agarose gel electrophoresis may be used to separate PCR products by size and charge for analysis or downstream use.

In this study, all PCR reactions were performed using Phusion Hot Start II DNA polymerase (New England Biolabs) on a thermal cycler with primers as specified below. Sybr safe dye was used for analysis and visualisation by agarose gel electrophoresis.

### 3.3.3 Restriction enzyme digestion

To manipulate DNA, restriction enzymes are used to recognize and cleave mostly palindromic nucleotide sequences in double-stranded DNA producing blunt or overhanging ends[167]. In this study, plasmid DNA or PCR products were digested using restriction endonucleases and resulting products visualized with Sybr Safe DNA stain by agarose gel electrophoresis. Generally 1 to 2 µg of DNA was transferred to an Eppendorf tube on ice, followed by 5 µL 10 × restriction buffer and 1 µL of each restriction enzyme. After adding dd$H_2O$ to a final volume of 50 µL, the tube was incubated at 37°C for 1 h and inactivated at 80°C for 20 min. The samples were analysed by agarose gel electrophoresis.

### 3.3.4 Ligation and cloning

DNA ligase is an enzyme that creates a phosphodiester bond between adjacent 3'-hydroxyl and 5'-phosphate termini to link a nicked DNA backbone [168]. Ligation is essential in molecular cloning where restriction digested plasmids and/or DNA fragment are each restriction digested before being allowed to anneal. In this study, ATP-dependent bacteriophage T4 DNA ligase was used. Digested plasmid and DNA insert were mixed in a 3:1 molar ratio, whereupon 2 µL 10 × ligase buffer, 1 µL T4 DNA ligase, 2 µL shrimp alkaline phosphatase (rSAP) and dd$H_2O$ were added for a final volume of 20 µL. The reaction was incubated at room temperature for 1 h, heat-inactivated at 65°C for 10 min, chilled on ice and finally used to transform *E. coli*.

### 3.3.5 Transformation and plating of *E. coli*

Bacteria are transformed by taking up foreign DNA or plasmid [169]. Bacterial transformation is a critical step in molecular biology as it allows a wide range of genetic studies. Bacteria may be induced to take up DNA by chemical, electrical or physical stresses [170]. Transformation of *E. coli* in this study invariably involved a heat shock using chemically competent cells (Section 3.2.1). 50 µL of competent *E. coli* cells thawed on ice and 2 µL plasmid DNA were combined in an Eppendorf tube. The tube was incubated on ice for 30 min, followed by a heat shock at 42°C for 60 s and placed back on ice for 2 min. 950 µL LB-medium was added to the cells and the mix incubated for 1 h at 37°C in a thermo-shaker. Transformed cells were spread on a LB-medium plate with the relevant antibiotic to selection cells that have taken up the intended plasmid. The inverted plate was incubated overnight at 37°C.

### 3.3.6 Isolation and purification of plasmid

For all plasmid related experiments, it was necessary to isolate and purify plasmid DNA from other cell debris from *E. coli* cells of interest. A single freshly transformed *E. coli* colony was picked and transferred to 5 mL LB-medium with the required antibiotic. The culture was allowed to growth at 37°C overnight in a shaker incubator. QIAGEN mini prep kit was used to isolate and purify plasmid DNA, following manufacture guidelines.

### 3.3.7 Site directed mutagenesis

Site directed mutagenesis (SDM) is a PCR-based procedure to insert, delete or replace individual nucleotides within a stretch of DNA [171]. SDM is often used to induce specific changes in a particular protein. It requires a set of complementary primers that are also complementary to the target DNA, except for the modified bases near the middle of both forward and reverse primers. During each amplification cycle, a primer anneals to the template and is extended to the full length of the plasmid, incorporating the desired mutation. As the modified plasmid serves as a template in the next round of amplification, many copies of the modified DNA are produced. To remove the original, unmodified plasmid, restriction endonuclease Dpn1 is added as it recognizes and cleaves Dam methylated DNA isolated from *E. coli*. The Dpn1 digested PCR product is used to transform *E. coli* cells, and plasmid re-isolated to confirm the intended modifications. For this study, SDM was used to remove a *BamH1* restriction site. The PCR reaction product was treated with 4 µL 10 × Dpn1 buffer and 1 µL Dpn1 enzyme for 1 h at 37°C before being used to transform *E. coli* cells. The

success of the procedure was checked by Sanger Sequencing and the correct plasmid used to produce the target protein.

### 3.3.8 Sanger Sequencing

Sanger sequencing is a PCR-based technique to identity or sequence DNA *in vitro* [172]. Fluorescently labelled, 2', 3'-dideoxynucleoside triphosphates (ddNTPs) are added to the normal 2'-deoxynucleoside triphosphates (dNTPs) in a low proportion. Individual DNA strands being elongated by DNA polymerase are terminated by the random incorporation of ddNTPs producing fragments colour-coded for the terminal nucleotide. By electrophoretically separating the fragment mix in a capillary, a sequence of colour signals will indicate the actual sequence of nucleotides in the original DNA [173].

All gene constructs generated for this project were verified by Sanger sequencing at the Sanger sequencing facility, University of Pretoria. Sequencing primers specific for pET28a and pGEX plasmid constructs were used (Table 3-5). Volumes and concentrations of reagents for the sequencing PCR are listed in Table 3-6.

**Table 3-5: Primers for Sanger sequencing experiments**

| Primer name | Sequence |
|---|---|
| T7 forward | 5'-TAATACGACTCACTATAGGG-3' |
| T7 reverse | 5'-GCTAGTTATTGCTCAGCGG-3' |
| pGEX forward | 5'-GGGCTGGCAAGCCACGTTTGGTG-3' |
| pGEX reverse | 5'-CCGGGAGCTGCATGTGTCAGAGG-3' |

**Table 3-6: Sanger sequencing PCR reaction mixture**

| Reagent | Concentration | Volume (µL) |
|---|---|---|
| Big dye | - | 1 |
| Sequencing buffer (5 x) | 0.5 x | 2 |
| Primer | 3.2 pmol | |
| Plasmid | 50-100 ng | 2 |
| Distilled water | | To 10 µL final volume |

To prevent the primers and Taq polymerase from impacting the capillary electrophoresis they were removed by ethanol precipitation. To this end, 3 µL 3M sodium acetate, 62.5 µL ethanol

and 15µL distilled water were added to the vial with the sequencing product mix. The vial was centrifuged at 15 000 x *g* for 30 min using a Thermo Fisher Scientific 5417C centrifuge. The supernatant was discarded, and the precipitated DNA was washed twice with 70% ethanol. The samples were air-dried for 20 min in the laminar flow and later analysed at the University of Pretoria Sanger sequencing facility.

## 3.4 Project specific procedures

A range of *etpA* constructs (Figure 3-4) were prepared in different plasmids to test the solubility and stability of corresponding EtpA fragments and to establish a purification strategy for each. Individual steps will be described in the sections below.

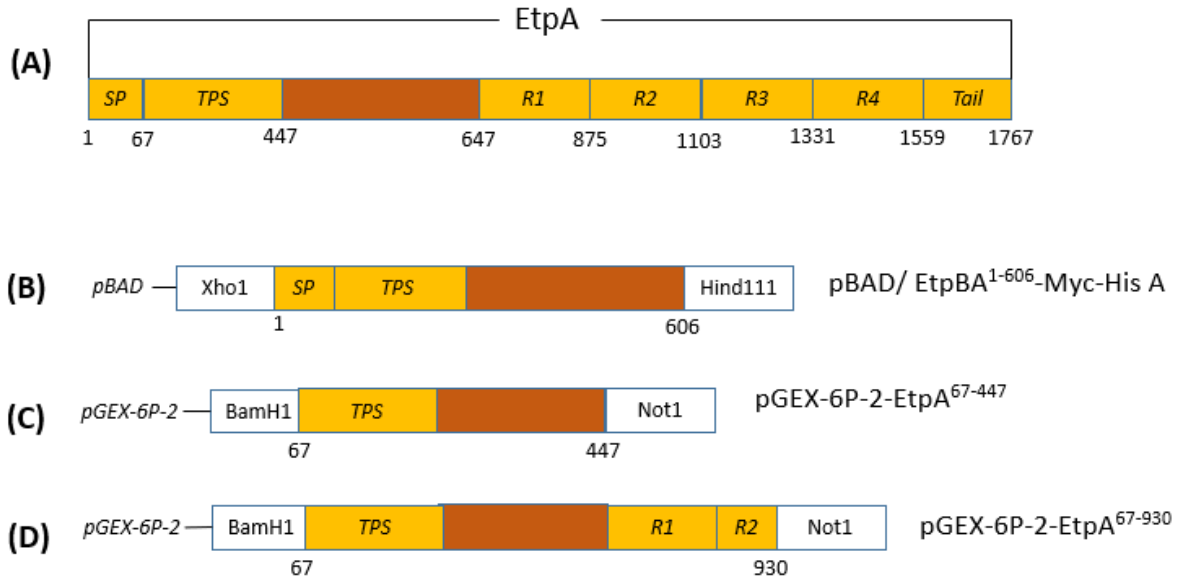



**Figure 3-4: EtpA constructs design**

**A)** Full-length EtpA annotation indicating the start and end of each region. **B)** A pBAD/ *EtpBA*$^{1-606}$*-Myc*-His A construct to produce EtpA$^{1-606}$-Myc-His A was generated by inserting the joint *etpB* / *etpA*$^{1-606}$ encoding gene construct between the Xho1 and Hind111 restriction sites (provided by Prof James M. Fleckenstein, Washington University, St Louis, MO, USA). **C)** A pGEX-6P-2-EtpA$^{67-447}$ gene fragment encoding residues 67-447 with BamH1 and Not1 restriction sites was PCR amplified from the pBAD/ *EtpBA*$^{1-606}$*-Myc*-His A construct and ligated into the BamH1 and Not1 sites of pGEX-6P-2 plasmid. **D)** A pGEX-6P-2-EtpA$^{67-930}$ plasmid encoding residues 67-930 constructed from pGEX-6P-2-EtpA$^{67-447}$ and pGEX-6P-2-EtpA$^{448-930}$ plasmids. The EtpA$^{448-930}$ included residues 448-647, the whole repeat 1 and the first 55 residues of repeat 2.

### 3.4.1 Cloning of *EtpA*$^{67-447}$ gene into pGEX-6P-2 vector

The *etpA*$^{67-447}$ gene fragment coding for residues 67 to 447 of EtpA was PCR amplified from a pBAD/EtpA$^{1-606}$-Myc-His-A plasmid (James M. Fleckenstein, Washington University, St Louis, MO, USA) using the following primers:

**Forward primer:** 5'-TCTCCGGGATCCCCCACTCGTACTTTCCCCGG-3'

**Reverse primer:** 5'-CACGATGCGGCCGCCGCGGATGCACTGACCGT-3'

The products of the PCR reaction were analysed by agarose gel electrophoresis to confirm the correct amplification of the *etpA*$^{67-447}$ fragment. As forward and reverse primers contained *BamH1* and *Not1* restriction sites (underlined), both the *etpA*$^{67-447}$ fragment and pGEX-6P-2 vector were *BamH1* and *Not1* restriction digested before ligation (Table 3-7).

**Table 3-7: Amplifying *etpA*$^{67-447}$ gene fragment by PCR**

| **Component** | **Stock concentration** | | **Amount (Total 50 µL)** |
|---|---|---|---|
| Phusion HF buffer | 5 x | | 10 µL |
| dNTPs | 10 mM | | 1 µL |
| Forward primer | 10 µM | | 2.5 µL |
| Reverse primer | 10 µM | | 2.5 µL |
| Plasmid DNA | 43 ng/µL | | 1 µL |
| Phusion Hot Start II DNA Polymerase | 2 U/µL | | 0.5 µL |
| dd$H_2O$ | | | 32.5 µL |
| | | | |
| **PCR steps** | **Temperature** | **Duration** | **Cycles** |
| Initial denaturation | 98°C | 30 s | 1 |
| Denaturation | 98°C | 10 s | |
| Annealing | 63°C | 30 s | 30 |
| Extension | 72°C | 60 s | |
| Final extension | 72°C | 10 min | 1 |
| | 4°C | hold | |
| **Product digest** | | | **(50 µL Total)** |
| NEB Buffer (CutSmart) | 10 x | | 5 µL |
| PCR product | 200 ng/µL | | 10 µL |
| Water | | | 33 µL |
| *BamH1* | 20000 U/mL | | 1 µL |
| *Not1* | 10000 U/mL | | 1 µL |

The *etpA*$^{67-447}$ gene fragment was ligated into pGEX-6P-2 vector and transformed into *E. coli* DH10 competent cells. A 3:1 molar ratio of insert to vector ligation is shown below (Table 3-8).

**Table 3-8: Ligation**

| **Component** | **Concentration** | **Amount (Total 20 µL)** |
|---|---|---|
| T4 DNA ligase reaction buffer | 10 x | 2 µL |
| pGEX-6P-2 (4.9 kb) | 45 ng/µL | 1 µL |
| Insert (1.1 kb) | 93 ng/µL | 0.3 µL |
| rSAP | 1000 U/mL | 2 µL |
| dd$H_2O$ | | 13.7 µL |
| T4 DNA ligase | 400000 U/mL | 1 µL |

To confirm the successful ligation and incorporation of the *etpA*$^{67-447}$ fragment with the correct sequence, a single colony of the freshly transformed *E. coli* DH10 cells was transferred to 5 mL LB-medium with 0.1 mg/mL final ampicillin and allowed to grow at 37°C overnight in

a shaker incubator. Isolated plasmid DNA was purified, digested with *BamH1* and *Not1* and samples analysed by agarose gel electrophoresis (Section 3.3.6, 3.3.3, 3.3.4 and 3.2.2).

### 3.4.2 Construction of pGEX-6P-2-EtpA$^{67-930}$ vector

The pGEX-6P-2-EtpA$^{67-930}$ vector containing *the etpA$^{67-930}$* gene fragment (designed from full-length *EtpA* gene) coding for EtpA$^{67-930}$ was constructed using both pGEX-6P-2-*EtpA$^{67-447}$* (this work) and pGEX-6P-2-*EtpA$^{448-930}$* (Gene Universal) vectors (Figure 3-4**Error! Reference source not found.** and Figure 3-5). The *EtpA$^{448-930}$* fragment which also includes the whole of repeat 1 and the first 55 residues of repeat 2 was synthesized and cloned between *BamH1* and *Not1* restriction sites. The *EtpA$^{67-447}$* gene was amplified by PCR from pGEX-6P-2-*EtpA$^{67-447}$* with primer having *BamH1* restriction endonuclease on the forward and reverse primers.

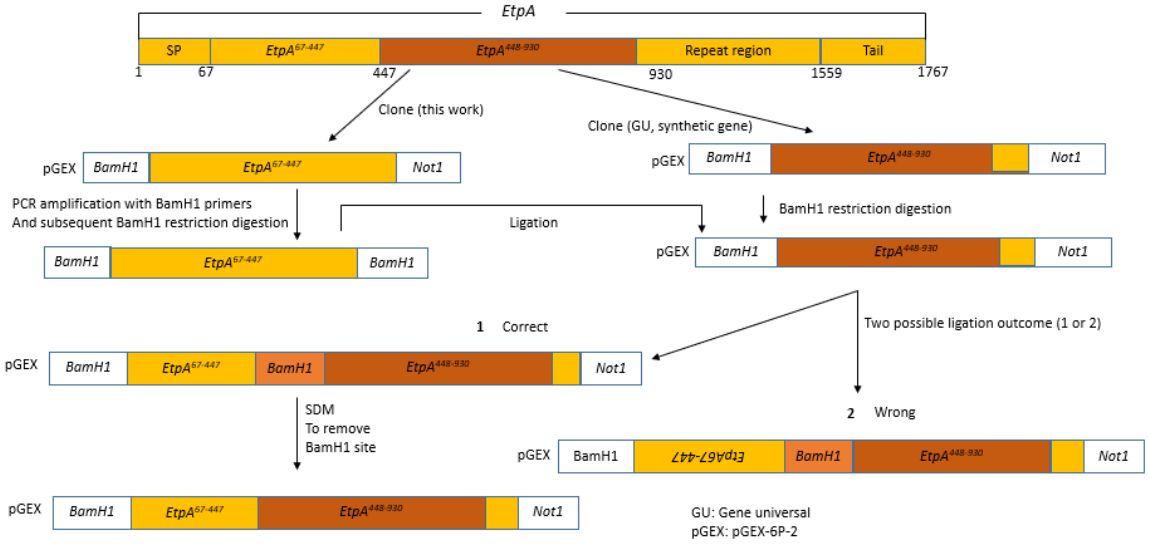



**Figure 3-5: EtpA$^{67-930}$ construct design**

An *etpA* gene fragment encoding residues 67-447 with *BamH1* restriction sites at either end was PCR amplified, *BamH1* digested and ligated into the *BamH1* site of the pGEX-6P-2-EtpA$^{448-930}$ plasmid. The BamH1 digested *etpA$^{67-447}$* PCR product was ligated into similarly digested pGEX-6P-2-EtpA$^{68-930}$ in a random orientation. The pGEX-6P-2-EtpA$^{67-930}$ plasmid with the correct *etpA$^{67-447}$* gene orientation were selected by Sanger sequencing. The additional *BamH1* site between the EtpA$^{67-447}$ and EtpA$^{448-930}$ gene fragments was removed by site-directed mutagenesis.

The primers and PCR parameters are shown below and in Table 3-9. Both PCR amplified *etpA$^{67-447}$* gene fragment and the pGEX-6P-2-*EtpA$^{448-930}$* vector were *BamH1* digested (Section 3.3.3) before ligation.

**Forward primer:** 5'-TCTCCGGGATCCCCCACTCGTACTTTCCCCGG-3'

**Reverse primer:** 5'-CACGATGGATCCCGCGGATGCACTGACCGT-3'

**Table 3-9: PCR procedures for pGEX-6P-2-etpA$^{448-93}$**

| Component | Concentration | Amount (Total 50 µL) | |
|---|---|---|---|
| Phusion HF buffer | 5 x | 10 µL | |
| dNTPs | 10 mM | 1 µL | |
| Forward primer | 10 µM | 2.5 µL | |
| Reverse primer | 10 µM | 2.5 µL | |
| Plasmid DNA | 43 ng/µL | 1 µL | |
| Phusion Hot Start II DNA polymerase | 2 U/µL | 0.5 µL | |
| dd$H_2O$ | | 32.5 µL | |
| **Cycle step** | **Temperature** | **Duration** | **cycles** |
| Initial denaturation | 98°C | 30 s | 1 |
| Denaturation | 98°C | 10 s | |
| Annealing | 63°C | 30 s | 30 |
| Extension | 72°C | 60 s | |
| Final extension | 72°C | 10 min | 1 |
| | 4°C | hold | |

The BamH1 digested *EtpA*$^{67\text{-}447}$ gene fragment was ligated into the pGEX-6P-2-EtpA$^{448\text{-}930}$ vector in a random orientation as indicated in Table 3-8 using 1 µL of 101 ng/µL vector and 1.6 µL of 32 ng/µL insert. The ligation reaction was used to transform *E. coli* DH10 competent cells and spread on a plate with 0.1 mg/mL ampicillin using sterile beads.

To confirm the success of the ligation and the correctness of the *EtpA*$^{67\text{-}447}$ gene, the plasmid was isolated, purified, and sequenced. After ligation of EtpA$^{67\text{-}447}$ gene fragment into pGEX-6P-2-*EtpA*$^{448\text{-}930}$ vector, an additional *BamH1* restriction site was produced, linking *EtpA*$^{67\text{-}447}$ and *EtpA*$^{448\text{-}930}$ genes fragments. The additional *BamH1* restriction site was removed by site-directed mutagenesis (SDM) (Section 3.3.7).

### 3.4.3 Removal of *BamH1* restriction site by site-directed mutagenesis

To remove the additional *BamH1* restriction site linking *EtpA*$^{67\text{-}447}$ and *EtpA*$^{448\text{-}930}$ genes fragments SDM was undertaken by PCR using the set of primers below. The PCR setup and parameters are indicated in the Table 3-10 PCR amplification was followed by digestion with Dpn1 enzyme. The mixture was used to transform *E. coli* DH5α cells and spread on plate with 0.1 mg/mL ampicillin. To confirm the success of SDM, plasmid DNA was isolated, purified, and sequenced (Section 3.3.8).

**Forward primer:** 5'-GTACGGTCAGTGCATCCGCGAAGAAAGGCAAGGCGGGTAAC-3'

**Reverse primer:** 5'-TTACCCGCCTTGCCTTTCTTCGCGGATGCACTGACCGTACC-3'

**Table 3-10: PCR for site-directed mutagenesis**

| Component | Stock concentration | Amount (Total volume 50µL) |
|---|---|---|
| Phusion HF buffer | 5 x | 10 µL |

| | | | |
|---|---|---|---|
| dNTPs | 10 mM | 1 µL | |
| Forward primer | 10 µM | 2.5 µL | |
| Reverse primer | 10 µM | 2.5 µL | |
| Plasmid DNA | 43 ng/µL | 1 µL | |
| Phusion Hot Start II DNA Polymerase | 2 U/µL | 0.5 µL | |
| $ddH_2O$ | | 32.5 µL | |
| | | | |
| **Cycle step** | **Temperature** | **Duration** | **cycles** |
| Initial denaturation | 98°C | 30 s | 1 |
| Denaturation | 98°C | 10 s | |
| Annealing | 70°C | 30 s | 30 |
| Extension | 72°C | 4 min | |
| Final extension | 72°C | 10 min | 1 |
| | 4°C | hold | |

### 3.4.4 Cloning of $^{ETEC}$*fliC* into pGEX-6P-2 vector

Flagellin (*fliC*) gene coding the FliC protein from ETEC was amplified (Table 3-11) from a pUC57-$^{ETEC}$*fliC* plasmid DNA purchased from Gene Universal (Newark, DE, USA) and cloned into pGEX-6P-2 vector between *BamH1* and *Not1* restriction endonuclease using primers and PCR set up below. The PCR reaction was analysed by agarose gel electrophoresis to confirm the amplification of the *fliC* gene. The primers used were:

**Forward primer:** 5'-TCGACGGATCCATGGCACAAGTCATTAATACCAAC-3'

**Reverse primer:** 5'-CATAGGCGGCCGCTTAACCCTGCAGCAGAGAC-3'

**Table 3-11: PCR procedures for pGEX-6P-2-*fliC***

| **Component** | **Stock concentration** | | **Amount (Total 50 µL)** |
|---|---|---|---|
| Phusion HF buffer | 5x | | 10 µL |
| dNTPs | 10 mM | | 1 µL |
| Forward primer | 10 µM | | 2.5 µL |
| Reverse primer | 10 µM | | 2.5 µL |
| Plasmid DNA | 35 ng/µL | | 1 µL |
| Phusion Hot Start II DNA polymerase | 2 U/µL | | 0.5 µL |
| $ddH_2O$ | | | 32.5 µL |
| | | | |
| **PCR steps** | **Temperature** | **Duration** | **cycles** |
| Initial denaturation | 98°C | 30 s | 1 |
| Denaturation | 98°C | 10 s | |
| Annealing | 63°C | 30 s | 30 |
| Extension | 72°C | 60 s | |
| Final extension | 72°C | 10 min | 1 |
| | 4°C | hold | |
| **Product digest** | | | **(Total 50 µL)** |
| NEB Buffer (CutSmart) | 10 | | 5 µL |
| PCR product | 225 ng/µL | | 10 µL |
| $ddH_2O$ | | | 33 µL |
| *BamH1* | 20000 U/mL | | 1 µL |
| *Not1* | 10000 U/mL | | 1 µL |

The PCR product and pGEX-6P-2 vector were digested separately with *BamH1* and *Not1* restriction endonuclease before ligation. The same procedure was used for pGEX-6P-2 digestion with 15 µL of 76 ng/µL plasmid used in a final volume of 50 µL.

The *fliC* gene was ligated into pGEX-6P-2 vector. A 3:1 molar ratio of insert to vector setup was used as outlined in Table 3-78 using 1.5 µL of 45 ng/µL vector (67.5 ng) plus 0.3 µL of 93 ng/µL insert (27.9 ng). The ligation mix was used to transform *E. coli* DH10 competent cells, and the transformed cells were spread on a 0.1 mg/mL ampicillin containing plate.

To confirm the incorporation of the *fliC* gene in the plasmid, the latter was isolated from the *E. coli* DH10 cells and double digested with *BamH1* and *Not1* restriction endonucleases. The presence of the *fliC* gene insert was confirmed by sequencing at the University of Pretoria Sanger sequencing facility.

### 3.4.5 Cloning of $^{ETEC}$*fliC* gene into pET28a vector

Flagellin (*fliC*) gene coding for $^{ETEC}$FliC protein was amplified (Table 3-12) from pGEX-6P-2-$^{ETEC}$*fliC* and cloned into pET28a vector between *Nde1* and *Xho1* restriction endonuclease using primers and PCR set up below. The PCR reaction was analysed by agarose gel electrophoresis to confirm the amplification the of *fliC* gene.

**Forward primer:** 5'-TCGACCATATGATGGCACAAGTCATTAATACCAAC-3'

**Reverse primer:** 5'-CATAGCTCGAGTTAACCCTGCAGCAGAGAC-3'

**Table 3-12: PCR procedures for pET28a-*fliC***

| **Component** | **Stock concentration** | | **Amount (Total 50 µL)** |
|---|---|---|---|
| Phusion HF buffer | 5 x | | 10 µL |
| dNTPs | 10 mM | | 1 µL |
| Forward primer | 10 µM | | 2.5 µL |
| Reverse primer | 10 µM | | 2.5 µL |
| Plasmid DNA | 35 ng/µL | | 1 µL |
| Phusion Hot Start II DNA Polymerase | 2 U/µL | | 0.5 µL |
| dd$H_2O$ | | | 32.5 µL |
| | | | |
| **PCR steps** | **Temperature** | **Duration** | **Cycles** |
| Initial denaturation | 98°C | 30 s | 1 |
| Denaturation | 98°C | 10 s | |
| Annealing | 63°C | 30 s | 30 |
| Extension | 72°C | 60 s | |
| Final extension | 72°C | 10 min | 1 |
| | 4°C | hold | |
| **Product digest** | | | **(Total 50 µL)** |
| NEB Buffer (CutSmart) | 10 x | | 5 µL |
| PCR product | 200 ng/µL | | 10 µL |
| Water | | | 33 µL |
| *BamH1* | 20000 U/mL | | 1 µL |

| | | |
|---|---|---|
| *Not1* | 10000 U/mL | 1 µL |

The PCR product and pET28a vector were digested separately with *Nde1* and *Xho1* restriction endonucleases before ligation. The same procedure was used to for pET28a where 20 µL of 76 ng/µL plasmid was used.

The $^{ETEC}$*fliC* gene was ligated into pET28a vector as outlined in Table 3-8 using 1.5 µL of 49 ng/µL plasmid and 0.3 µL of 95 ng/µL insert.

The ligation mix (5 µL) was used to transform *E. coli* DH10 competent cells. The transformed cells (50 µL) were spread on to kanamycin antibiotic (0.05 mg/mL final concentration) containing plate.

To confirm the success of the ligation and the presence of the *fliC* gene in the plasmid, the latter was isolated, purified and double digested with *Nde1* and *Xho1* restriction endonucleases. The presence of the *fliC* gene insert was confirmed by sequencing at the University of Pretoria sequencing facility, South Africa.

## 3.5 Protein production, purification and analysis

### 3.5.1 Production and purification of 3C protease

The 3C protease used to cleave GST-tag from GST-fusion proteins was produced and purified in-house. A starter culture was prepared by inoculating 20 mL LB medium with 0.1 mg/mL ampicillin from a glycerol stock of *E. coli* Bl21 cells harbouring pGEX-6P-2-*3C* protease plasmid and grown overnight at 37°C in a shaker incubator. The 20 mL culture was transferred to 1 L LB-medium with 0.1 mg/mL ampicillin and grown at 37°C and 180 rpm to an $OD_{600}$ of 0.6. Protein production was induced with 0.1 mM IPTG and production allowed to continue at 25°C overnight with 180 rpm shaking. Before and after induction, 1 mL cell culture was collected to analyse target protein induction. The cells were pelleted by centrifugation at 12 000 x *g* for 2 min and the pellet resuspended in SDS sample buffer to lyse the cells where the volume of buffer was proportional to the culture $OD_{600}$. Following overnight incubation, the bacterial cell culture was centrifuged for 15 min at 5200 x *g* and 4°C in a Sorvall Lynx 6000 centrifuge using an F9-6X1000 LEX rotor. The supernatant was discarded, and the cell pellet resuspended in 20 mL PBS at 4°C. The cells were lysed by sonication on ice for eight 30 s on/off cycles using a Qsonica sonicator. The lysate was centrifuged for 1 h at 16000 x *g* and 4°C in a Sorvall LYNX 6000 centrifuge using an F21-8X50Y rotor to separate the soluble protein fraction

from cell debris and inclusion bodies. The soluble supernatant and insoluble pellet were collected. 100 µL 8 M urea was added to the pellet fraction to solubilise insoluble proteins and all resulting samples analysed by SDS-PAGE.

For affinity purification, the soluble supernatant was mixed with glutathione sepharose beads and incubated with agitation for 1 h at 4°C. Unbound proteins were eluted with PBS in a gravity flow column. The GST-3C fusion protein was eluted from the column with 50 mM Tris pH 8.0, 80 mM NaCl, and 20 mM reduced glutathione. Production and purity of GST-3C protease was confirmed on SDS-PAGE. The protein was dialysed against 20 mM Tris pH 8.0, 150 mM NaCl, 1 mM EDTA and 20 % glycerol. Aliquots were made and stored at -80°C.

### 3.5.2 Production and purification of EtpA$^{1-606}$

To produce EtpA$^{1-606}$ protein, the sequenced pBAD/EtpBA$^{1-606}$-Myc-His A plasmid with the correct *EtpBA*$^{1-606}$ sequence was used to transform *E. coli* TOP10 cells. A single colony was transferred to 20 mL LB-medium supplemented with 0.1 mg/mL ampicillin, grown overnight at 37°C and 160 rpm. The culture was transferred to 0.4 L LB-medium with 0.1 mg/mL ampicillin and incubated at 37°C with agitation at 180 rpm. Production of EtpA$^{1-606}$-Myc-His$_6$ protein was induced at an $OD_{600}$ of 0.5 with 0.002 % (w/v) arabinose with shaking at 22°C for 18 h. Before and after induction, 0.2 mL of the cell culture was collected, pelleted by centrifugation at 12000 x *g* for 2 min using a 5417C centrifuge. Secretion of EtpBA$^{1-606}$ into the supernatant was investigated by SDS-PAGE. Following overnight incubation, bacterial cell culture was centrifuged for 20 min at 10000 x *g* and 4°C in a Sorvall Lynx 6000 centrifuge with an F9-6X1000 LEX rotor. The supernatant containing secreted proteins was collected, treated with 0.5 mM PMSF and transferred to 50 mL tubes. The supernatant was mixed with pre-equilibrated Ni-NTA beads and incubated at 4°C for 1 h with mild agitation. In a gravity flow column, unbound proteins were eluted with 20 mM Tris pH 7.9, 200 mM NaCl. Target protein EtpA$^{1-606}$-Myc-His$_6$ was eluted with 20 mM Tris pH 7.9, 500 mM NaCl and 200 mM imidazole. Production and purity of EtpA$^{1-606}$-Myc-His$_6$ was assessed by SDS-PAGE.

### 3.5.3 Production and purification of EtpA$^{67-447}$

A single colony of *E. coli* BL21 cells transformed with pGEX-6P-2-EtpA$^{67-447}$ plasmid was transferred to 20 mL LB-medium with 0.1 mg/mL ampicillin and incubated overnight at 37°C and 160 rpm and transferred to 1 L LB-medium with 0.1 mg/mL ampicillin and grown at 37°C and 180 rpm to an $OD_{600}$ of 0.8. Protein production was induced with 0.1 mM IPTG and

continued at 22°C with 180 rpm shaking overnight. Before and after induction, 1 mL of the cell culture was collected, and pelleted by centrifugation at 12000 x *g* for 2 min using a 5417C centrifuge. The pellets were suspended in an equivalent amount of SDS sample buffer to lyse the cells depending on the $OD_{600}$ to analyse the induction of the target protein. Following overnight incubation, the bacterial cell culture was centrifuged for 15 min at 5200 x *g* and 4°C in a Sorvall Lynx 6000 centrifuge using an F9-6X1000 LEX rotor. The supernatant was discarded and the cell pellet resuspended in 20 mL PBS per 1 L culture at 4°C. The cells were lysed by sonication on ice for seven 30 s on/off cycles using a Qsonica sonicator. The lysate was centrifuged at 16000 x *g* and 4°C for 1 h in a Sorvall LYNX 6000 centrifuge with a F21-8X50Y rotor to separate the soluble protein from the cell debris and inclusion bodies. The soluble supernatant and insoluble pellet were separated and 100 µL 8 M urea added to the pellet fraction to solubilise insoluble proteins. Samples of individual steps were analysed by SDS-PAGE.

For affinity purification, soluble supernatant was mixed with pre-equilibrated glutathione sepharose (GS) beads and incubated with agitation for 1 h at 4°C. Unbound proteins were eluted with PBS in a gravity flow column. The GST-$EtpA^{67-477}$ fusion protein was cleaved by adding 0.4 mg 3C protease and incubating for 24 h at 4°C with agitation. Released $EtpA^{67-447}$ was eluted with PBS. Production, purity and successful fusion protein cleavage were assessed by SDS-PAGE. The protein was dialysed against 20 mM Tris pH 8.0, 20 mM NaCl for further purification by ion exchange chromatography (IEC).

For IEC, a 5 mL HiTrap Q HP column was connected on an Äkta pure chromatography system. The column was equilibrated with 5 column volumes (CV) of buffer A (20 mM NaCl in 20 mM Tris pH 8.0) at a flow rate of 1 mL/min. The dialysed protein was loaded onto the column where the target protein binds due to its $Hi_6$-tag. Unbound proteins were eluted with 3 CV of buffer A. Bound proteins were eluted with a linear gradient of 5 to 1000 mM NaCl in 20 mM Tris pH 8.0. Samples from the eluted fractions were analysed by SDS-PAGE.

Size exclusion chromatography (SEC) was used to further purify and to determine the elution profile of $EtpA^{67-447}$. A 24 mL Enrich SEC 650 (10/300) column was equilibrated with 2 CV of 20 mM NaCl in 20 mM Tris pH 7.6 at a flow rate of 0.5 mL/min. 5 mg of $EtpA^{67-447}$ protein was loaded and eluted at a flow rate of 0.5 mL/min. Eluted fractions were analysed by SDS-PAGE.

### 3.5.4 Production of GST-EtpA$^{67-930}$

The pGEX-6P-2-*EtpA*$^{67-930}$ plasmid was used to transform *E. coli BL21* cells for production of GST-EtpA$^{67-930}$ fusion protein.

#### 3.5.4.1 Optimizing temperature and IPTG concentration

To optimize GST- EtpA$^{67-930}$ production, three temperatures, 37, 30, and 22°C, and two IPTG concentrations, 0.1 and 0.5 mM, were tested. A single colony harbouring the pGEX-6-P1-*EtpA*$^{67-930}$ plasmid was used to inoculate 20 mL LB-medium with 0.1 mg/ml ampicillin and incubated overnight at 37°C and agitation at 160 rpm. A 1 mL of the culture was transferred to 150 mL LB-medium supplemented with 0.1 mg/ml ampicillin, grown at 37°C and 180 rpm to an $OD_{600}$ of 0.6. The culture was transferred into two 50 mL flasks. Protein production was induced with 0.1 and 0.5 mM IPTG, respectively, and incubated at 37°C with 180 rpm shaking overnight. Before and after induction, 1 mL cell culture was collected, pelleted by centrifugation at 12000 x *g* for 2 min using a 5417C centrifuge. To compare the induction levels of the target protein, the pellets were suspended in equivalent volumes of SDS sample buffer to lyse the cells depending on the $OD_{600}$. Following overnight incubation, bacterial cell cultures were centrifuged at 5200 x *g* and 4°C for 15 min in a Sorvall Lynx 6000 centrifuge with an F9-6X1000 LEX rotor. The supernatants were discarded, and the cell pellets resuspended in 8 mL PBS per 50 mL culture at 4°C. The cells were lysed by sonication on ice for four 30 s on/off cycles using a Qsonica sonicator. The lysate was centrifuged at 16000 x *g* and 4°C for 30 min using a Sorvall LYNX 6000 centrifuge with the F21-8X50Y rotor to separate the soluble protein fraction from cell debris and inclusion bodies. The soluble supernatant and insoluble pellet fractions were separated. 500 µL 8 M urea was added to the pellet fraction to solubilise insoluble proteins. Samples from critical fractions were analysed by SDS-PAGE. The procedure was repeated for 22 and 30°C, and for the two different IPTG concentrations.

### 3.5.5 Solubilisation of GST-EtpA$^{67-930}$ inclusion bodies

When multiple recombinant proteins are expressed at high levels in *E. coli*, highly aggregated proteins known as inclusion bodies are formed [174]. The desire protein is frequently expressed at high translational rate when high temperatures, high inducer concentrations and robust promoter systems are used during protein expression. A significant obstacle to the production and purification of recombinant proteins using *E. coli* as the host is the formation of inclusion bodies [175, 176]. To yield the active proteins, inclusion bodies require a lengthy

processing procedure that includes cell isolation, solubilisation, refolding, and purification [177]. Solubilisation of inclusion bodies and refolding of the solubilized protein into its native state are the most important steps in the recovery of active protein from inclusion bodies. Inclusion bodies from *E. coli* have been isolated and solubilised successfully to produce active proteins using a number of isolation and solubilisation methods [178, 179]. Isolation methods include the use of mechanical cell disrupture such as sonication and chemical disruption techniques such lysozyme [180]. High and mild concentration of denaturants and chaotropes such as urea and guanidine hydrochloride have been used to solubilize inclusion bodies [175, 181]. Use of high pH buffer in combination with 2 M urea have been used successfully for solubilisation of inclusion bodies [182, 183].

Here urea was used to solubilize GST-EtpA$^{67\text{-}930}$ protein from inclusion bodies after attempts to produce soluble protein proved unsuccessful. Inclusion bodies containing GST-EtpA$^{67\text{-}930}$ from a 50 mL induced culture were resuspended in 40 mL PBS pH 8.0 with 2 M urea. The sample was allowed to freeze at -20°C overnight and thaw the next morning. The sample was centrifuged for 1 h at 16000 x *g* and 4°C in a Sorvall LYNX 6000 centrifuge with a F21-8X50Y rotor to separate the solubilized protein fraction and the insoluble remnants. Half of the solubilised protein fraction was dialysed overnight against PBS pH 8.0 to remove the urea and potentially regain natively folded protein. The second half was diluted 10-fold with PBS pH 8.0. GS beads were added, and the mix incubated for 1 h at 4°C with gentle agitation. Unbound proteins were eluted with PBS in a gravity flow column. To confirm that any GST-EtpA$^{67\text{-}930}$ protein bound to the GS beads, a 50 µL sample was collected and analysed by SDS-PAGE.

### 3.5.6 Production and purification of His$_6$-$^{ETEC}$FliC protein.

A single colony of *E. coli* BL21 cells transformed with pET28a-$^{ETEC}$*fliC* plasmid was transferred to 20 mL LB-medium with 0.05 mg/mL kanamycin and permitted to grow overnight at 37°C and 160 rpm. The culture was transferred to 1 L LB-medium with 0.05 mg/mL kanamycin and grown at 37°C and 180 rpm to an $OD_{600}$ of 0.8. Protein production was induced with 0.1 mM IPTG and continued at 22°C with 180 rpm shaking overnight. Before and after induction, 1 mL of the cell culture was collected, pelleted by centrifugation at 12000 x *g* for 2 min using a 5417C centrifuge. To assess the production of target protein, the pellet was suspended in a volume of SDS sample buffer equivalent to the $OD_{600}$ and incubated overnight. The bacterial

cell culture was centrifuged for 15 min at 5200 x *g* and 4°C in a Sorvall Lynx 6000 centrifuge with an F9-6X1000 LEX rotor. The supernatant was discarded, and the cell pellet resuspended in 20 mL lysis buffer (20 mM Tris pH 7.9, 400 mM NaCl, 20 mM imidazole) per 1 L culture at 4°C. The cells were lysed by sonication on ice for seven 30 s on/off cycles using a Qsonica sonicator. To separate the soluble protein from the cell debris and any inclusion bodies, the lysate was centrifuged for 1 h at 16000 x *g* and 4°C in a Sorvall LYNX 6000 centrifuge with a F21-8X50Y rotor. The soluble supernatant and insoluble pellet were separated. 100 µL 8 M urea was added to the pellet fraction to solubilise membrane proteins.

For affinity purification, the soluble supernatant was mixed with Ni-NTA beads and incubated with agitation for 1 h at 4°C. Unbound proteins were eluted with lysis buffer in a gravity flow column. Fused $His_6$-$^{ETEC}$FliC protein was eluted from the Ni-NTA column with elution buffer (20 mM Tris pH 7.9, 500 mM NaCl, 200 mM imidazole). Production and purity of $His_6$-$^{ETEC}$FliC protein was confirmed by SDS-PAGE. The protein was dialysed against buffer A (20 mM Tris pH 7.9, 20 mM NaCl) for downstream experiments.

### 3.5.7 Production and purification of GST-$^{ETEC}$FliC

A single colony was used to inoculate 20 mL LB-medium with 0.1 mg/mL ampicillin and grown overnight at 37°C and 160 rpm. The culture was transferred to 1 L LB-medium with 0.1 mg/mL ampicillin and grown at 37°C and 180 rpm to an $OD_{600}$ of 0.8. Protein production was induced with 0.1 mM IPTG and continued at 22°C with 180 rpm shaking overnight. Before and after induction, 1 mL of the cell culture was collected, pelleted by centrifugation at 12 000 x *g* for 2 min using a 5417C centrifuge. The pellets were suspended in an equivalent amount of SDS sample buffer to lyse the cells depending on the $OD_{600}$ to analyse induction of target protein. Following overnight incubation, bacterial cell culture was centrifuged for 15 min at 5200 x *g* and 4°C in a Sorvall Lynx 6000 centrifuge with an F9-6X1000 LEX rotor. The supernatant was discarded, and the cell pellet resuspended in 20 mL PBS per 1 L culture at 4°C. The cells were lysed by sonication on ice for seven 30 s on/off cycles using a sonicator. The lysate was centrifuged for 1 h at 4°C and 16000 x *g* using the Sorvall LYNX 6000 centrifuge with the F21-8X50Y rotor to separate the soluble protein fraction from cell debris and inclusion bodies. The soluble supernatant was separated from the insoluble pellet. 100 µL 8 M urea was added to the pellet fraction to solubilise aggregated or membrane proteins.

For affinity purification, soluble supernatant was mixed with glutathione sepharose beads and incubated with agitation for 1 h at 4°C. Unbound proteins were eluted with PBS in a gravity flow column. The GST-ETEC-FliC fusion protein was cleaved by adding 0.7 mg of 3C protease and incubating for 24 h at 4°C with agitation. Target $^{ETEC}$FliC protein was eluted with PBS. Production and purity of $^{ETEC}$FliC was confirmed by SDS-PAGE. The protein was dialysed against 20 mM Tris pH 8.0, 20 mM NaCl for further purification with ion exchange chromatography (IEC).

For IEC, a 5 mL HiTrap Q HP column was connected on an Äkta pure chromatography system. The column was equilibrated with 5 column volumes (CV) of 20 mM NaCl in 20 mM Tris pH 8.0 at a flow rate of 1 mL/min and the dialysed protein was loaded. Unbound proteins were eluted with 3 CV of buffer A. Bound proteins were eluted using a linear gradient of 5 to 1000 mM NaCl in 20 mM Tris pH 8.0. Eluted fractions were analysed by SDS-PAGE to check for the success of IEC.

### 3.5.8 Thermal unfolding of EtpA$^{1-606}$

All thermal unfolding studies were carried out in 10 mM $Na_3PO_4$ pH 7.4, 150 mM NaCl. Circular dichroism (CD) spectra were obtained at the CSIR (Pretoria, South Africa) using a Chirascan spectrophotometer in a cuvette of 1 mm path length. Readings were recorded from 180 to 280 nm with scanning steps of 1 nm at a time-per-point of 0.25 s. Four spectra were collected and the average was plotted. A 1 mg/mL of purified EtpA$^{1-606}$ was heated from 20 to 90°C at 10°C intervals in a temperature-controlled compartment. Different absorbance for each denaturation point were plotted against temperature to obtain a melting temperature curve for EtpA$^{1-606}$ protein. The transition mid-temperature (Tm) of EtpA$^{1-606}$ was obtained by a plot of CD data extracted at 222 nm as a function of temperature and fitted by a sigmoidal curve.

### 3.5.9 Urea induced unfolding and refolding of EtpA$^{1-606}$

The CD spectra were used to assess the stability of EtpA$^{1-606}$ incubated with 0 to 9 M urea. Spectra were recorded on a Chirascan spectrophotometer at 20°C between 180 and 280 nm in a 1-mm path-length cuvette at a rate of 0.25 s/nm. Spectra were corrected for the respective protein-free solutions and an averages of four repeat spectra were used for analysis. Buffer A consisting of 10 mM $Na_3PO_4$ pH 7.4 and 150 mM NaCl was used throughout. Sample were prepared by mixing a 1 mg/mL solution of EtpA$^{1-606}$ in buffer A with the

appropriate volume of a 9 M urea stock also in buffer A and incubating for 24 h at 20°C with gentle agitation (Table 3-12: PCR procedures for pET28a-*fliC*). To assess the reversibility of urea-induced unfolding, 1 mL samples of $EtpA^{1-606}$ exposed to the respective urea concentrations for 24 h were diluted 10-fold with buffer A, dialysed against buffer A for 24 h at 20°C with gentle agitation and further urea depleted by cycles of dilution with buffer A followed by concentration to 0.8 mg/mL in an Amicon Ultra-15 concentrator with a 50 kDa molecular weight cut-off. The fraction of unfolded protein was assessed by plotting elipticities at 222 nm against the urea concentration and fitting a sigmoidal curve.

**Table 3-13: Urea unfolding of $EtpA^{1-606}$**

| Urea concentration (M) | Volume of 5 mg/mL protein (µL) | 12 M Urea (µL) | Buffer (µL) |
|---|---|---|---|
| 0 | 400 | 0 | 1600 |
| 1.5 | 400 | 250 | 1350 |
| 3 | 400 | 500 | 1100 |
| 4.5 | 400 | 750 | 850 |
| 6 | 400 | 1000 | 600 |
| 7.5 | 400 | 1250 | 350 |
| 9 | 400 | 1500 | 100 |
| | | | To a final volume of 2 mL |

## 3.6 Structure determination

### 3.6.1 Crystallization, data collection and processing

In a crystallization experiment, the goal is to supersaturate the protein and initiate nucleation and crystal growth. The degree of saturation can determine if a protein will crystallize or not. A fully dissolved protein that is undersaturated will never crystallize. On the other hand, depending on the degree of supersaturation, nucleation or crystal growth may happen. Screening and optimization are the two main processes in the crystallization procedure. The former entails determining the physical, chemical, and biological conditions that might promote crystal growth, even though these conditions might not always produce high calibre crystals for X-ray diffraction. Commercial screens with typical crystallization conditions are available and have been used with greater success. In order to improve the quality of the crystals for X-ray diffraction, conditions that provide leads during the screening phase are further investigated or adjusted during optimization.

With enough and pure proteins in hand, the hanging-drop vapour-diffusion method was used to screen for potential crystallization conditions. Initial screening was done randomly using available commercial crystallization screen kits (Compas, Ammonium sulphate, Cryos, PACT, PEGs and Procomplex Suites; Qiagen, Venlo, Netherlands). To begin with, 24-well plates were grease ringed and labelled to match with the 96 well unique reagents found in each screen kit. While protein was kept on ice, 500 µL of unique screen kit reagent were added into each well. With the unique screen kit reagent into each well, 2 µL of protein was mixed with 2 µL from a well on a cover glass slide for a 1:1 ratio. The cover glass slide was overturned over the well to seal the drop over the well. This process was repeated for all the screen kits reagents used. The plates were stored at 12 and 22°C for to allow for crystallization to hopefully happen. The plates were continually checked under the microscope for any crystal formation. Crystals for X-ray diffraction were cryoprotected in 10 mM MES pH 5.5, 25 mM NaCl, 5 % (v/v) glycerol and cryocooled in liquid nitrogen.

Diffraction data were recorded remotely on an Eiger2 XE 16M detector on beamline i04 at the DIAMOND light source (Oxfordshire, UK). Images were auto processed, scaled and merged using Xia2 XDS program suite [184].

### 3.6.2 Structure determination and refinement

The crystal structure of EtpA$^{67-447}$ was determined by molecular replacement with PHASER-MR in PHENIX [185] using the crystal structure of HMW1A-PP (PDB code 20DL) with 34 % sequence identity as a searched model. The model was re-built and refined by Autobuild in PHENIX. The model was improved manually using WinCoot and refined using PHENIX refine [186]. The refining cycle involves fitting the model's atomic locations and the related B-factors, which are parameters that represent the atom's thermal motion to the observed electron density map. This typically results in a better set of phases. Parameters such as stereochemistry, hydrogen bonds and allocation of bond lengths and angles are essential features used to determine the quantity of the refined model. However, R- and Rfree factors are the primary parameters that need to be managed following each cycle of model improvement. The data resolution affects both factors. The R-factor measures the degree of agreement between the built crystallographic model and the X-ray diffracted data, indicating how well the revised model predicts the measured data. A subset of reflections that were left out of the structure refinement are used to calculate Rfree.

### 3.6.3 Structure modelling with AlphaFold

Full-length theoretical models of EtpA and the four TpsA members FHA (*B. pertussis*) [187], HpmA (*P. mirabilis*) [99] , HMW1 [188, 189] and HxuA (*H. influenza*) (4RM6) were generated using AlphaFold [190]. Most protein structures modelling programs require homologous protein structure to predict the structure of proteins, and in cases where there is no homologous structure, predictions from these programs are usually presented with low atomic accuracy. AlphaFold uses evolutionarily related sequences, multiple sequence alignment and a representation of amino acid residue pairs to predict protein structure with high atomic accuracy without a known homologous protein structure being required [190]. The amino acid sequence of each protein was divided into fragments of ~1200 residues where the full-length protein sequence is longer than 1200 residues. The sequence for each fragment was separately input on to the server to generate the modelled structure. The modelled fragments were then combined to produce a full-length structure using PyMol.

## 3.7 Molecular interaction studies

### 3.7.1 Molecular interaction of $EtpA^{67-447}$ and FliC.

To investigate the interaction of $EtpA^{67-447}$ with FliC, purified $EtpA^{67-447}$ and $His_6$-FliC were both buffer exchanged with buffer A (150 mM NaCl in 20 mM Tris pH 7.4). Approximately 65.5 nmol of $EtpA^{67-447}$ and 130.5 nmol of $His_6$-FliC were mixed in a molar ration of 1:2 and incubated overnight at 4°C. Following overnight incubation, 300 µL of Ni-NTA beads pre-equilibrated with buffer A were added and incubated at 4°C for 1 h with gentle agitation. The sample was poured into a gravity flow column and unbound proteins were eluted. The column was washed with buffer A to remove any unbound proteins. The beads sample was collected, treated with SDS-loading dye and proteins denatured at 95°C for 5 min. The sample was then analysed by SDS-PAGE. The relative amount of $EtpA^{67-447}$ would indicate specific interaction of the two proteins.

### 3.7.2 Molecular interaction of $EtpA^{1-606}$ and FliC

To investigate the interaction of $EtpA^{1-606}$ with FliC, the procedure above was essentially repeated except that $EtpA^{1-606}$-$His_6$ now bound to the column matrix while any retained FliC would indicate the specific interaction of the two proteins.

### 3.7.3 Interaction of EtpA$^{67-447}$ and EtpA$^{1-606}$ with FliC

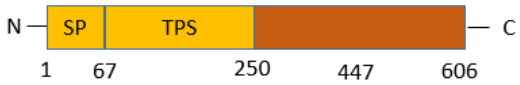
To investigate the interaction of EtpA$^{67-447}$ and FliC by SEC, 65.5 nmol of EtpA$^{67-447}$ was mixed with 130.5 nmol of FliC in a molar ratio of 1:2. The mixture was incubated overnight at 4°C with gentle agitation. Following overnight incubation, the mixture was analysed on SEC650 (10/300) column equilibrated with 150 mM NaCl in 20 mM Tris pH 7.4. Peaks fractions were analysed by SDS-PAGE to deduce their protein composition.

To analyse the interaction of EtpA$^{1-606}$ and FliC with SEC, the procedure outlined in section 3.7.3 above was repeated.

# 4 Results

## 4.1 EtpA$^{1-606}$

Initially EtpA$^{1-606}$, an N-terminal fragment of full-length EtpA, was investigated. This fragment comprises the N-terminal signal peptide, the TPS domain and a part of the C-terminal repeat domain as indicated below.




### 4.1.1 Production and purification

The protein EtpA$^{1-606}$ was produced in *E. coli* using arabinose to induce expression. The plasmid construct also encoded for EtpB, such that EtpA$^{1-606}$ was secreted across both inner and outer cell membranes. The protein was purified by Ni-NTA IMAC and size exclusion chromatography as described in Section 3.5.2. Note that although EtpA$^{1-606}$ would presumably lose its signal peptide during its Sec-dependent secretion across the inner bacterial membrane, the loss of signal peptide was not demonstrated experimentally. Correspondingly the protein is referred to as EtpA$^{1-606}$ throughout, instead of the possibly formally more correct EtpA$^{67-606}$. Induction with both 0.002 and 0.0002 % (w/v) arabinose yielded comparable amounts of protein, visible as faint yet single bands in SDS PAGE analysis of the cell supernatant (Figure 4-1A). The size matches the expected protein size of 63 kDa. As the protein band was not observed in the uninduced sample (Figure 4-1A, uninduced lane) and corresponds to the expected size of the fusion protein EtpA$^{1-606}$-His$_6$, the correct protein appears to be produced. It is secreted across both inner and outer bacterial membranes. Further concentration and purification by Ni-NTA affinity chromatography confirmed the purity of the secreted protein (Figure 4-1B).

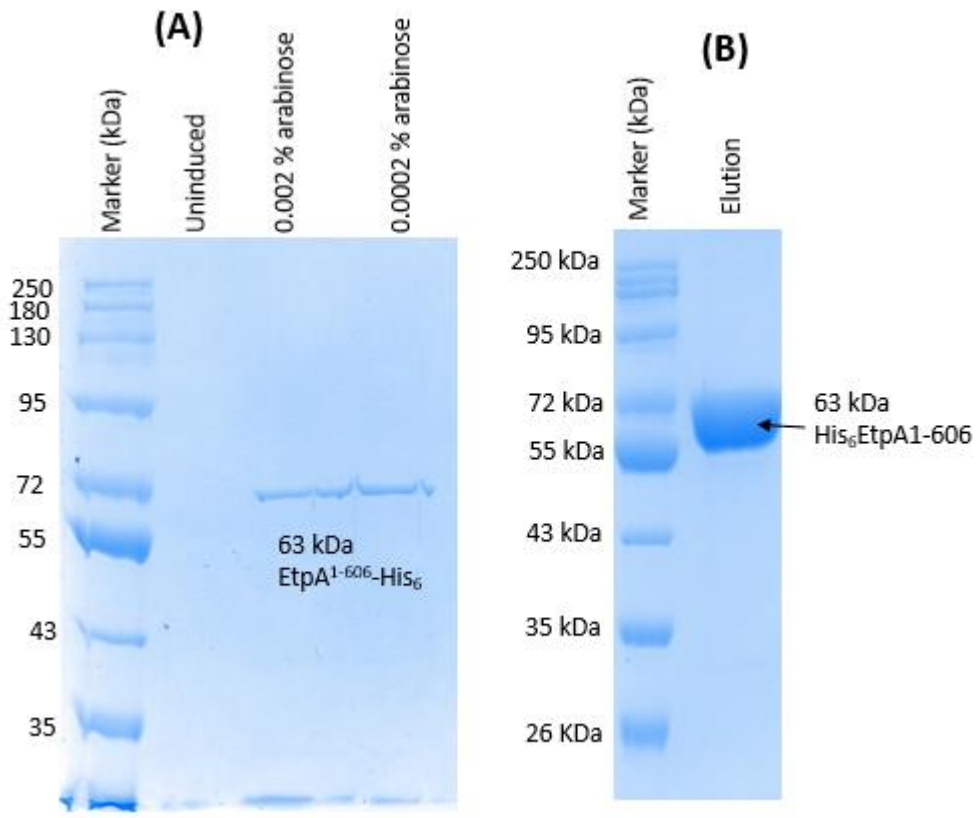



**Figure 4-1: Production and purification of EtpA$^{1\text{-}606}$-His$_6$**

**A)** SDS-PAGE for production of EtpA$^{1\text{-}606}$-His$_6$ protein induced by arabinose at two concentrations.
**B)** SDS-PAGE after Ni-NTA affinity chromatography confirming the presence of EtpA$^{1\text{-}606}$-His$_6$.

### 4.1.2 Size and oligomerisation

To investigate the oligomerisation state of secreted EtpA$^{1\text{-}606}$ appropriate samples were analysed by native PAGE and SEC (Section 3.2.3).

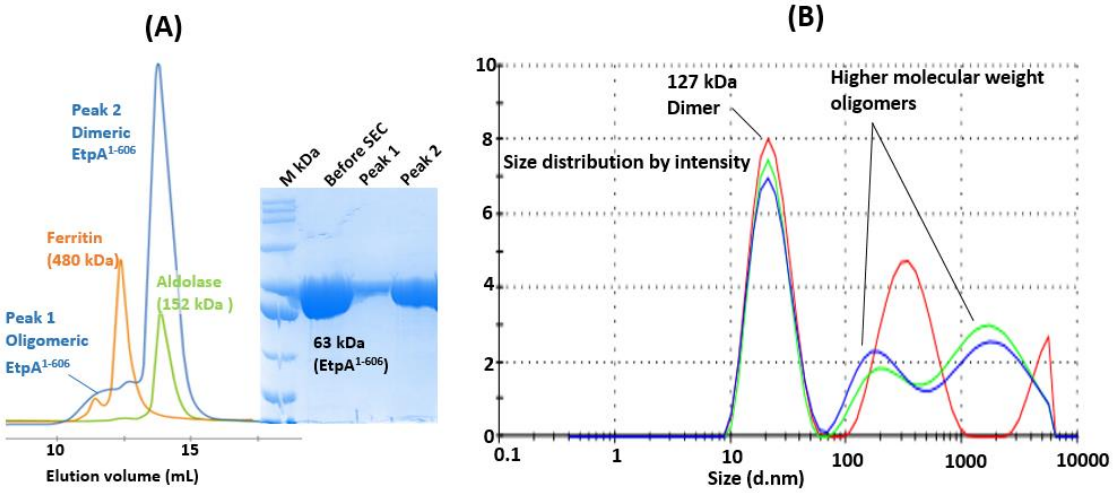



**Figure 4-2: Analysis of EtpA$^{1\text{-}606}$ oligomerisation**

**A)** Main peak for EtpA$^{1-606}$ in size exclusion chromatography corresponds to a dimer. Standards were aldolase (152 kDa, green) and ferritin (480 kDa, orange). SDS-PAGE (insert) of EtpA$^{1-606}$ peak fractions before and after SEC.
**B)** Dynamic light scattering profile for different models confirm EtpA$^{1-606}$ dimers and higher molecular weight oligomers.

Analysis of EtpA$^{1-606}$-His$_6$ by SEC resulted in a single major peak (Peak 2) with a retention volume of 13.8 mL, plus a broad preceding shoulder (Peak 1) with a retention volume of 11.56 mL (blue line in Figure 4-2A). Samples from each fraction analysed by SDS-PAGE (insert) again yielded single bands at 63 kDa. This implies that both peaks correspond to two different oligomerisation states of EtpA$^{1-606}$-His$_6$ (Figure 4-2A). Calibrating the SEC column with 152 kDa aldolase and 480 kDa ferritin, allowed the mass of the protein oligomers to be broadly estimated at ~504 and ~160 kDa, respectively. Dividing these values by 63 kDa implies that peak 1 could correspond to an EtpA$^{1-606}$ octamer whereas peak 2 corresponds most closely to an EtpA$^{1-606}$ dimer. Analysing the size distribution of EtpA$^{1-606}$ oligomers by dynamic light scattering (DLS) yielded three peaks (Figure 4-2B). Again the main peak with a mass of 127 kDa implied a possible EtpA$^{1-606}$ dimer. The remaining two HMW oligomer peaks are significantly larger and imply large and very large aggregates of EtpA$^{1-606}$.

Overall, EtpA$^{1-606}$ was thus successfully produced and purified as a secreted protein. However, size analysis indicates that it mostly forms dimers in addition to higher molecular mass oligomers. Investigation of the distribution over time indicate that the latter accumulate implying that EtpA$^{1-606}$ tends towards non-specific oligomerisation.

Variability in the degrees of oligomerisation is a significant hurdle in crystallising a protein. Proteins should ideally be monodisperse i.e. have a single oligomerisation state that remains constant over days or weeks. Sometimes buffer conditions can be optimized to enhance monodispersity. In the case of EtpA$^{1-606}$ no buffer could be identified that prevented the oligomerisation. As a result, a smaller fragment EtpA$^{67-447}$ was produced by cloning a smaller fragment of the encoding gene into a different plasmid. This will be discussed in Section 4.2 below.

### 4.1.3 Circular dichroism spectroscopy and thermal unfolding of EtpA$^{1-606}$

Circular dichroism (CD) spectroscopy was used to assess the dominant secondary structural elements of EtpA$^{1-606}$ and to study its thermal stability.

The CD spectrum of EtpA$^{1-606}$ revealed a broad minimum around 222 nm (Figure 4-3A). As this spectrum is well defined with distinct maxima and minima implies that the protein is well folded. However, it differs from known spectra of mainly α-helical or β-strand dominated structures (Figure 4-3B). Instead, the curve complies with known CD spectra of β-helical proteins. This interpretation was supported by automated analysis with programs K2D2 and CAPITO [191, 192].

Next the stability of EtpA$^{1-606}$ was investigated by recording CD spectra for protein heated to different temperatures at 10°C intervals (Figure 4-3C). Unexpectedly, the CD spectrum shows only minor changes for temperatures between 20 and 90°C. Clearly EtpA$^{1-606}$ is highly thermostable with the overall fold being maintained throughout this temperature range. Only at a temperature above 90°C does the protein unfold rapidly losing almost all secondary structure (lime-green line). Plotting the ellipticities at 222 nm from individual temperature curves against each temperature (Figure 4-3D) yielded a melting point ($T_m$) of 94°C for EtpA$^{1-606}$ again confirming the high thermostability of this protein.

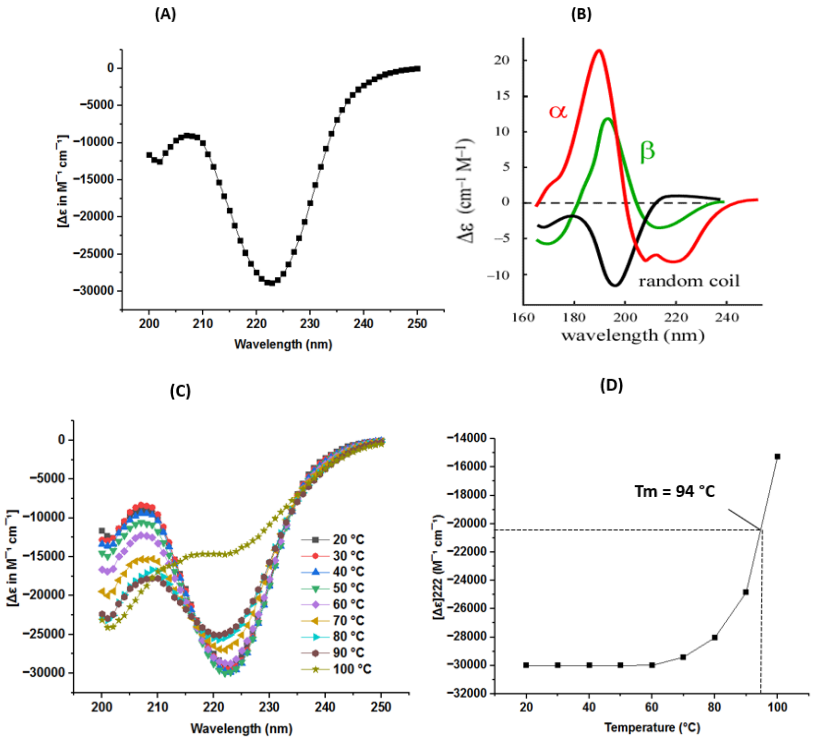



**Figure 4-3: Temperature induced-unfolding of EtpA$^{1-606}$ by CD spectroscopy**

**(A)** CD spectrum for EtpA$^{1-606}$ reveals a broad minimum around 222 nm.
**(B)** Standard CD spectra for proteins.
**(C)** Thermal unfolding data between 20 and 100°C.
**(D)** Plot of ellipticities at 222 nm against temperature yields a $T_m$ of 94°C.

### 4.1.4 Urea induced unfolding and refolding

To assess the stability of EtpA$^{1-606}$ with respect to chemical stresses, CD spectra were recorded for protein incubated in urea concentrations ranging from 0 to 9 M (Figure 4-4). Each panel represents a different urea concentration and in each case, CD spectra for untreated (fully folded, black), urea-treated (unfolded, red) and refolded samples (blue) are compared.

Little to no change is observed for panels A to C of (Figure 4-3) implying that the secondary and tertiary structure of EtpA$^{1-606}$ is not affected by urea concentrations up to 4.5 M urea. In panel D the red curve is observed to have a slightly shallower minimum compared to panels A to C, implying that a slight loss of secondary structure is incurred at a urea concentration of 6 M. Interesting though, the blue curve is again quite similar to the black curve, indicating that the loss of secondary structure documented by the red curve is reversed when the urea is removed. Panels E and F continue this trend with much more extensive loss of local structure being documented by the shallower red curve, through the overall β-helical fold

being maintained as the maxima and minima of the curve only show a small change of wavelength. Again, though, the blue lines resemble the black line meaning that the urea induced unfolding is reversible upon removal of the chaotrope.

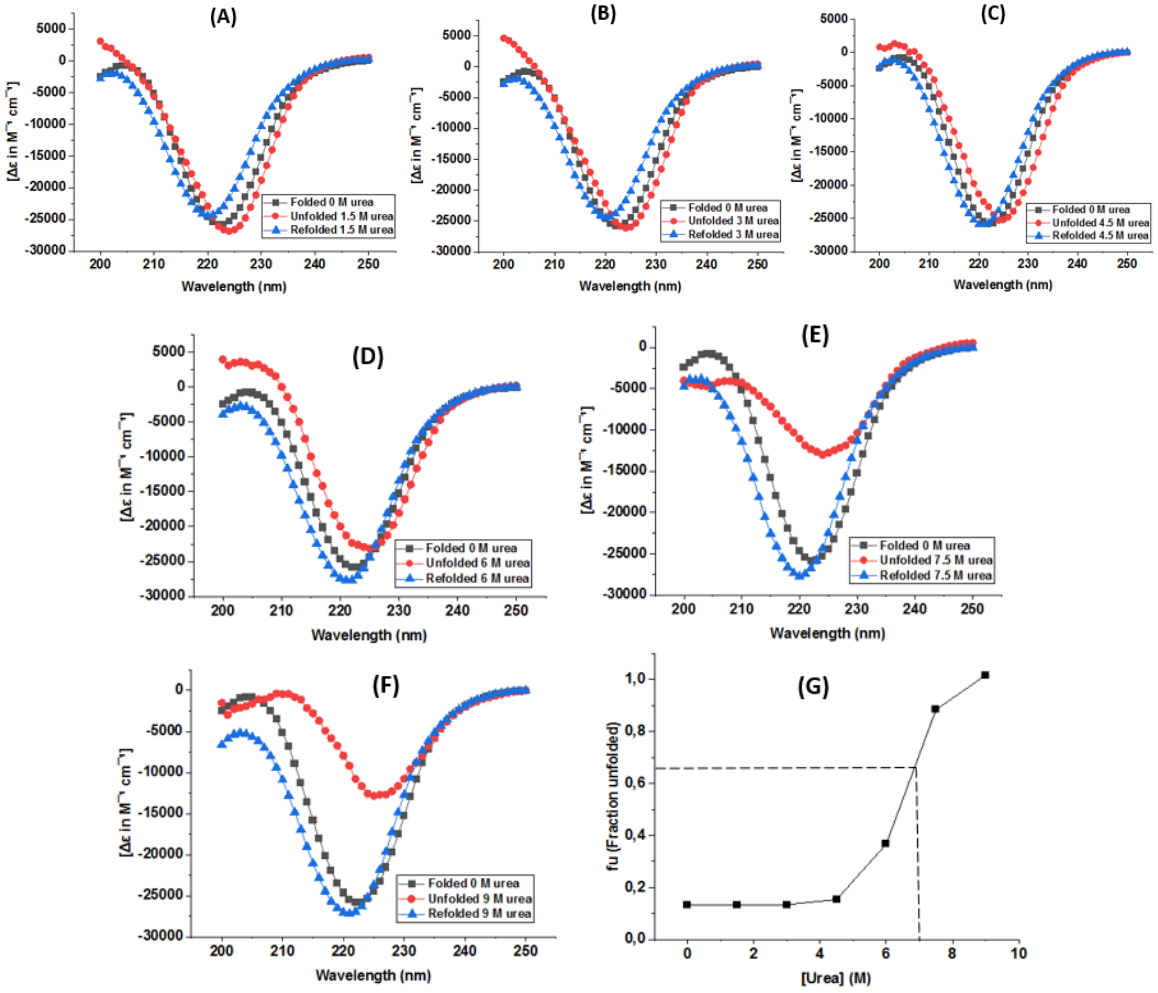


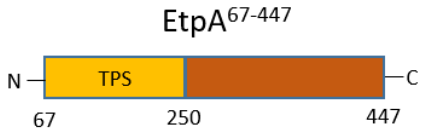

**Figure 4-4: Unfolding and refolding of EtpA$^{1-606}$ induced by urea**

**(A to F)** Superposition of the CD spectra for the native state (0 M urea, black square), unfolded state (red circle), and refolded state (blue triangle) for 1.5 to 9 M urea respectively. **(G)** Fraction of unfolded curve obtained after unfolding the protein.

The fraction of unfolded protein (fu) was determined from each panel for the corresponding urea concentration. Plotting fu against the underlying urea concentration yielded an unfolding curve for EtpA$^{1-606}$ (Figure 4-3G). Its inflection point identifies the critical urea concentration for EtpA$^{1-606}$ to be ~7 M.

## 4.2 EtpA$^{67-447}$

As demonstrated in Section 4.1 above, the protein fragment EtpA$^{1-606}$, an N-terminal fragment of full-length EtpA, could be produced by a transformed laboratory strain of *E. coli*. The protein was secreted into the medium and purified as described. However, this protein proved difficult to work with due to its tendency to aggregate into large oligomers. In its place EtpA$^{67-447}$, a shortened N-terminal fragment of full-length EtpA was designed. Note that this construct is not only shorter than the one encoding EtpA$^{1-606}$. In addition, it also lacks both the *etpB* gene and the N-terminal secretion signal, such that this protein is produced cytoplasmically rather than being secreted.




### 4.2.1 Cloning of *etpA*$^{67-447}$

A DNA fragment encoding EtpA$^{67-447}$ was amplified by PCR and ligated into the plasmid pGEX-6P-2. Transforming *E. coli* DH5α cells with this plasmid, re-isolating the plasmid and digesting with restriction enzymes BamH1 and Not1 yielded a 1.1 kb DNA band (Figure 4-5). The size of the observed band corresponds to the expected size of desired gene fragment, such that this preliminary analysis indicated that the fragment was successfully amplified and cloned as planned.

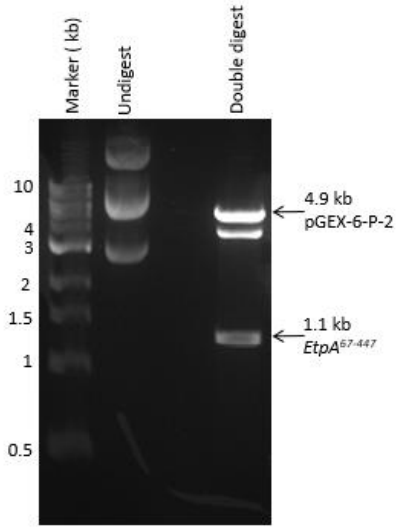



**Figure 4-5: Cloning of fragment *etpA*$^{67-447}$**

Agarose gel of restriction enzyme digest of pGEX-6P-2-*EtpA*$^{67-447}$ plasmid with *BamH1* and *Not1*

### 4.2.2 Production and purification of EtpA$^{67-447}$

*E. coli* Bl21 cells were transformed with the pGEX-6P-2-*EtpA*$^{67-447}$ plasmid to achieve the production of a GST-EtpA$^{67-447}$ fusion protein. A 63 kDa protein band that corresponds to the size of GST-EtpA$^{67-447}$ fusion protein is absent from the uninduced sample in an SDS-PAGE analysis but is observed in the sample after IPTG induction (Figure 4-6A). The 63 kDa protein band is further observed in both the insoluble and soluble fractions (Figure 4-6A), with a more intense band in the soluble fraction implying that the GST-EtpA$^{67-447}$ fusion protein is predominantly produced as a soluble protein, though a small proportion of the protein is sequestered into insoluble inclusion bodies.

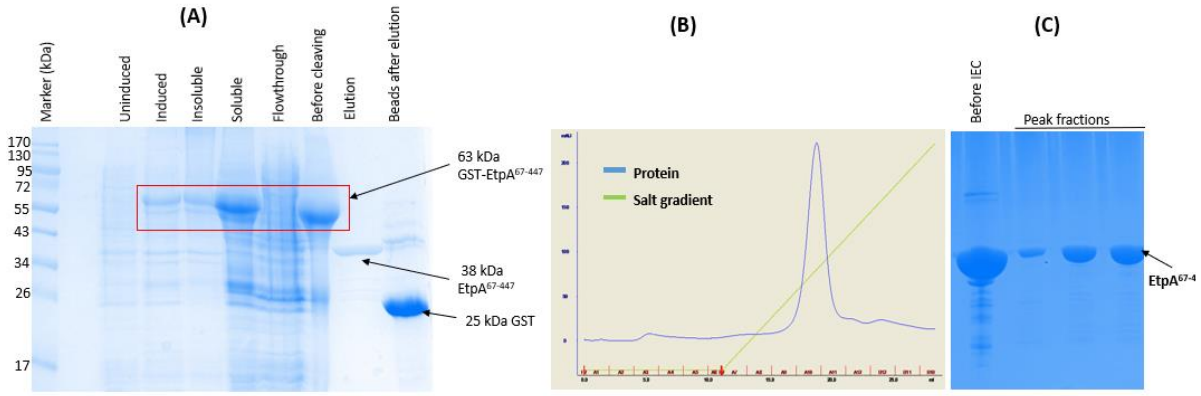



**Figure 4-6: Cloning, production and purification of EtpA$^{67-447}$**

**A)** SDS-PAGE of production and affinity purification by glutathione sepharose as well as 3C cleavage. **B)** Chromatogram for ion exchange purification (IEC). **C)** SDS-PAGE of protein samples obtained from IEC peak.

To purify GST EtpA$^{67-447}$ fusion protein, the soluble fraction was mixed with glutathione sepharose (GS) beads. After incubating, collecting the flow through, and eluting any unbound proteins, the 63 kDa GST-EtpA$^{67-447}$ fusion protein band is observed in the sample denoted "before cleaving" but not in the flow through (Figure 4-6A), implying that it quantitatively bound to the GS beads and was successfully purified by affinity chromatography. 3C protease was used to separate GST from EtpA$^{67447}$. This resulted in a 38 kDa protein being eluted that matches the expected size of EtpA$^{67-447}$. A small amount of GST-EtpA$^{67-447}$ fusion protein and cleaved EtpA$^{67-447}$ remain attached to the beads after elution, alongside a large amount of GST (right-hand lane in Figure 4-6A). 3C protease thus successfully cleaved GST-EtpA$^{67-447}$ fusion protein allowing the untagged EtpA$^{67-447}$ to elute.

As faint protein bands contaminated the EtpA$^{67-447}$ fraction after GS affinity chromatography (Figure 4-6A) and downstream experiments such as crystallization require as pure a protein sample as possible, further purification with ion exchange chromatography (IEC) was attempted. The results indicate the presence of a single peak after IEC (Figure 4-6B). Analysis of protein samples from the peak on SDS-PAGE revealed purer protein bands that correspond to EtpA$^{67-447}$ after IEC (Figure 4-6C).

### 4.2.3 Crystallization, data collection and refinement

Lead crystallization conditions for EtpA$^{67-447}$ at 12°C were investigated using the hanging-drop vapour-diffusion method of the Cryos screen kit (Qiagen) and combining 2 µL reservoir solution with µL EtpA$^{67-447}$ solution (15 mg/mL in 10 mM MES pH 5.5, 25 mM NaCl, 5 % (v/v) glycerol). Condition number seven yielded protein crystals. It combined 0.1 M sodium citrate pH 5.6, 20 % (w/v) PEG 4000, 20 % (v/v) isopropanol. Other crystallization screen kits did not produce any crystals. While lead crystallization conditions normally need to optimized to improve crystal size and quality, this was not required for EtpA$^{67-447}$ as the crystal plates grown within two weeks were sufficiently large, single crystals that did not require further improvement (Figure 4-7A and B). Similar platelets were also obtained for protein concentrations ranging from 13 to 18 mg/mL while keeping other parameters unchanged.

To record diffraction data for EtpA$^{67-447}$, crystals were individually retrieved with a nylon loop, directly plunged into liquid nitrogen and couriered at 100 K to the Diamond Light Source (Oxfordshire, UK). Diffraction images were recorded remotely on beamline i04 using an Eiger2 XE 16M detector. The crystals diffracted X-rays to a resolution of 1.8 Å producing a typical protein diffraction pattern (Figure 4-7C). The data set of 360 diffraction images was auto processed, scaled, and merged using Xia2 XDS program suite [184]. The evaluation of the diffraction data indicated the space group to be triclinic P1, the lowest symmetry space group. This, in part, explains the relatively low completeness of 97.4 %. In addition, it had a multiplicity of 3.4 and an R-merge of 0.097 (Table 4-1), indicating an internally consistent and usable set of structure factor amplitudes.

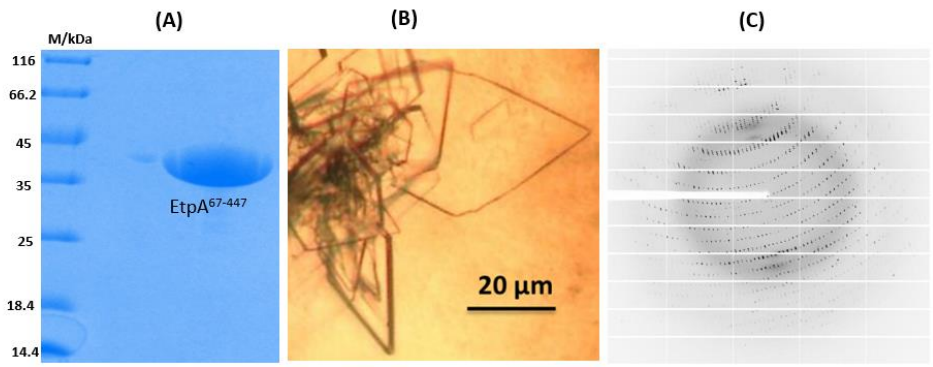



**Figure 4-7: Crystallization and X-ray diffraction of EtpA$^{67-447}$**

**A)** EtpA$^{67-447}$ protein used for crystallization. **B)** Crystals obtained with EtpA$^{67-447}$ protein. **C)** Diffraction pattern from an EtpA$^{67-447}$ crystal.

The crystal structure of EtpA$^{67-447}$was solved using the crystal structure of HMW1A-PP (PDB code 20DL) as a model. The two protein fragments share a sequence identity of 34 %. The molecular replacement calculation was achieved using the PHASER-MR option of PHENIX [185]. Crystal density and solvent content analyses indicated a probable four EtpA$^{67-447}$ molecules per asymmetric unit. A single MR solution with a high log likelihood gain (LLG) value of 9565 and a similarly high translation function Z score (TFZ) of 9.2 was obtained with the expected four molecules per asymmetric unit, implying a likely correct solution. Analysis of the MR solution in Coot indicated a good packing of the molecules both in the asymmetric unit and throughout the crystal, again confirming the probable correctness of the solution.

Rigid body refinement of the initial model in Phenix resulted in R-work = 0.28 and R-free = 0.32 indicating that the model matched the electron density map. Phenix Autobuild was then used to automatically rebuild the structure of EtpA$^{67-447}$. This procedure adapts the amino acid sequence of the original HMW1A-PP to that of EtpA$^{67-447}$ and adjusts their positions to match the resulting electron density map. The model was further improved by alternating interactive sessions in Coot [186] in which the model was manually aligned with the electron density and automated global refinement in Phenix Refine [185]. Water molecules were also added and their positions refined. Incorrectly placed water molecules were removed manually.

While the electron density map revealed the same global structure for the four molecules in the asymmetric unit, differences in the position or orientation of the residue main- and side-chain atoms could be observed at the level of some side chain and main chain residues. As an example, the side chain of Lys41 in monomer A and D is surrounded by electron density

(Figure 4-8A and D), those for Lys41 in monomers B and C were not (Figure 4-8B and C). Regions of missing electron density are a typical feature of experimental density maps. They result from errors in the experimentally determined structure factor amplitudes (including missing reflections), errors in the derived model, or disorder in the molecules making up a crystal. By restraining equivalent residues to adopt comparable conformations in all four chains, such limitations can be overcome, without preventing observable differences in chemically identical chains exposed to different crystal packing forces.

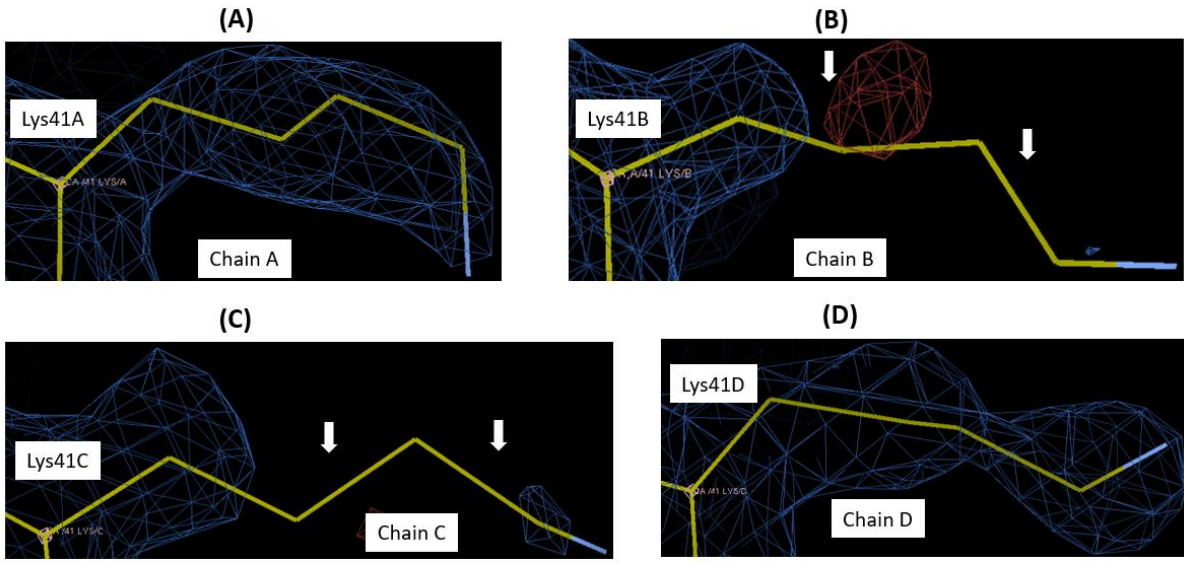



**Figure 4-8: Structural comparisons of the four independent EtpA$^{67-447}$ monomers**

**A to D**: Comparing the side chain Lys41 in the four chains. While electron density is visible for all atoms in chain A and D in this region, electron density is missing in the side chain of chain B and C. White arrows indicate regions of missing electron density.

A final R-work of 0.22 and an R-free of 0.26 for the diffraction data shell of highest resolution are typical for protein structures at a resolution of 1.8 Å and indicate that the refined model overall matches the experimental data well.

The EtpA$^{67-447}$ crystal structure was validated using appropriate tools in Coot and Phenix [186] [185] as well as the PDB deposition tool. Validation included analyses of the backbone geometry via a Ramachandran plot [186] (Figure 4-9), rotamer evaluations, spatial clashes, temperature factor outliers and water molecule positions and contacts. The Ramachandran plot revealed that 93.2% of residues are located on the most favoured region, indicating that the EtpA$^{67-447}$structure is chemically and structurally plausible. Three residues, Gly214B (chain B), Gly214C and Leu216A, were identified as outliers in the Ramachandran plot, though the

electron density in each case confirmed the combination of phi and psi angles for these residues. The EtpA$^{67-447}$ structure was deposited in the Protein Data Bank (PDB) under ID: 8CPK. Data reduction and structure refinement statistics are indicated in Table 4-1.

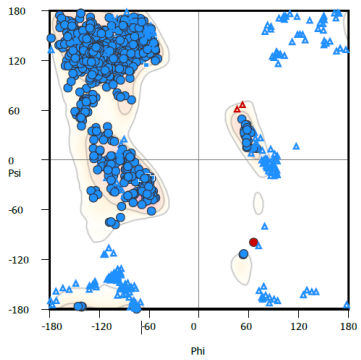



**Figure 4-9: Ramachandran plot for the EtpA$^{67-447}$ crystal structure**

Of 1400 residues in four protein chains, 1305 or 93.2% lie in favoured regions of the phi/psi torsion angle space (blue circles and blue triangles), 92 or 6.57% in the allowed regions (blue triangles with white dots) and 3 or 0.21% in the outlier region (red circle and red triangle).

**Table 4-1: Data collection and refinement statistics for EtpA$^{67-447}$**

| **Data reduction statistics** | | **Refinement statistics** | |
|---|---|---|---|
| Wavelength (Å) | 0.9795 | Reflections used for R-free* | 71355 (6137) |
| Resolution range (Å)* | 64.23 - 1.79 (1.85 - 1.79) | R-work* | 0.22 (0.27) |
| Space group | P1 | R-free* | 0.26 (0.35) |
| Unit cell a, b, c, α, β, γ (Å,°) | 37.7, 64.1, 123.9, 91.6, 90.9, 90.0 | No. of non-hydrogen atoms | 11438 |
| | | No. of waters | 642 |
| Images collected | 3600 | No. of amino acid residues | 1516 |
| Total reflections* | 382586 (35589) | RMS (bonds)(Å) | 0.008 |
| Unique reflections* | 111805 (106543) | RMS (angles)(°) | 1.21 |
| Multiplicity* | 3.4 (3.3) | Ramachandran favoured, additionally allowed, outliers (%) | 95.0, 4.2, 0.8 |
| Completeness (%) * | 97.4 (95.9) | | |
| Mean I/sigma(I)* | 7.8 | | 29.8 |
| Wilson B-factor | 23.7 | Average B-factor protein | 37.30 |
| R-merge* | 0.097 | Average B-factor solvent | |
| R-meas | 0.115 | | |
| *Values in parentheses refer to the outer shell highest in resolution. | | | |

## 4.2.4 Structural overview of EtpA$^{67-447}$

The asymmetric unit of the EtpA$^{67-447}$ crystals was found to contain four symmetrically independent monomers (Figure 4-10A). Each molecule consists of a right-handed β-helix (Figure 4-10B) of 13 turns. Each β-helical turn consists of two or, more commonly, three roughly coplanar β-strands creating a triangular right-handed β-helix. Strands from adjacent turns align to create the parallel β-sheets PB1, PB2 and PB3 with 13, 11 and 14 β-strands,

respectively. Three β-strands of EtpA$^{67-447}$, β13, β20 and β21, create a small extra-helical β-sheet. β-Strands β1 to β8 create a tapered end of the β-helix slightly tilted from the main β-helical axis. While β-strands β1 and β2 form part of PB1 and PB2, they align antiparallel to

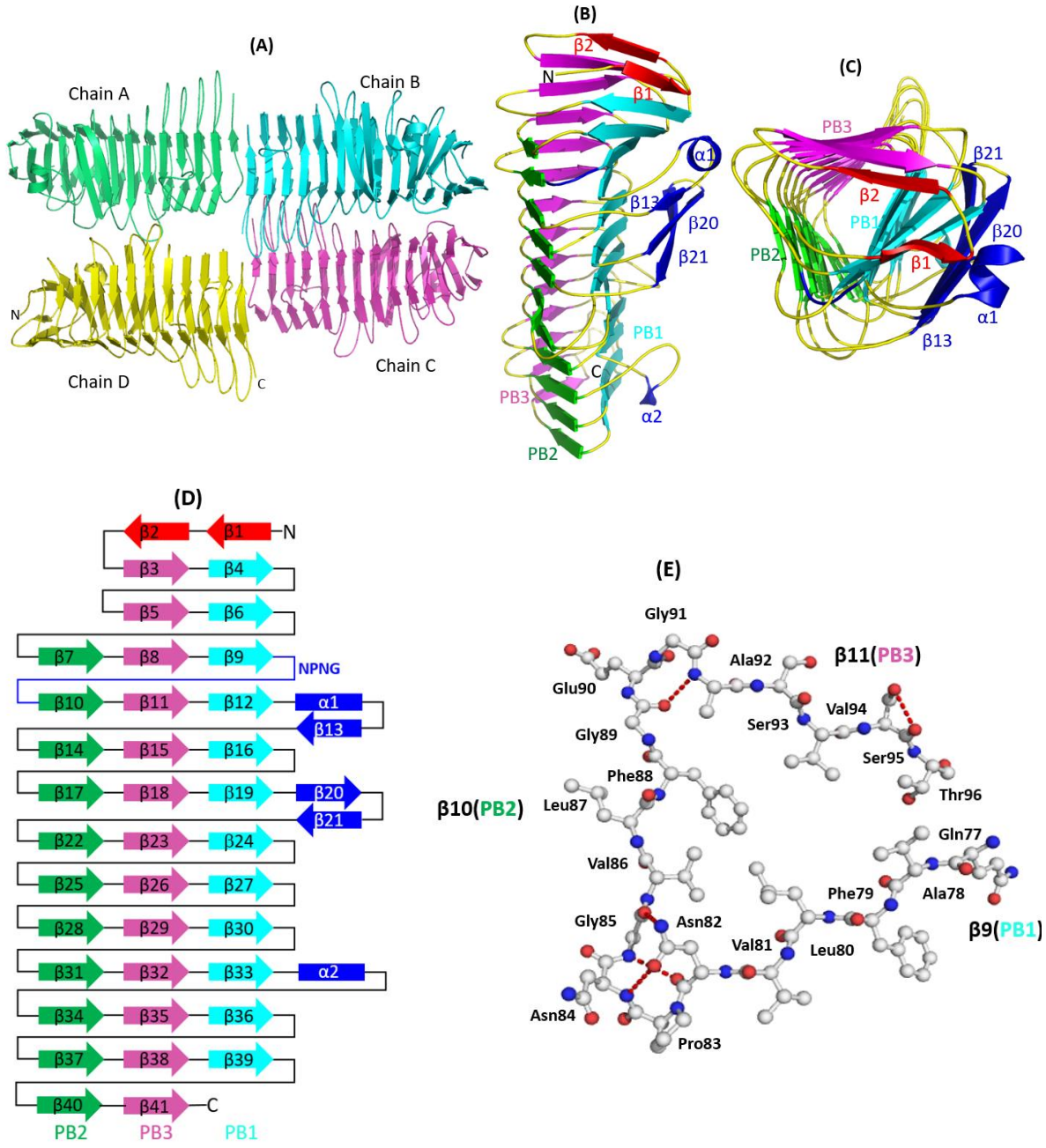



**Figure 4-10: Crystal structure of EtpA$^{67-447}$**

**A)** Ribbon representation of the four EtpA$^{67-447}$ molecules occupying the asymmetric unit of the crystal unit cell. Each chain is marked by a unique colour. **B)** Side view of the EtpA$^{67-447}$ chain A. β-Strands β1 and β2 are coloured red, β-strands of β-sheets PB1, PB2, and PB3 in orange, magenta, and green, and the extra-helical motif α1/β13/β20/β21 and conserved NPNG loop in blue. **C)** N-terminal view highlighting β-strands β1 and β2 that cap and protect the N-terminal end of the hydrophobic core. **D)** Topology diagram. **E)** Ball-and-stick view of β-helix turn 4 with the conserved NPNG motif in loop β9-β10. Red, dotted lines mark turn-stabilising hydrogen bonds.

the otherwise parallel β-sheets.

The 38 β-turns of the β-helix mostly contain four residues including one or two glycine residues. Hydrogen bonds between the glycine main chain atoms and serine side chains oriented towards the β-helix interior stabilize the fold. The loop connecting β-strands β9 and β10 forms a type 1 β-turn that harbours a conserved NPNG motif in TpsA proteins shown to be critical to protein folding [193] . The loop is stabilized by hydrogen bonds between the loop main chain and side chains atoms of residues Asn82, Asn84, Gly85, Gly89, Gly91 and Ser95 (Figure 4-10E). PB1 and PB2 β-strands are respectively three and seven residues in length, with the β-strands of PB2 decreasing in length towards the C-terminus. The interior of the β-helix is dominated by hydrophobic, aliphatic residues, especially valine. N-terminally, β-strands β1 and β2 linked by a hairpin turn shield the hydrophobic core of the β-helix. Two loops connecting physically adjoining β-strands β12 and β14 as well as β19 and β22, are extended to create two small domains outside the main β-helix. The first consists of α-helix α1 and β-strand β13, while the second one includes β20 and β21 (Figure 4-10B). The β-strands of these extra-helical domains broadly run (anti-)parallel to the main β-helical axis. Helix α1 is positioned by hydrogen bonds between Asn108 (α1) and Arg59 of PB1 β-strand β6. Similar hydrogen bonds anchor β-hairpin strands β20 (Arg169 and Thr171) and β21 (Gln183 and Thr185) to β13 (Lys161). A third extra-helical motif involves α-helix α2 forming part of the extended loop connecting β33 (PB1) to β34 (PB3) (Figure 4-10D).

Superimposing the four symmetry-independent chains of EtpA$^{67-447}$ indicates that both the secondary and tertiary structures are highly consistent. Structural differences are small with root-mean-square values between identical atoms of 0.256, 0.253 and 0.125 Å for chains B, C and D relative to chain A (Figure 4-11A). A single noticeable difference involved β-strand β2 (orange, Figure 4-11A), which is slightly longer in chains A and B compared to chains C and D. Interactions between distinct EtpA$^{67-447}$ chains include reciprocal hydrogen bonds Thr219 (B) - Asn218 (D) and Asn218 (B) - Thr219(D) (Figure 4-11B). Chains A and C are linked the hydrogen bonds Ser51(A) - Asp48(C) and Asp203(A) - Asn176(C) (Figure 4-11C).

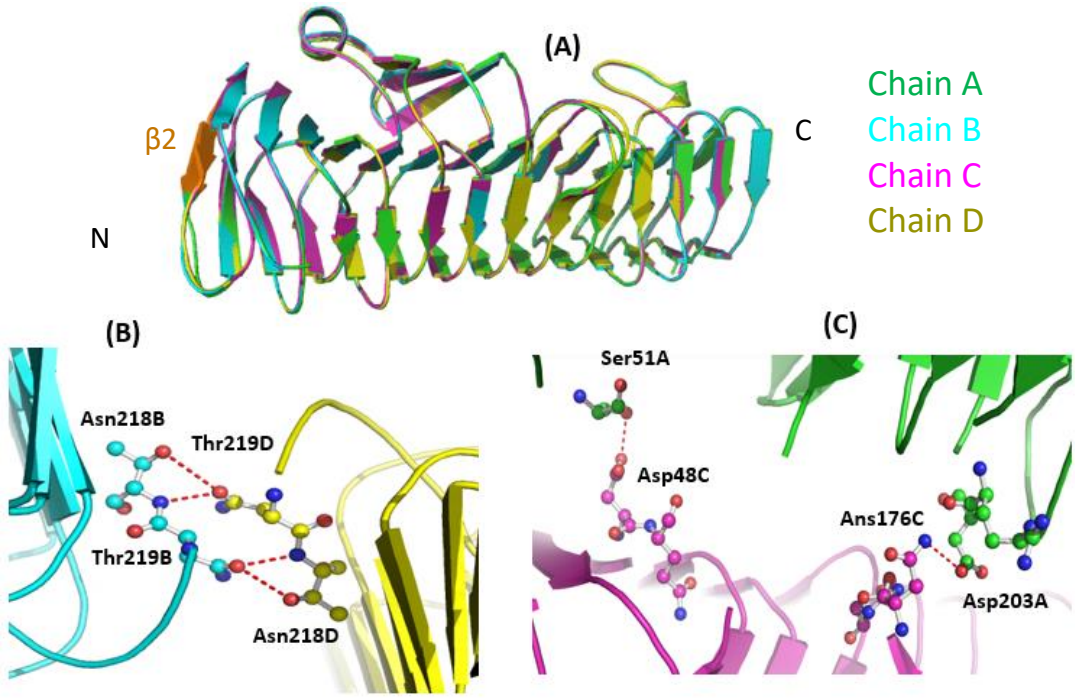



**Figure 4-11: Structural comparison of the four symmetry independent EtpA$^{47-447}$ molecules**

Chains A, B, C and D are coloured green, cyan, magenta and yellow, respectively. Hydrogen bonds are indicated as red dotted lines.
**A)** A superposition of the four symmetry independent chains in the asymmetric unit. **B)** Residues at the interface of chains B and C. **C)** Residues at the interface of chains A and C.

## 4.3 Interaction of EtpA$^{67-447}$ and $^{ETEC}$FliC

### 4.3.1 Cloning of $^{ETEC}$*fliC* gene

To investigate the expected interaction of EtpA with the main flagellin protein FliC from ETEC, a synthetic gene was purchased and cloned into both pGEX-6P-2 and pET28a vectors to produce the fusion proteins GST-$^{ETEC}$FliC and His$_6$-$^{ETEC}$FLIC, respectively.

Restriction enzyme digestion of purified plasmids after ligation and transformation released a 1.5 kb band from all plasmid constructs (Figure 4-12). The 1.5 kb size correspond to the size of $^{ETEC}$*fliC* gene circumstantially confirming the successful incorporation of $^{ETEC}$*fliC* into both pGEX-6P-2 and pET28a vectors. The results were further verified by sequencing.

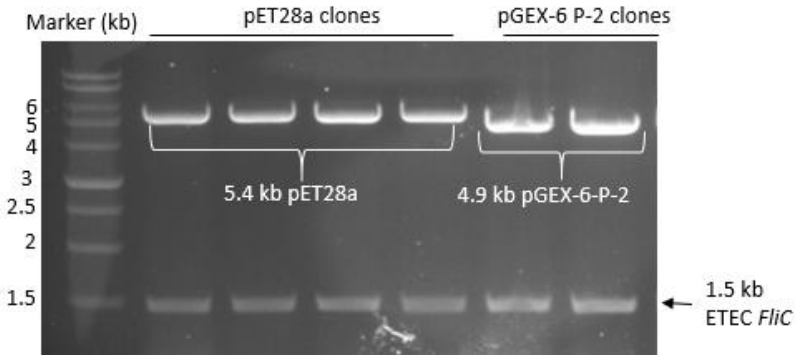



**Figure 4-12: Restriction digest of pGEX-6P-2 and pET28a vectors to release $^{ETEC}$*fliC* genes**

The size of the $^{ETEC}$*fliC* gene insert and those of the vectors after double digestion are indicated on the agarose gel.

### 4.3.2 Production and purification of ETEC flagellin (FliC)

The two vectors pGEX-6P-2-FliC and pET28a-FliC, were used to transform *E. coli* Bl21 cells for the production of $^{ETEC}$FliC proteins (Section 3.56 and 3.57). Protein production was induced with 0.1 mM IPTG, and production at 22°C was allowed to continue for 16 h.

**Glutathione sepharose affinity purification**

The GST-FliC protein produced by transforming *E. coli* Bl21 cells with pGEX-6P-2-FliC vector was purified by glutathione sepharose (GS) affinity chromatography using GS beads. Although no obvious bands stand out for the induced and soluble protein fraction in Figure 4-13, a large band is observed for the protein fraction bound to the GS-resin after contaminating proteins had been removed (lane "Before cleaving"). This band matches the expected size of the GST-$^{ETEC}$FliC fusion protein. The following lane demonstrates the eluted fraction after incubating the fusion protein with 3C protease for 24 h. While uncleaved fusion protein is observed at 71 kDa, dominant bands at 51 and 25 kDa, corresponding to $^{ETEC}$FliC and GST tag, respectively (Figure 4-13). Both the GST-fusion protein and GST should normally have remained bound to the GS beads but in this case the beads appear to have been saturated with fusion protein, leading to "bleeding" of protein from the column. These contaminants were removed selectively by re-exposing the protein sample to fresh GS-beads. GST-FliC was thus successfully produced, cleaved, and eluted from the column.

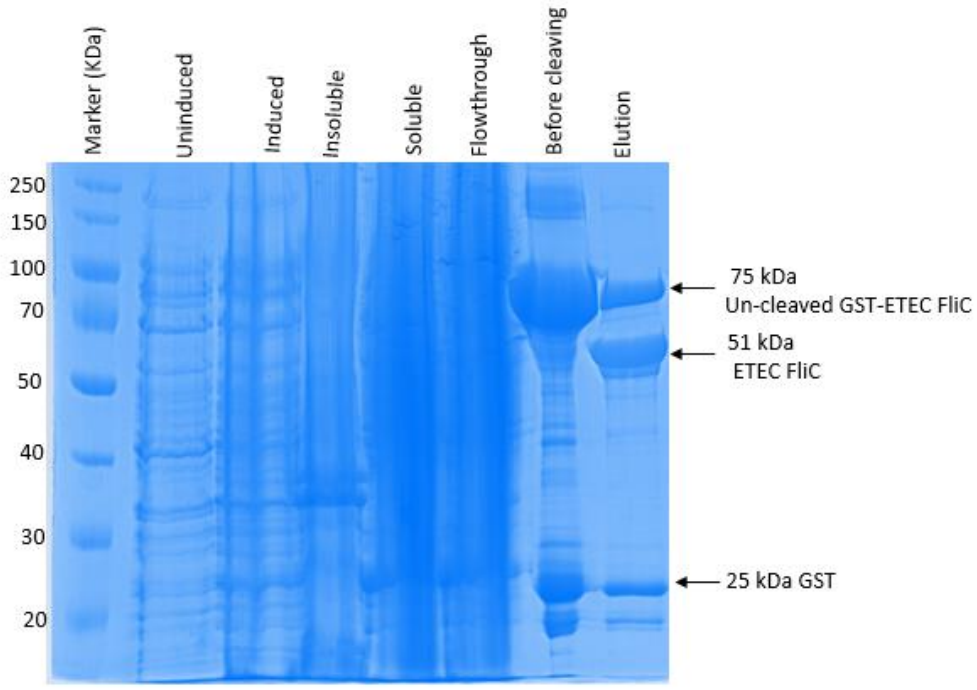



**Figure 4-13: Production and GS affinity purification of GST-FliC**

The SDS-PAGE indicates various protein samples collected during production and purification. The sizes of the target protein bands are also indicated.

### Ion exchange chromatography with $^{ETEC}$FliC

After GS affinity chromatography (Figure 4-13), FliC was further purified by IEC prior to SEC complexing experiments. The corresponding chromatogram indicates a single peak after IEC (Figure 4-14A). Analysing peak fractions by SDS-PAGE revealed FliC protein of higher purity after IEC (Figure 4-14B). Further purification was, however, still required.

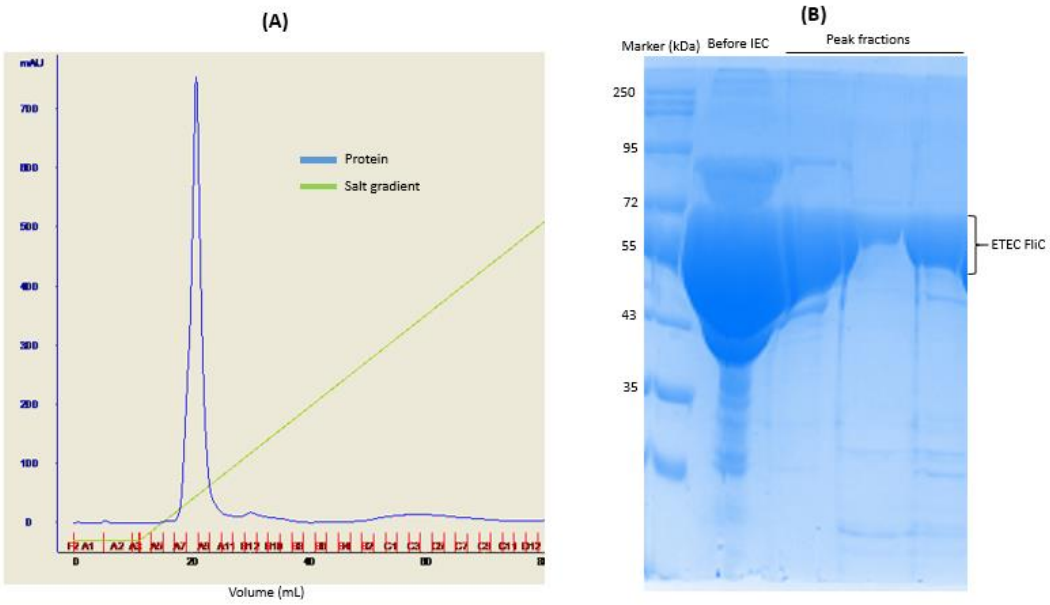



**Figure 4-14: Production and Ni-NTA affinity purification of $His_6$-FliC**

**A)** Chromatogram obtained after IEC chromatography.
**B)** SDS-PAGE of peak fractions obtained after IEC

### Ni-NTA affinity chromatography of $His_6$FliC

The $His_6$-FliC protein produced by transforming *E. coli* Bl21 cells with pET28a-FliC vector was purified by Ni-NTA affinity chromatography. The protein with an N-terminal $His_6$-tag was produced to use as a bait in affinity studies with $EtpA^{67-447}$. The 51-kDa band is visible in the SDS-PAGE analysis (Figure 4-15) for the induced, insoluble, and soluble fractions indicating that $His_6$-$^{ETEC}$FliC protein was successfully produced. Allowing the soluble fraction to interact with Ni-NTA beads, washing and eluting the target protein, yielded a strong 51-kDa band in the elution fraction (Figure 4-15) that again corresponded to the size of $His_6$-$^{ETEC}$FliC protein. Due to major contaminants, the partially purified protein required additional purification steps.

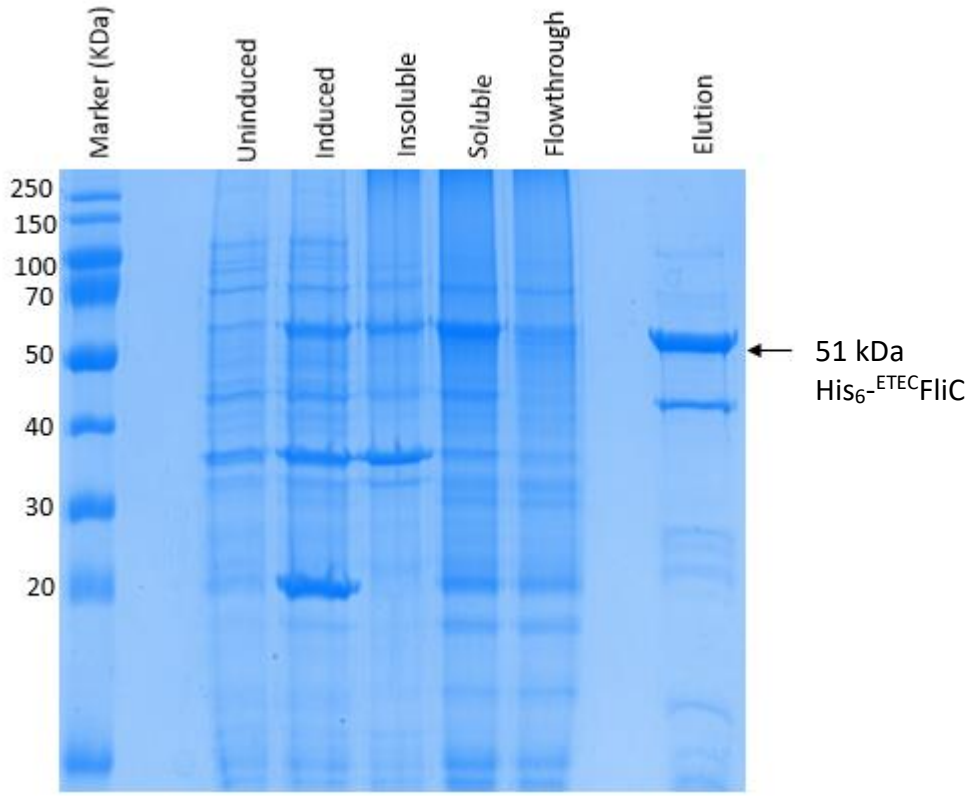



**Figure 4-15: Production and Ni-NTA affinity purification of $His_6$FliC**

The SDS-PAGE indicates various protein samples collected during production and purification. The size of the target $His_6$FliC protein band is also indicated.

### 4.3.3 Size exclusion chromatography

To investigate the interaction of EtpA N-terminal fragments with $^{ETEC}$FliC, SEC was used. Individual proteins were first run separately as internal references. The chromatogram indicates a single peak with elution volume around 14.7 mL for $^{ETEC}$FliC (Figure 4-16A). The $^{ETEC}$FliC samples after SEC (Figure 4-16B) reveal high protein purity, eliminating the danger of non-specific interactions. For EtpA$^{67-447}$, a single peak with elution volume around 16.5 mL is observed after SEC (Figure 4-16C). Though additional protein bands are observed alongside EtpA$^{67-447}$ (Figure 4-16D), their low concentration ensured minimal interference during the interaction studies.

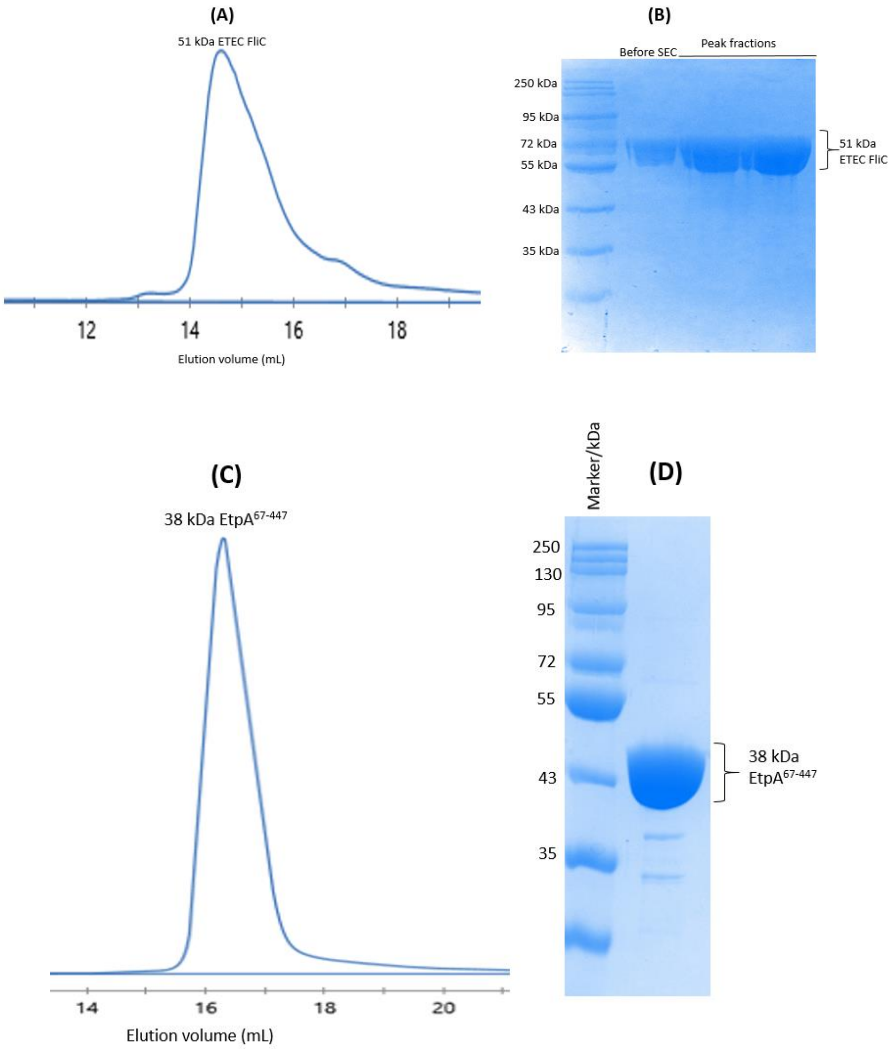



**Figure 4-16: SEC chromatography profile for $^{ETEC}$FliC and EtpA$^{67-447}$**

**A)** $^{ETEC}$FliC SEC chromatogram. **B)** SDS-PAGE of peak fractions obtained after SEC with $^{ETEC}$FliC. **C)** EtpA$^{67-447}$ SEC chromatogram. **D)** SDS-PAGE of peak fractions obtained after SEC with EtpA$^{67-447}$.

### 4.3.3.1 Interaction of EtpA$^{67-447}$ with $^{ETEC}$FliC

To ascertain the interaction of EtpA$^{67-447}$ with FliC, the two proteins were mixed in stoichiometric amounts and incubated overnight. The mixture was analysed on a SEC column and the results were overlaid with those for the individual proteins.

The results indicate the presence of three peaks for the complex experiment (orange line in Figure 4-17A).

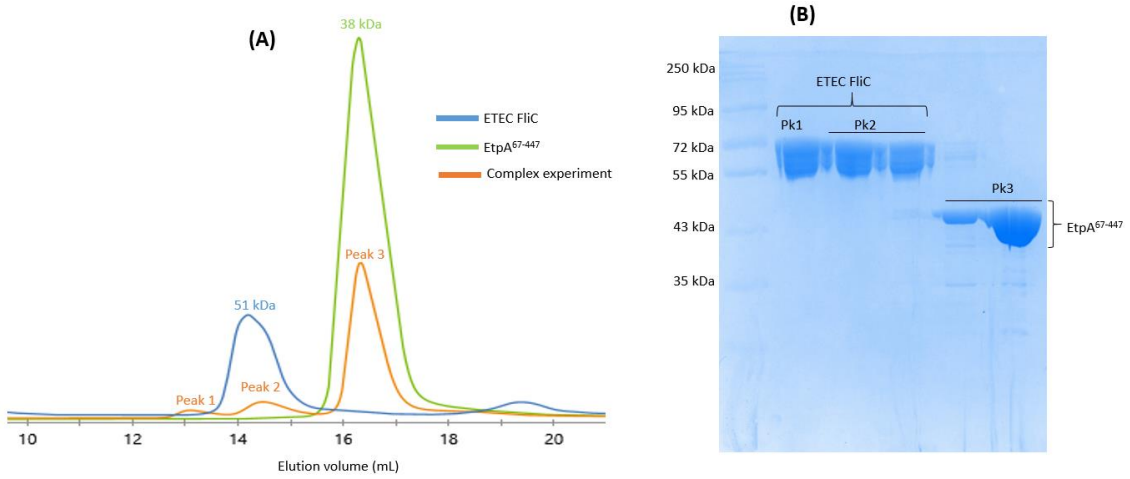



**Figure 4-17: SEC for interaction studies between EtpA$^{67-447}$ and $^{ETEC}$FliC**

**(A)** SEC chromatogram. **(B)** SDS-PAGE of complex experiment peak fractions.

The retention volume of peak 3 (orange) precisely matches that of EtpA$^{67-447}$ (green), implying that it corresponds to uncomplexed EtpA$^{67-447}$ protein. Correspondingly, the SDS-PAGE analysis of peak fractions reveal only protein of around 43 kDa, as previously observed for EtpA$^{67-447}$ confirming the interpretation.

Peak 1, though very small, preceded the other two peaks and was not previously observed for the individual protein samples. Its reduced retention volume could indicate that the two proteins had indeed formed a complex though only to a very limited degree and hence implying a weak interaction. However, the SDS-PAGE analysis for this peak only reveals a band corresponding to $^{ETEC}$FliC protein, while no band corresponding to EtpA$^{67-447}$ is observed. This would thus imply that peak 1 could correspond to a $^{ETEC}$FliC dimer, rather than the expected complex.

Peak 2 (orange) is somewhat more enigmatic. Its retention volume is larger than that recorded for $^{ETEC}$FliC (green) by itself, possibly implying that $^{ETEC}$FliC dynamically switches between a dimer and monomer state, but that this equilibrium is disrupted in the presence of EtpA$^{67-447}$. The SDS-PAGE correspondingly only reveals $^{ETEC}$FliC in this sample. The last fraction of this peak does appear to show a faint band corresponding to EtpA$^{67-447}$ (Figure 4-17B). This could either be due to a slight spill of the Peak 3 sample into the previous lane or again imply a very weak interaction of the two proteins. Overall though, the data presented does not support a direct interaction of the two proteins. At best there is a slight indication of very weak interaction. But this would need to be confirmed by other techniques.

### 4.3.3.2 Interaction of EtpA$^{1-606}$

To investigate the potential interaction of $^{ETEC}$FliC with the slightly longer EtpA$^{1-606}$, SEC was again used as in the preceding section. Both individual proteins were SEC purified For EtpA$^{1-606}$ the elution profile indicated a main peak (Peak 2, retention volume of 13.9 mL, Figure 4-18A) preceded by a broad shoulder marked as Peak 1, with a retention volume of 12 mL. SDS-PAGE analysis confirmed all peaks to belong to EtpA$^{1-606}$ with an expected mass of ~63 kDa. While Peak 2 presumably is the predominant dimeric protein, Peak 1 was presumably caused by higher oligomers of EtpA$^{67-447}$.

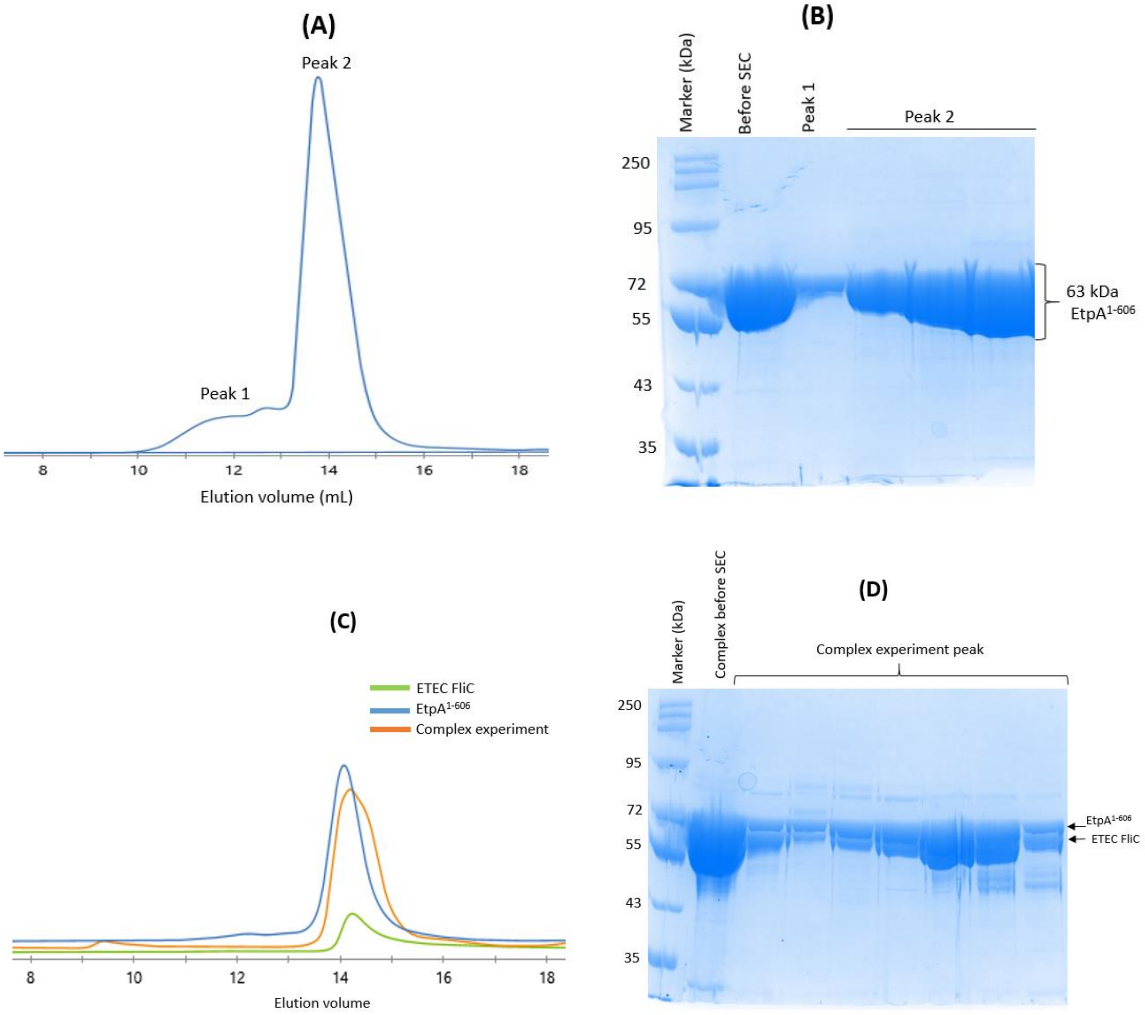



**Figure 4-18: SEC for interaction studies between EtpA$^{1-606}$ and $^{ETEC}$FliC**

**(A)** EtpA$^{1-606}$ SEC Chromatogram. **(B)** SDS-PAGE of peak fraction obtained after EtpA$^{1-606}$ SEC. **(C)** SEC chromatogram obtained after complexing experiment. **(D)** SDS-PAGE of complex experiment peak fractions.

The chromatogram for the complexing experiment (orange curve, Figure 4-18C) revealed a single broadened peak that incorporates both the EtpA$^{1-606}$ and $^{ETEC}$FliC main peaks. SDS-PAGE indicates the presence of both both EtpA$^{1-606}$ and $^{ETEC}$FliC proteins throughout the peak (Figure 4-18D). However, the presence of both proteins cannot be taken to indicate a complex formation of the two proteins as this would have created a peak to the left of the two individual peaks i.e. having a smaller retention volume. As such a peak is not seen, there is no evidence that the two proteins directly interact.

### 4.3.4 Molecular pulldown assay

The potential interaction of EtpA N-terminal fragments with FliC was further investigated using a molecular pulldown assay. In this case the bait protein GST-$^{ETEC}$FliC was non-covalently bound to GS beads via the GST tag. Binding of EtpA$^{1-606}$ was monitored by allowing it to bind before allowing unbound protein to elute.

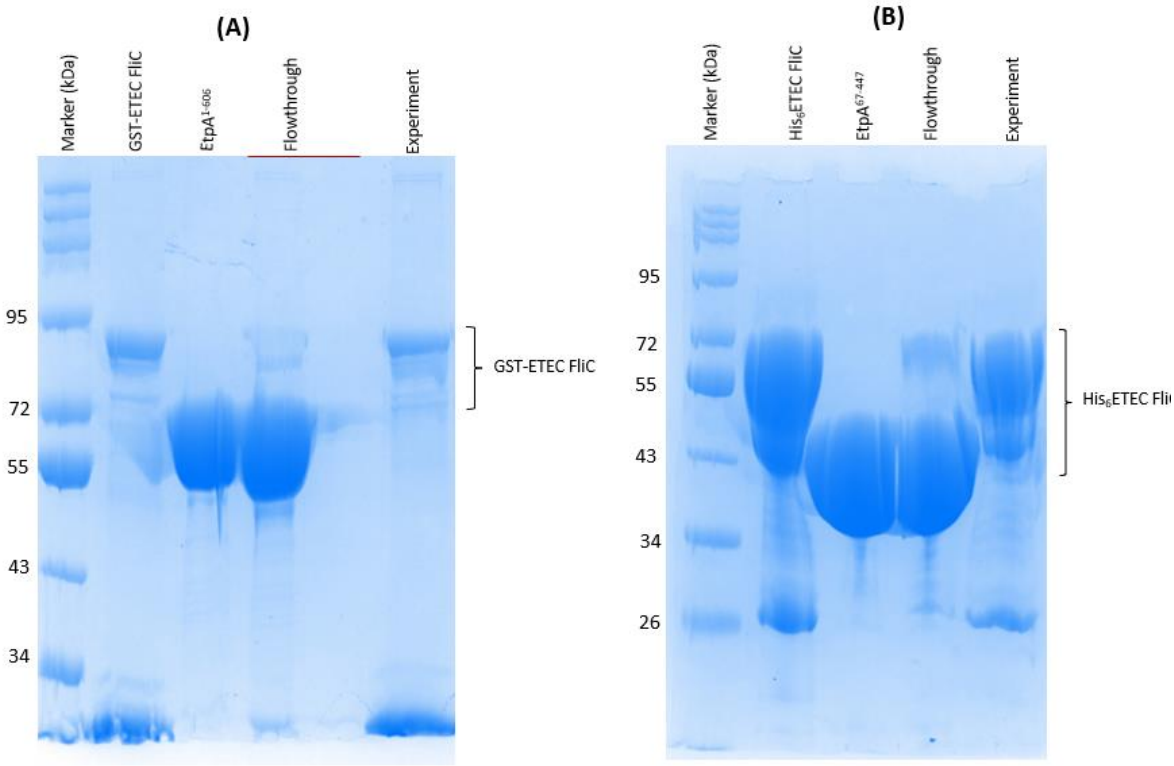



**Figure 4-19: Binding assays for EtpA N-terminal fragments and $^{ETEC}$FliC**

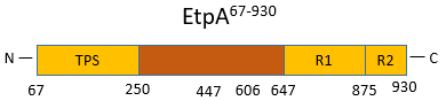
**(A)** Pulldown assay between GST-$^{ETEC}$FliC as bait and EtpA$^{1-606}$. **(B)** Pulldown assay between His$_6$-$^{ETEC}$FliC as bait and EtpA$^{67-447}$.

The SDS-PAGE indicates that EtpA$^{1-606}$ eluted essentially quantitatively accompanied by a small amount of GST-$^{ETEC}$FliC (Figure 4-19A). In the experiment only GST-$^{ETEC}$FliC protein band

is observed (Figure 4-19A) but no EtpA$^{1-606}$. The fact that EtpA$^{1-606}$ elutes quantitatively indicates very weak to no interaction between EtpA$^{1-606}$ and GST-$^{ETEC}$FliC.

A comparable assay was undertaken for EtpA$^{67-447}$. Here His$_6$-$^{ETEC}$FliC was bound to Ni-NTA beads as a molecular bait. However, the eluted protein contained most EtpA$^{67-447}$, with a small contamination of His$_6$-$^{ETEC}$FliC (Figure 4-19B). In the experiment no EtpA$^{67-447}$ was retained leaving only His$_6$-$^{ETEC}$FliC (Figure 4-19B). Again, if the proteins do interact directly, binding would be very weak i.e. with a $K_D$ in the 100 µM range or weaker.

## 4.4 EtpA$^{67-930}$




Apart from the shorter N-terminal fragments of EtpA described above, an attempt was made to produce a longer fragment extending from residue 67 to 930, denoted EtpA$^{67-930}$. This fragment was designed to include a part of the central region of the protein. At the genetic level, fragments encoding residues 67 to 447 and 448 to 930 were combined.

### 4.4.1 Cloning

A plasmid pGEX-6P-2-*EtpA$^{448-930}$* was purchased from Gene Universal (Newark, USA). Due to the cloning procedure, only a *BamHI* restriction site was available at the 5' of the 448-930 encoding insert to include the fragment coding for residues 67 to 447. Forward and reverse amplification primers were thus designed for the latter where both contained *BamHI* restriction sites (Section 3.4.2). This would allow the insertion of the 5' fragment into the *BamHI* linearised vector – albeit in two possible orientations. A plasmid construct with the correct orientation was identified by Sanger sequencing, whereupon the central *BamH1* restriction site was removed by site-directed mutagenesis. The final construct was once more checked by agarose gel electrophoresis (Figure 4-20).

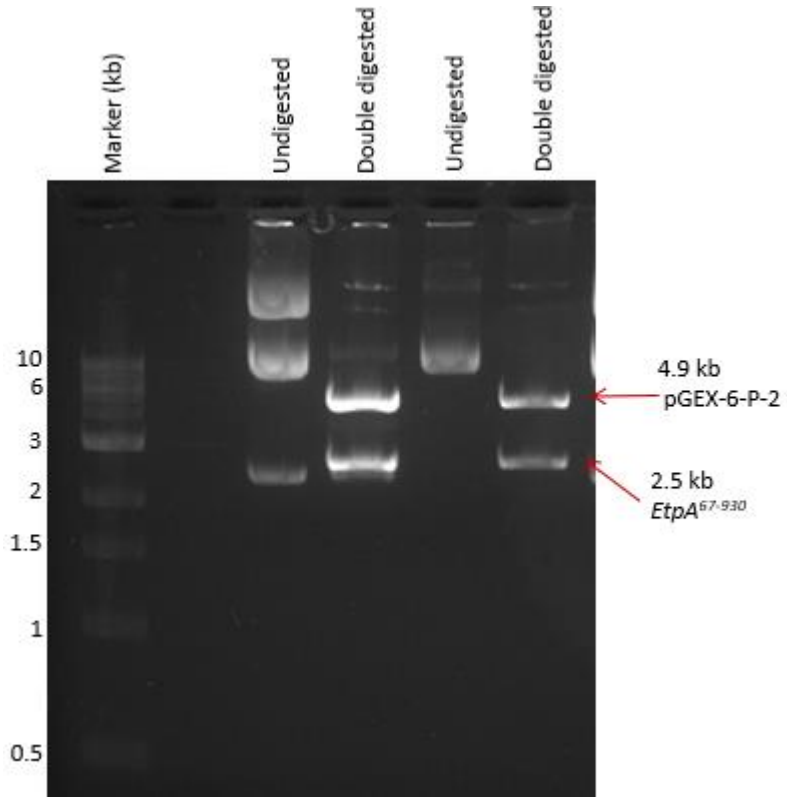



**Figure 4-20: Restriction digest of presumed pGEX-6P-2-EtpA$^{67-930}$ plasmids**

The size of the *EtpA*$^{67-930}$ insert and that of pGEX-6P-2 vector after the double digestion are indicated on the agarose gel.

In the double *BamH1* and *Not1* restriction experiment, a 2.5 kb band was observed that corresponds to the size of *EtpA*$^{67-930}$ gene fragment as well as a 4.9 kb band that corresponds to the size of pGEX-6P-2 vector. The generation of a pGEX-6P-2-*EtpA*$^{67-930}$ plasmid thus proved successful.

### 4.4.2 Production trial

The pGEX-6P-2-EtpA$^{67-930}$ plasmid was used to produce GST-EtpA$^{67-930}$ fusion protein at 22, 30 and 37°C using 0.1 and 0.5 mM IPTG. An expected 112 kDa band corresponding to the GST-EtpA$^{67-930}$ fusion protein was observed in SDS PAGE gels for all temperatures and IPTG concentrations (Figure 4-21A and B). Unexpectedly, the intensity of the 112 kDa band was not dramatically affected by variations in temperature and IPTG concentration. Overall most protein was produced at 37°C and 0.5 mM IPTG.

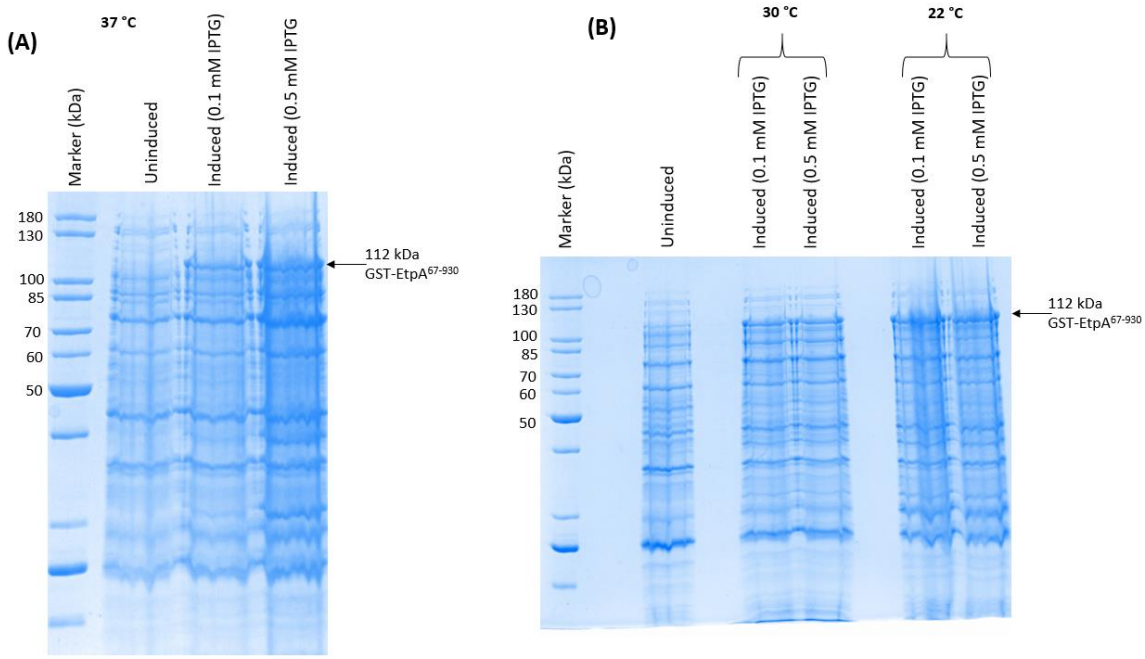



**Figure 4-21: Production trial for EtpA$^{67-930}$**

**(A)** Production trial at 37°C with 0.1 mM and 0.5 mM IPTG. **(B)** Production trial at 22 and 30°C with 0.1 mM and 0.5 mM IPTG.

### 4.4.3 Solubility test

During the previous optimization, the total protein production was analysed rather than the solubility and hence its usefulness. Hence the proportion of soluble versus insoluble protein was tested at 15, 22 and 30°C inducing with 0.1 mM IPTG (Figure 4-22A).

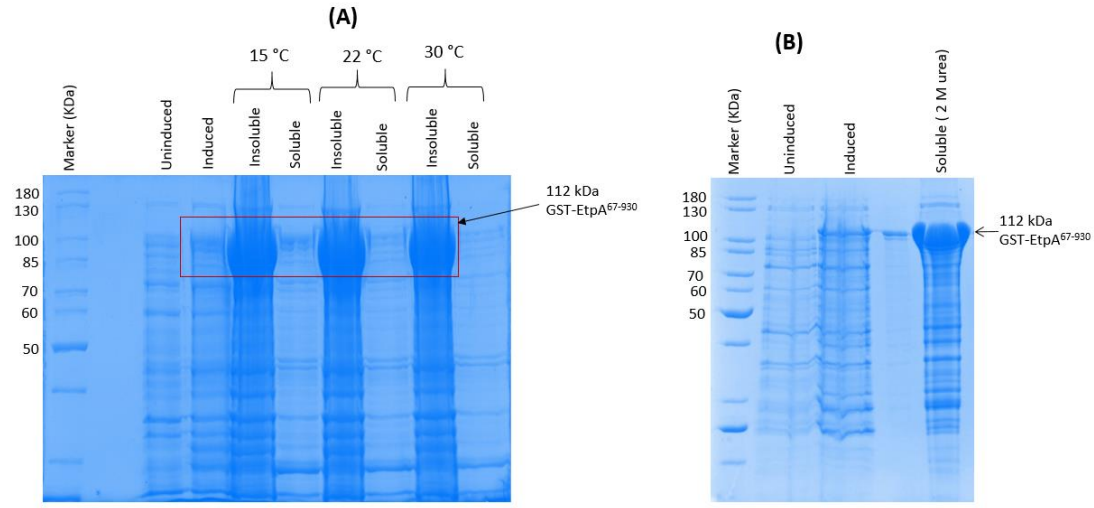



**Figure 4-22: Solubility test and solubilisation of GST-EtpA$^{67-930}$**

**(A)** SDS-PAGE for solubility at different temperatures. **(B)** SDS-PAGE of solubilisation with urea.

Although the expected 112-kDa band for the GST-EtpA$^{67-930}$ fusion protein is observed for all temperatures, the overwhelming larger proportion of the protein is observed in the lanes labelled “Insoluble” indicating that most of the protein was produced in an insoluble form. Fortunately the same band is also observed in fractions labelled as “soluble”, indicating that some protein, alas a small proportion was produced in a soluble form at these temperatures. The largest proportion of soluble protein is observed at 15°C compared to 22 and 30°C, though the proportion of soluble protein is still low. Protein inclusion bodies were solubilised with 2 M urea yielding the expected 112 kDa band of GST-EtpA$^{67-930}$ Figure 4-22B). To refold the solubilized protein, half of the sample was diluted 10-fold with PBS, mixed with GS beads and incubated at 4°C with gentle agitation while the second fraction was dialysed a ten-fold larger volume of PBS. The diluted sample failed to bind to GS beads indicating that the 0.2 M urea did not allow the GST tags to fold sufficiently for binding. The dialysed faction precipitated about 30 min into the dialysis process implying that the transition was possibly too fast. Next buffers with different pH values and salt concentrations were used in stepwise dialyses to remove the urea more gradually. However, all attempts resulted in protein precipitation. No workable procedure to refold EtpA$^{67-930}$ was thus achieved.

## 4.5 AlphaFold models

Full-length EtpA is a large protein of ~170 kDa. This project initially started with a reduced, N-terminal fragment encompassing residues 1 to 606 (EtpA$^{1-606}$) encoded by a plasmid provided by Prof James M. Fleckenstein, Washington University, St Louis, MO, USA. This fragment was produced and characterised, followed by the X-ray crystal structure of the yet shorter EtpA$^{67-447}$. The high molecular weight limited our ability to clone and produce full-length EtpA. A synthetic gene was purchased to extend the N-terminal fragment up to residue 930. This was added to the vector encoding EtpA$^{67-447}$to produce EtpA$^{67-930}$. However, this cytoplasmically produced protein was mainly found to be insoluble, a possible bacterial response to a very large or hydrophobic protein. This prevented any further characterization of EtpA$^{67-930}$ (see above). To obtain relevant structural information for this protein, AlphaFold was used to generate a theoretical structure. In addition, structures of four related TpsA proteins were also modelled, to allow some comparison of EtpA to other proteins for which structural information was previously largely lacking.

### 4.5.1 AlphaFold model of full-length EtpA

A structural model of full-length EtpA was generated on the AlphaFold server [190]. As part of the output, AlphaFold provides a local Distance Difference Test (lDDT) scores for each amino acid residue indicating its reliability in terms of local or long-range interactions as well as its stereo-chemistry. lDDT values range between 0 and 100, for low to high confidence [194] and is colour coded from blue to red in AlphaFold plots.

For EtpA, lDDT values for the first 70 residues are low (Figure 4-23A). These residues constitute the N-terminal signal peptide which is hydrophobic, and inherently unfolded. In the AlphaFold models this represented as an unstructured loop of undefined secondary structure interrupted by two β-strands (Figure 4-23C). lDDT scores were above 70 for most residues beyond the N-terminal signal peptide with some exceptions between residues 200 to 300 and 400 to 600 (Figure 4-23B). The residues around position 200 are modelled into an extended loop in red that extends from the first α-helix in green (Figure 4-23C). Residues between positions 400 and 600 also have lower lDDT scores. They were modelled as the start of the C-terminal repeat domain, shown in dark blue in Figure 4-23C. Interestingly, the residues connecting the N- (yellow) and C-terminal (cyan/magenta) were modelled as a long loop structure plus an α-helix that interrupts the continuity of the β-helical design by being positioned between adjacent β-helical turns (red in Figure 4-23C). The inserted α-helix creates a distinct kink into the otherwise linear structure of the β-helix. Such regions between domain boundaries may be prone to increased flexibility [194].

lDDT values slightly below 60 are also observed between position 600 and 800, position 900 and 1000 and around position 1200. These residues form part of repeating loops (magenta in Figure 4-23C) that create the start and end of the four 228-residues repeats found in EtpA and correspond to conserved STSGNAINL sequence motifs (magenta in Figure 4-23C). They cause a constriction of the β-helix (arrows) creating boundaries between the related repeats, resulting in structural flexibility and slightly less than perfect stereochemistry. Extra-helical domains observed for the EtpA N-terminal region are absent in the C-terminal domain.

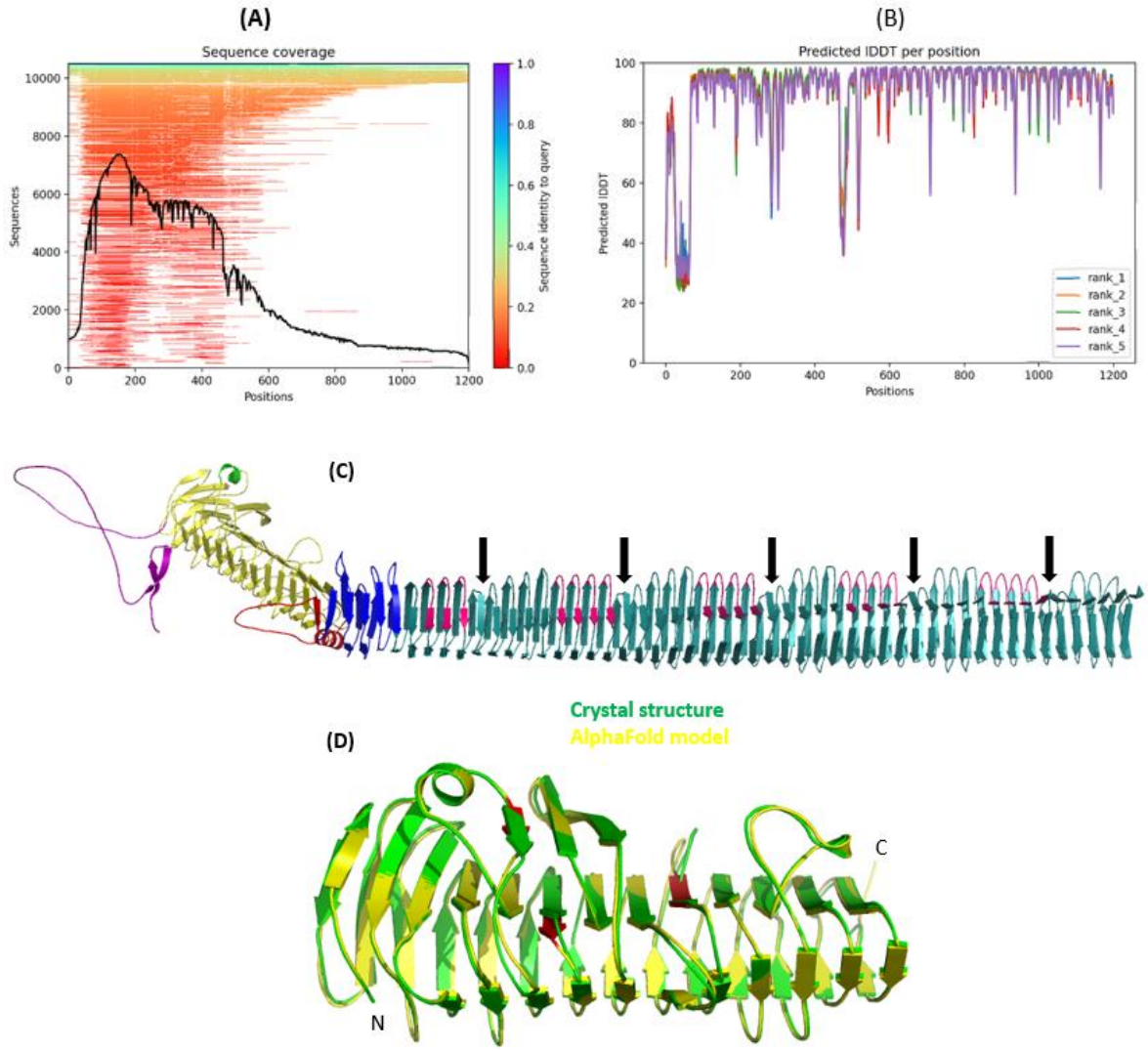



**Figure 4-23: AlphaFold model of full-length EtpA**

**A)** Sequence coverage of protein sequences related to EtpA. **B)** Calculated lDDT scores indicating the reliability of the predicted structure for individual residues. **C)** Modelled structure of full-length EtpA. The N-terminal domain in yellow harbours the TPS domain. The signal peptide and the first α-helix are rendered in green and purple respectively. The repetitive C-terminal domain is in dark cyan. An intervening α-helix, linked to the interdomain kink (in blue), and the associated loop are highlighted in red. Conserved STSGNAINL motifs associated with C-terminal repeats are shown in magenta. Indentations following major repeats are marked by black arrows. **D)** Structural alignment of the experimental TPS domain crystal structure in green and the AlphaFold model in yellow. Structural differences in β-strands are marked in red.

The TPS domain represented by the EtpA$^{67-447}$ crystal structure closely matches the N-terminal part of the AlphaFold model (Figure 4-23D). Three β-strands are slightly shorter or longer in the crystal structure than in the model (red in Figure 4-23D). The overall N-terminal domain fold is nevertheless almost identical between the two structures.

**Side chain comparison of Modelled EtpA$^{67-447}$and the crystal structure**

As outlined above, a structural superposition of the EtpA$^{67-447}$crystal structure and the corresponding AlphaFold model revealed no major difference except for three β-strands that are slightly shorter or longer in either model (Figure 4-24D).

To check whether the overall similarity of the fold also extends to the conformation of side chains, these were compared for these three β-strands β13, β16 and β27 (Figure 4-23D).

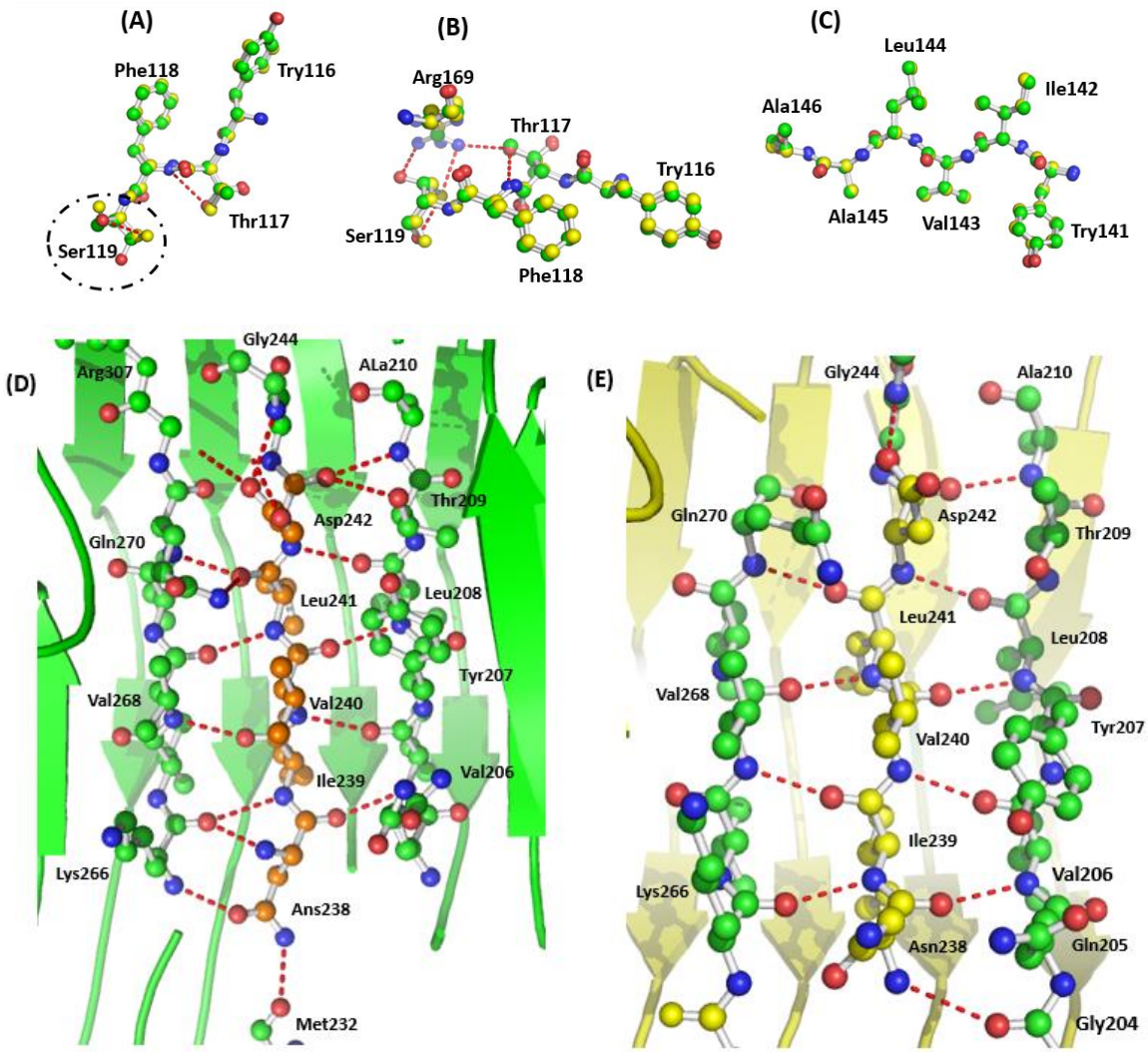



**Figure 4-24: Comparing side chain residues of EtpA$^{67-447}$crystal and modelled structure**

**A)** β-Strand β13 residues analysis with doted circles emphasising on the Ser119. **B)** β-Strand β13 residues showing side chain bonding with Arg169. **C)** β-Strand β16 residues analysis.
**D)** Crystal structure showing boding between β-strand β27 (orange) to other residues (green) on the structure. **(B)** AlphaFold model showing bonding between β-strand β27 (yellow) to other residues (green) on the structure. Hydrogen bodings are indicated in dotted lines. Crystal structure and AlphaFold model residues are indicated in green and yellow respectively for A-C.

The β-strand β13 of the crystal structure encompasses the residues Try116, Thr117, Phe118 and Ser119. In the AlphFold model Ser119 forms part of the loop connecting β13 to β14 However, except for the side chain of Ser119, all other residues adopt highly comparable conformations. A hydrogen bond between Thr117-OD1 to Phe118-N is conserved in both (Figure 4-24A). Similarly the side chain of Ser119 forms hydrogen bond to Arg169 in both models, despite the slight change in orientation on the side chain (Figure 4-24B).

Next we compared the side chains for residues in β-strand β16 (Figure 4-10D) to those on the AlphaFold model. In the crystal structure β-strand β16 consist of the five residues Try141, Ile142, Val143, Leu144 and Ala 145, but includes Ala146 in the AlphaFold model, which forms part of the loop connecting β16 to β17 in the crystal structure (Figure 4-10D). The side chain of β16 residues align perfectly relative to each other (Figure 4-24C).

The third β-strand that was slightly longer in the AlphaFold model is β-strand β27 (Figure 4-10D). It encompasses Ile239, Val240, Leu241 and Asp242 in the crystal structure, while AlphaFold extended the β-strand to include Asn238. In the crystal structure, the side chain of Asn238 forms hydrogen bonds to the main chain atoms of Met232, Lys266 and Val206 (Figure 4-24D). In the AlphaFold model, Asn238 instead forms hydrogen bonds the main chain atoms of Gly204 and Val206 (Figure 4-24E). In both models, the hydrogen bonding pattern for residues Ile239 (to Lys266 and Val268), Val240 (to Val208 and Leu208), and Leu241 (to Val268 and Qln270) is conserved (Figure 4-24D and E). For Asp242 a hydrogen bond to Gly244, Leu208 and Ala210 are conserved in both models (Figure 4-24E), though the crystal structure indicates two additional hydrogen bonds to Thr209 and Arg307 (Figure 4-24E).

A comparison of the conserved NPNG loop again confirms a high degree of similarity for both models with conserved hydrogen bonds between Asn82 and both Asn84 and Gly85 (Figure 4-25). Apart from the minor conformational differences and hydrogen bonding patterns, no major differences were noticeable for the side chains of the crystal structure of EtpA$^{67-447}$and the AlphaFold model. In part this may be due to similar TpsA N-terminal domains having been used in training AlphaFold. Nevertheless the similarity in structures is noticeable.

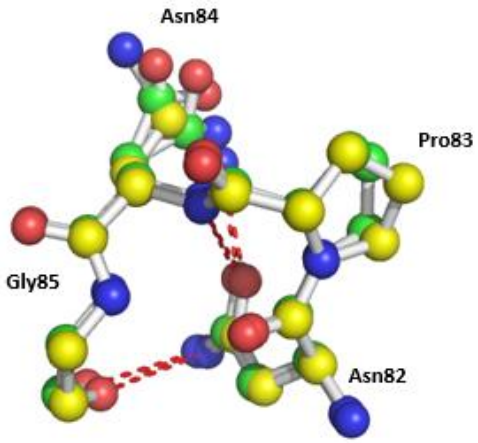



**Figure 4-25: Analysis of conserved NPNG motif residues for EtpA$^{67-447}$**

Crystal structure (green) and the AlphaFold model (yellow). Hydrogen bonds are indicated in red dots.

### 4.5.2 AlphaFold model of full-length HMW1A

The high molecular weight 1A (HMW1A) protein from *H. influenza* was also modelled using AlphaFold. Results for the first 1200 residues indicate very low sequence identity to query for most of the sequence (Figure 4-26A).

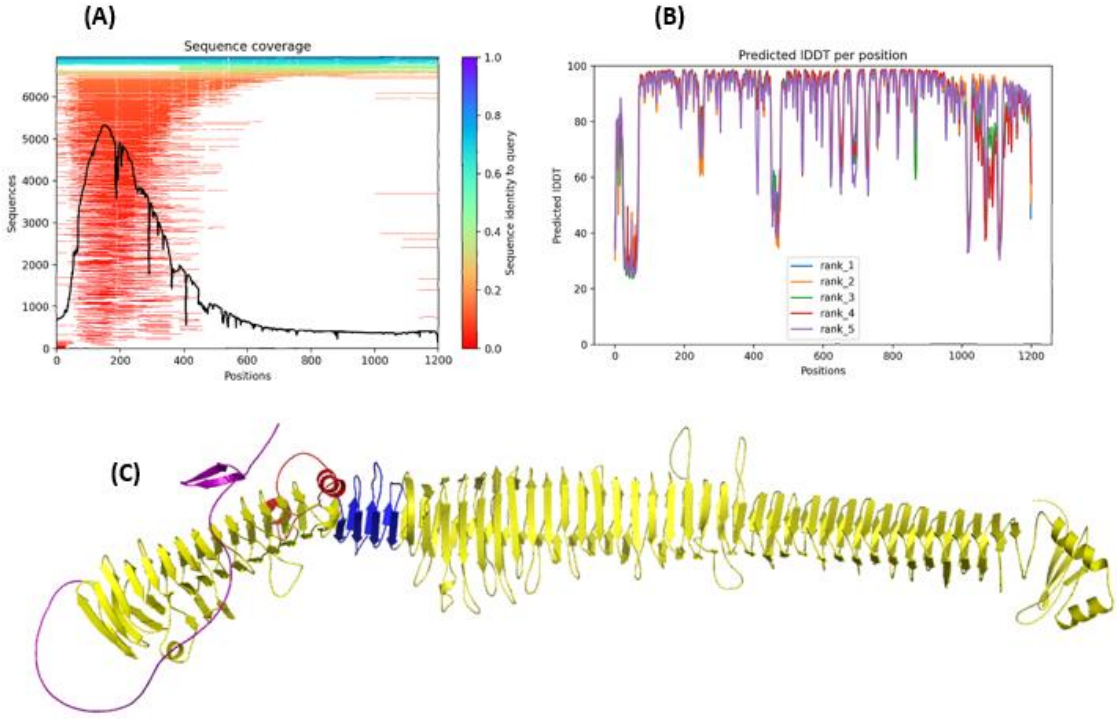



**Figure 4-26: AlphaFold model of HMW1A**

**A)** Sequence coverage. **B)** Calculated lDDT score per residue. **C)** Modelled full-length HMW1A structure. The signal peptide is coloured in purple, an interdomain α-helix, linked to the interdomain kink (in blue), and the associated loop are highlighted in red.

The lDDT scores for residues 1 to 80, 400 to 500, and 1000 to 1200 are quite low (Figure 4-26B). The residues involved constitute the protein secretion peptide in purple, an α-helix with an extended loop in red and the inter-domain kink in blue (Figure 4-26C). Overall, full-length HMW1A is modelled as a continuous β-helix with a noticeable inter-domain kink (Figure 4-26C). In this regard, the structure of HMW1A is highly reminiscent of EtpA.

### 4.5.3 AlphaFold model of full-length FHA

As for the other TspA structures, significant sequence identity for the 1276-residue filamentous hemagglutinin (FHA) protein from *B. pertussis* is mainly limited to the N-terminal domain i.e. up to residue 400 (Figure 4-27A). Only a single related sequence was identified for the C-terminal end.

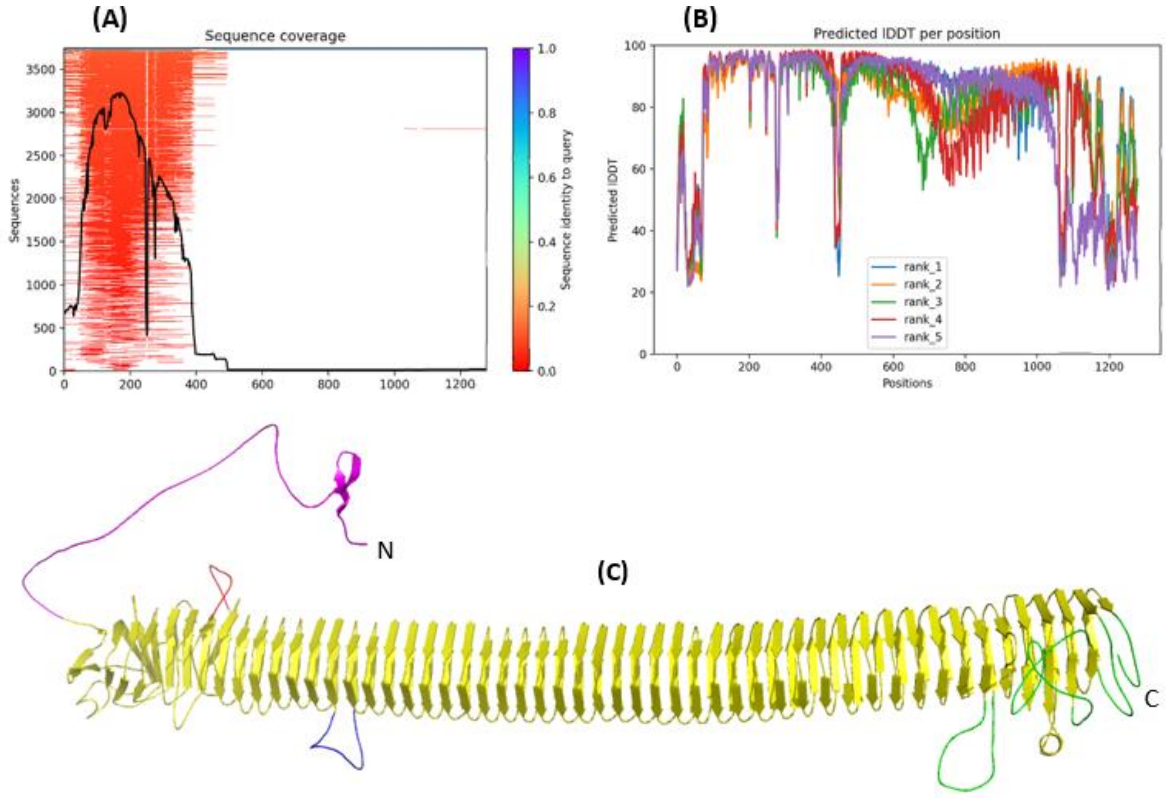



**Figure 4-27: AlphaFold model of FHA**

**A)** Sequence coverage. **B)** Predicted lDDT score per position. **C)** Ribbon representation of the modelled FHA first fragment. The signal peptide is coloured in purple. Extended loops involved with low lDDT scores are indicated in red, blue and green.

Mostly lDDT scores were high, except for the N-terminal 80 residues, for narrow bands of residues near positions 250 and 500 as well as resides beyond position 1000. The poorly defined residues at positions 250 and 500 are modelled as extra-β-helical loops (blue and

green in Figure 4-27C) as well as C-terminal β strands. The limited information on how these extended loops could interact with the main β-helical structure, is reflected in the low lDDT scores. This could also correlate with increased flexibility. Overall full-length FHA protein was modelled as a single, continuous β helix, which lacks the interdomain kink plus associated α-helix observed for EtpA and HMW1A (Figure 4-27C).

### 4.5.4 AlphaFold model of full-length HpmA

Results for the modelled 794 residues hemolysin (HpmA) protein from *P. mirabilis* indicate a high sequence coverage to query for the first 250 positions and decreases for the rest of the protein (Figure 4-28A). The overall lDDT score is very high with the only exception being at the first 50 positions (Figure 4-28B) which contain the signal peptide sequence. The overall predicted lDDT score is very high, producing a β-helical structure with β-sheets that extends from its core (Figure 4-28C).

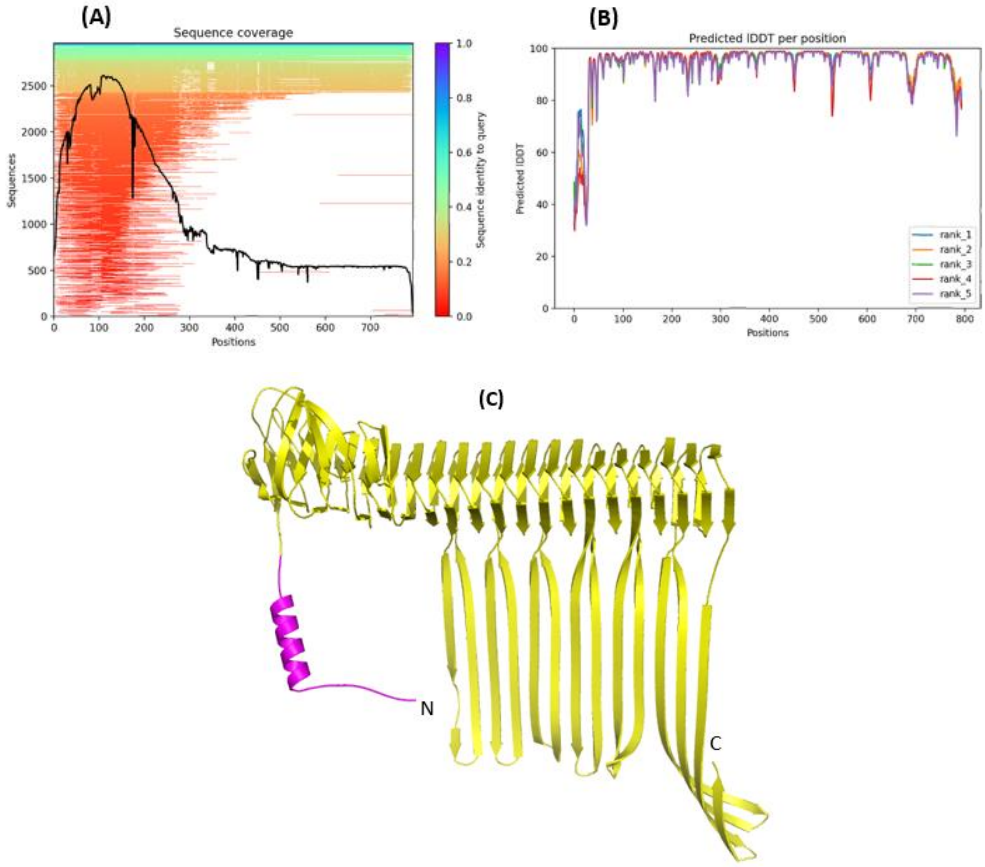



**Figure 4-28: AlphaFold model of HpmA**

**A)** Sequence coverage, **B)** Predicted lDDT score per position. **C)** Ribbon representation of modelled HpmA. The signal peptide is coloured in purple.

### 4.5.5 AlphaFold model of full-length HxuA

Modelling results for the 884-residue hemopexin-binding protein (HxuA) from *H. influenza* indicate many related sequences for the first 100 to 300 amino acids but no to almost no related sequences for residues 400 to 884 (Figure 4-29A).

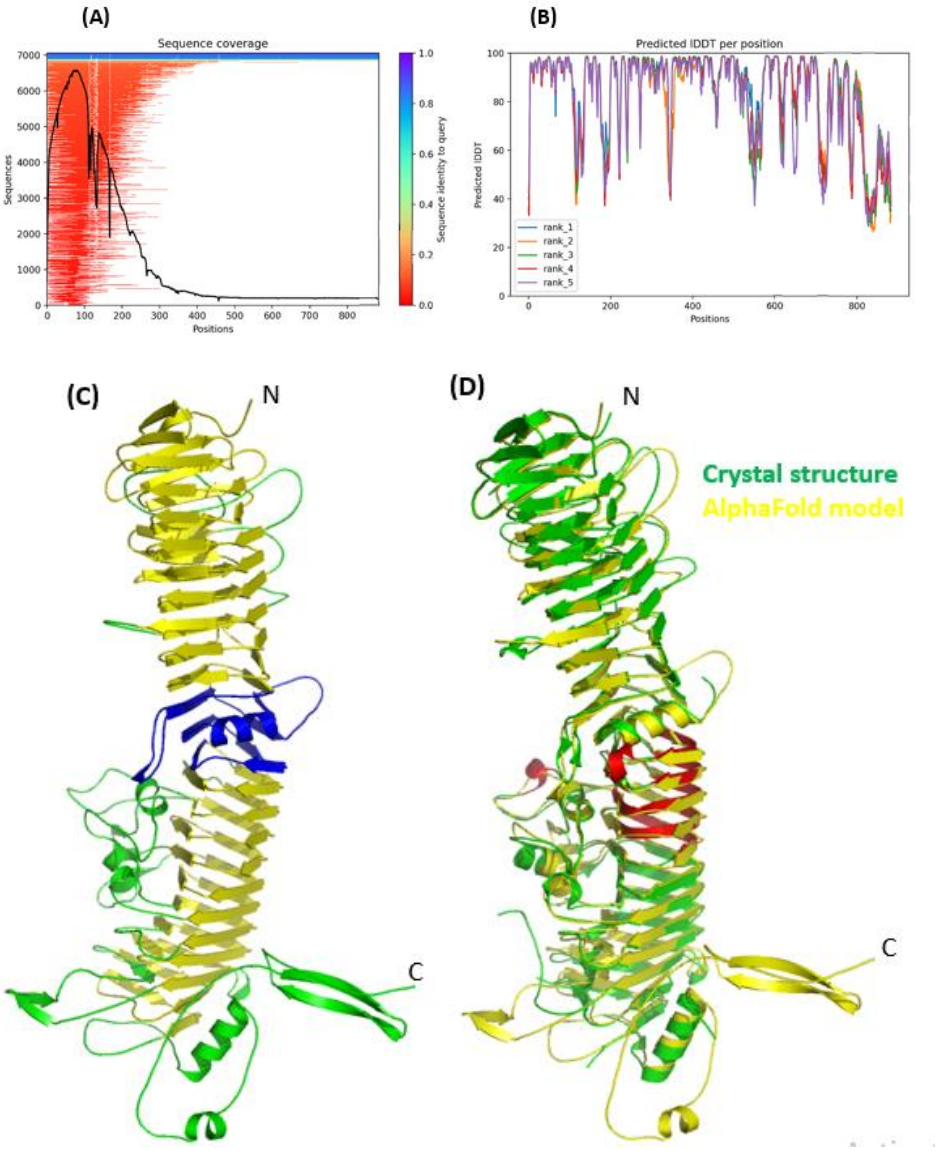



**Figure 4-29: AlphaFold model of HxuA**

**A)** Sequence coverage, **B)** Predicted lDDT score per position. **C)** Ribbon representation of modelled HxuA. The kink region is indicated in red while other flexible regions with low lDDT scores are indicated in green. **D)** Structural alignment of the experimental crystal structure in green and the AlphaFold model in yellow. Structural differences in α-helices and β-strands are marked in red.

Although the overall predicted lDDT score is over 80 %, many individual residues, especially towards the C-terminus, have scores of only 60 % or below (Figure 4-29B). These residues are modelled as extended loops (green, Figure 4-29C). The flexibility of these extended loops

could have contributed to the low predicted lDDT scores at these positions. Very low lDDT scores are also observed between residues 300 and 400 (Figure 4-29). The model includes a slight kink in this region consisting of an α-helix connected to β-strands by an extended loop (blue, Figure 4-29C). This kink region is a potential domain boundary connecting the N- and C-terminal domains of the protein. The overall modelled structure of HxuA thus consist of a β-helix with a slight kink and extended loops (Figure 4-29C). Structural alignment of the experimental structure and the modelled structure revealed some noticeable differences between the two structures in which two α- helices and five β-strands on the experimental crystal structure are modelled into loops on the AlphaFold structure (red, Figure 4-29D). These differences, however, did not result in significant differences between the two structures. This may have been expected as the HxuA structure was presumably part of the PDB database used to train AlphaFold.

# 5 Discussion

This work aimed to structurally and biophysically characterize the protein EtpA from enterotoxigenic *E. coli* (ETEC) as well as confirm and quantify its physical interaction with flagellin (FliC) from the same subspecies. Various protein fragments were designed, corresponding gene constructs prepared in various plasmids, and protein produced and purified. While some constructs could be produced successfully, others proved problematic either in terms of soluble production, purification or oligomerization. Physical characterization by un- and refolding, X-ray crystallography and AlphaFold nevertheless yielded a significant amount of complementary structural and biophysical information on EtpA.

## 5.1 EtpA N-terminal domain and related structures

The N-terminal fragment EtpA$^{1\text{-}606}$ was secreted to the medium. It proved soluble and remarkably stable with a $T_m$ of 94°C. The chemical denaturation induced by urea of EtpA$^{1\text{-}606}$ proved fully reversible, implying a clear folding path and supporting its inherent stability. A stable β-helical fold thus appears particularly suited to a secreted protein allowing it to survive and perform its role in infection in an unpredictable and possibly highly variable environment. While similar stability and/or folding studies for related TpsA proteins have not yet been undertaken, the β-helical autotransporter effector domain of the type 5a secretion protein pertactin from *Bordetella pertussis*, the causative agent of whooping cough, has been investigated in some detail [87]. In pertactin, a related C-terminal domain was also found to be very stable. It was shown to provide a template for the efficient folding of the extended β-helical protein [87, 195]. Note that compared to TpsA proteins (type Vb secretion system), the secretion of autotransporter effector domains from Type Va secretion systems starts at the C-terminus and proceeds towards the N-terminus. Analogous to pertactin, the stable TPS domain would be important for efficient secretion and vectorial folding of full-length EtpA.

β-Helical proteins beyond those from type V secretion systems include meso- and thermo-stable pectate lyases [196] or heat and chaotrope resistant “gene product 5” (Gp5), a spike-shaped trimeric bacteriophage T4 protein [197]. Like EtpA, Gp5 has a well-defined repeat, refolds spontaneously, and forms oligomers [197]. The β-helical fold thus clearly ensures the

production of stable proteins in hostile, extracellular environments that can survive for extended periods of time.

The crystal structure of EtpA$^{67-447}$, as presented here, is the fifth TpsA TPS structure overall. Related TPS domains resolved structurally include those of HMW1-PP, the TPS of high molecular weight adhesin (HMW1A) from *H. influenza*, hemopexin-binding protein (HxuA) also *H. influenza* [91], FHA30 the TPS of filamentous hemagglutinin adhesin (FHA) from *B. pertussis [94]*, and HpmA265, the TPS of hemolysin (HpmA) from *P. mirabilis* [93]. Their sequence alignment and structures are shown in Figure 5-1.

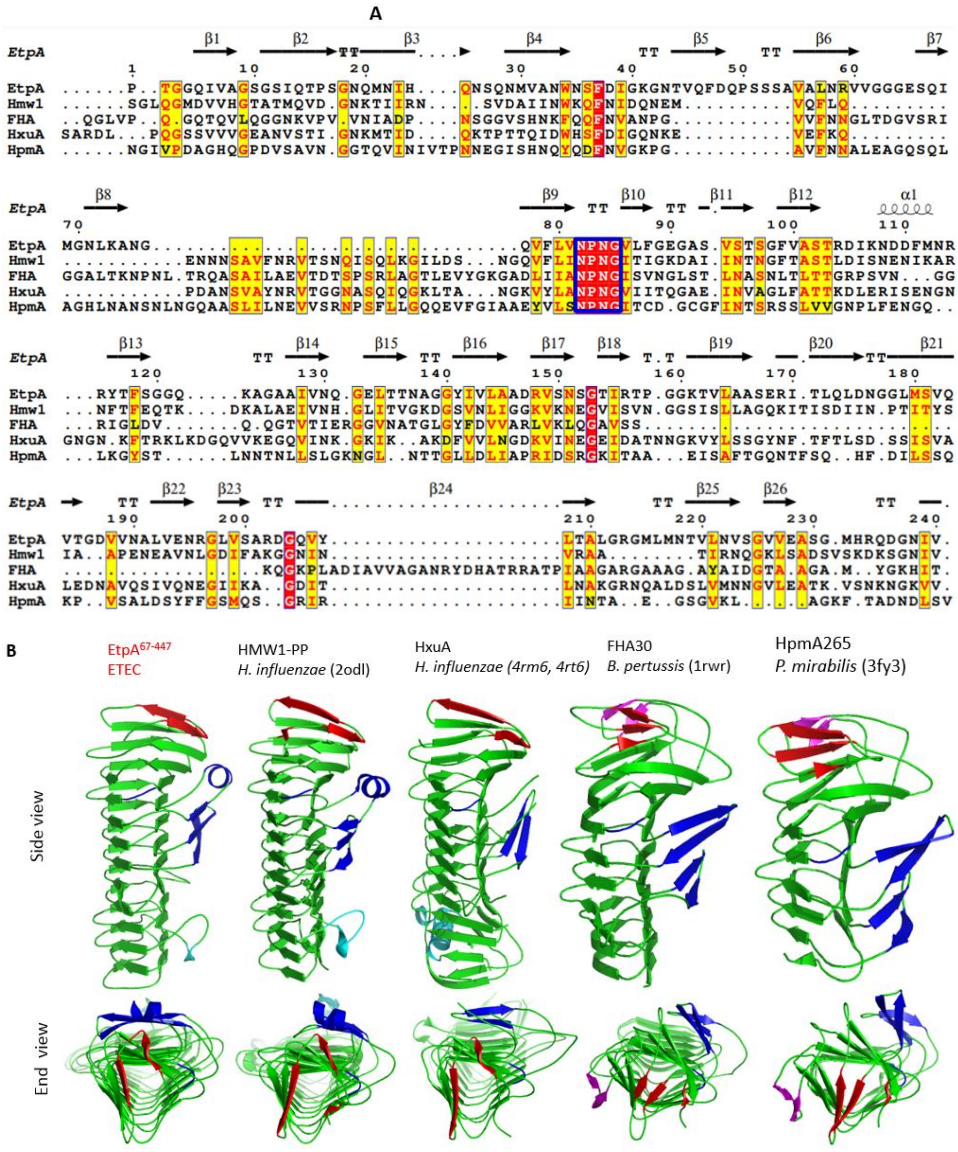



**Figure 5-1: Comparison of selected TpsA**

**A)** A partial, structure-based sequence alignment of the TPS domain of EtpA$^{67-447}$, HxuA301, HMW1-PP, HpmA265 and Fha30. Conserved and mostly conserved residues are shown as white or red coloured text on a red or yellow background respectively. A blue box highlights the conserved NPNG motif.

**B)** TpsA TPS domain crystal structures. Above: Lateral views, below: N-terminal views. The β strands of the N-terminal cap are shown in red, the conserved NPNG motifs and the extra-helical domains in blue, partly-conserved NPNL motifs in magenta, additional extra-helical elements in cyan.

Despite low sequence identities between EtpA and other TPS domains of 23 % for HpmA265 to 35 % for HxuA (Figure 5-1A), the domains all form structurally analogous right-handed β-helices (Figure 5-1B). The β-helices all share tightly packed, hydrophobic cores and distinctive aromatic clusters in the first helical turn: Trp34, Phe47, Phe79, Phe88 and Phe99 in EtpA. N-terminal cap structures shield the hydrophobic cores of all the proteins but differ with respect to the number and arrangement of β-strands (Figure 5-1B). Thus in EtpA$^{67-447}$ and HxuA two β-strands serve this purpose, while in Hmw1-PP, FHA30 and HpmA265 three β-strands are involved. Structural differences between TpsA members presumably reflect their distinct co-evolution with their respective TpsB outer membrane transporters [89].

Another noticeable shared feature of TpsAs are similarly positioned extra-helical domains (blue in Figure 5-1B), though the number and arrangement of secondary structural elements and the insertion points within the β-helix varies extensively. They are consistently located on the outside of the PB1 β-sheet, possibly implying a common role as for homotypic interactions or in biofilm formation [198]. These extra-helical motifs could also provide evolutionary hotspots to generate additional binding sites especially as the rigidity of the β-helix itself may prevent the simple generation of additional functions here without impacting on the overall stability.

EtpA would furthermore need to interact with other bacterial and host structures as part of its role of promoting ETEC adherence and intestinal colonization. The interaction of EtpA with flagellin and glycans in which some of the extra-helical motifs may be required in promoting toxin delivery is undoubtedly one such case. Further functional characterisation of TpsA proteins will be required to reveal the role of individual secondary structural elements in recognition, secretion, folding and additional functions.

A further similarity of TPS domains involves a conserved NPNG motif. Substituting the first asparagine in the motif dramatically reduces secretion rates of FHA and ShlA [193, 199]. A related NPNL motif shared by FHA30 and HpmA265 is, however, absent in EtpA$^{67-447}$, HMW1-PP and HxuA.

## 5.2 Full-length modelled structures of TpsA proteins

The large size and the need for specific TpsB transporters have historically complicated the structural analysis of full-length TpsAs. As a result, only HxuA, the smallest member of the family by far, ever yielded a full-length structure [92]. Correspondingly, the roles of the C-terminal domains are currently not well understood. In HxuA this domain binds hemopexin via its extra-helical motifs [92].

AlphaFold now provides a method to generate structural models of full-length TpsAs (Figure 5-2), expanding the previously available experimental TPS domain structures.

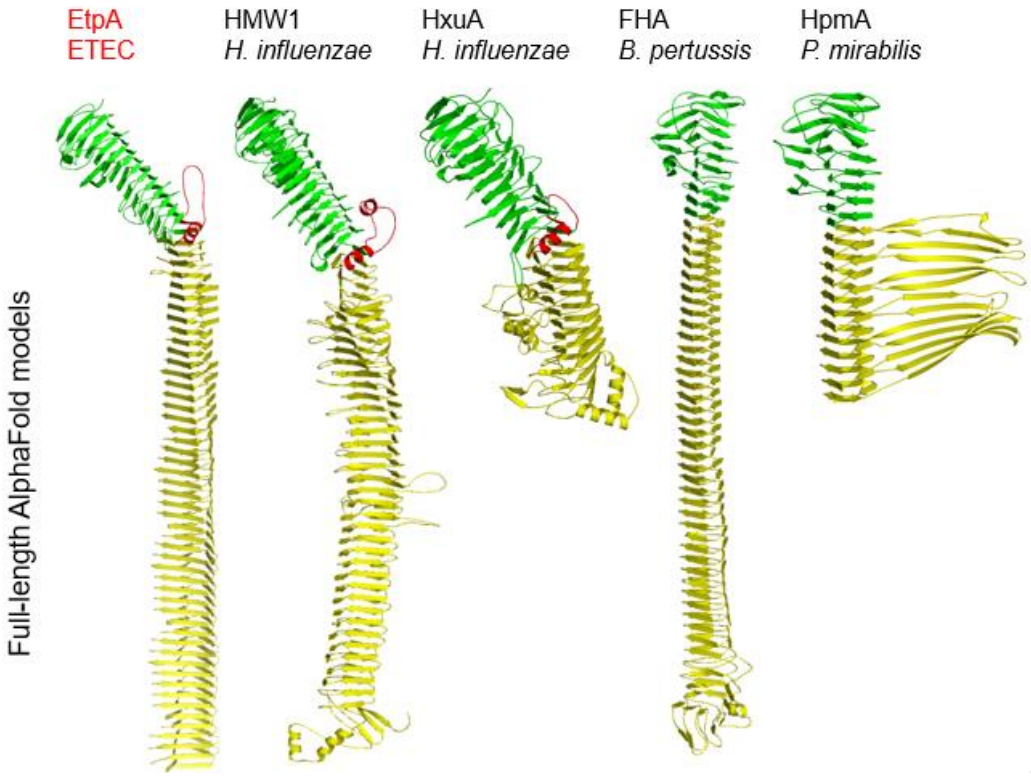



**Figure 5-2: AlphaFold models of full-length TpsA proteins**

TPS containing N-terminal and C-terminal domains are in green and yellow, respectively, interdomain α-helices and adjoining loops in red.

Apart from EtpA, structural models were generated for HMW1, HxuA, FHA and HpmA. In all these proteins, the β-helical fold that defines their TPS domains (Figure 5-1A) extends into the C-terminal domains (Figure 5-2). Nevertheless, both the length and repeat-pattern for the C-terminal β-helices differ appreciably. In HpmA, the C-terminal β-helix appears to be accompanied by a large, extended β-sheet extending laterally away from the β-helix (Figure 5-2). While this unusual fold is supported by RosettaFold [200], its correctness would ideally need to be confirmed experimentally as an isolated β-sheet would be prone to rapid

degradation. In EtpA, HMW1A and HxuA, an α-helix as part of a longer loop insert between the N- and C-terminal domains (red in Figure 5-2) creating a localised kink in the β-helix – more noticeable in EtpA but much less so in HMW1 and HxuA (Figure 5-2). HpmA and FHA lack this α-helix, resulting in essentially linear assemblies (Figure 5-2). In EtpA and HMW1 the β-helix constricts noticeably at the start of the C-terminal domain while the β-helical diameter continues largely unchanged in HxuA, HpmA and FHA, contributing to the single, continuous, and linear β-helix. Apart from the interdomain kink and the constriction of the β-helix, the EtpA C-terminal domain is most similar to HMW1A and FHA in particular with respect to the length and linearity of the C-terminal domain. Functionally, EtpA shares adherence and agglutination properties with HMW1 and FHA [150] [201, 202].

The shape of proteins created from fused repeats may range from linear to circular, helical or twisted based on repeating contributions from each repeat unit [203]. Bends and twists are caused by gradual shifts in repeat proteins through multiple offsets leading to circular or helical rotation around the helix axis. By contrast, kinks create a localized offset of the helical axis. While bends and twists generally have little effect on protein stability, kinks have been shown to affect the stability of some proteins [204]. Correspondingly, kinks may have contributed to the paucity of crystal structures for β-helical proteins. EtpA and HMW1 that contain kinks may also prove less stable overall than the more linear counterparts. Antigen 43a, an adhesin from *E. coli* involved with cell aggregation and biofilm formation, possesses a kink like that of EtpA and HMW1 [205], implying that the kink of EtpA may be of functional importance. Kinks in α-helices are observed in membrane and soluble proteins with kink angles ranging from 6 to 65° [206]. TpsA proteins show some variation in terms of kinks and twists, but less in terms of bends especially if compared to other superhelical proteins such as leucine-rich repeat proteins where curved structures provide a clasped binding surface for protein-protein interactions [207]. Interestingly, antigens 43a (4KH3) and 43b (7KOB) share 90 % sequence identity yet demonstrate different bending curvatures [205].

Apart from subtype Vb secreted proteins such as TpsAs, β-helical structures are also typical for passenger domains of the closely related subtype Va secretion systems or autotransporters (ATs). Examples include antigens 43a (4KH3) and 43b (7KOB) as well as SepA (5J44) [205, 208]. Other β-helical proteins include tail spike protein of *E. coli* bacteriophage HK620 (2VJI) [209] and the AFP antifreeze protein [210]. A common feature of these β-helical proteins is their physical stability in a challenging extracellular environment. Potentially, the

thermodynamics of β-helix formation may aid secretion by providing a ratchet mechanism helping to move the protein through the pore in a single direction and offsetting a lack of accessible energy outside the outer membrane [211].

TpsA proteins typically contain complex sequence repeats with those in EtpA being among the longest. Repeat domains, including those of TpsAs, are assumed to evolve by internal gene duplications and recombination processes [203]. Genetically, such repeats may presumably be easily extended through multiple recombination events creating longer proteins with large surfaces for possible interaction with other proteins or surfaces. Protein repeats of various patterns are common in proteins and critically contribute to their function [212]. Repeats may vary from single amino acid repeats, to short repeats of 20 to 40 amino acids, and larger repeats of more than 100 amino acids [212]. Such repeats can include distinct domains of defined structure and function linked by disordered loops [213] or directly fused repeating units as observed in EtpA. Many proteins larger than 500 amino acids contain extensive repeats [214]. The complexity of repeats, however, means that no single annotation system can capture all resulting properties. Functions known to arise from protein repeats include protein-protein interactions, nucleotide-binding, signal transduction, antiviral response and virulence [203]. Thus while leucine-rich repeats are involved in protein-protein interactions, zinc finger transcription factors bind DNA, *S. pyogenes* proteins with DUF1542 domain repeats enable adhesion and antibiotic resistance [215]. EtpA may have a similar function. Some repeat-rich proteins give rise to antigenic behaviours [216]. The EtpA repeats could account for its antigenic traits allowing it to interact with glycans in blood group A individuals [152] with depressions between successive EtpA repeats providing binding pockets for molecules such glycans and flagellin.

An early hypothesis for TpsA translocation and folding suggested that the TPS domain would remain bound to the POTRA domains while the remaining protein was translocated and folded on the cell surface [140] . A latter model suggested that TPS domains initiate folding upon secretion [89]. This suggestion is supported by the stable TPS fold, the efficient secretion of truncated N-terminal domains of TpsA proteins, TPS domain-initiated TpsA folding *in vitro* [139, 143], and the accessibility of N-termini of stalled FHA constructs at the cell surface [93, 141, 142]. We propose that the TPS domain may furthermore serve as a template to facilitate folding of the C-terminal domain.

## 5.3 Investigating the interaction with flagellin

Protein-protein interactions are a critical component of most infectious diseases. A typical example is the interaction between the severe acute respiratory syndrome coronavirus (SARS-CoV-2) spike protein and the human angiotensin converting enzyme-2 (ACE-2) [217, 218]. In this project the interaction of two N-terminal domains of EtpA and flagellin from ETEC was investigated. However, both SEC and molecular pulldown assays failed to show any direct interaction between the two N-terminal domain fragments and flagellin. The finding differ from that from a previous study in which molecular pulldown assay was used to report that the first 560 residues of EtpA are enough to interact with flagellin [13]. Our findings could possibly suggest that other factors might be required for the interaction to happen. Our SEC findings with EtpA$^{1-606}$ were not conclusive considering that the protein eluted with a similar elution volume with flagellin, making it difficult to resolve any shift in the peaks that could have been because of a complex between the two proteins. This therefore open the doors for other interaction techniques such as microscale thermophoresis (MST) to be applicable. As for those other factors that might be required for the interaction between the two proteins to happen, future work can be done to investigate them. Factors such as varying incubation time, different incubation temperature and buffers should be part of such future investigations. This project has generated several constructs that would facilitate such future investigations.

## 5.4 Production of EtpA extended fragment

Characterization of most TpsA proteins to date has been limited to their TPS domain, with HxuA from *H. influenza* being the only full-length protein of this family that has been studied. We attempted to characterize EtpA protein from ETEC beyond its TPS domain by producing a fragment that incorporates the N-terminal domain (residues 67-447) and the central region (residue 448-930) giving rise to EtpA$^{67-930}$. Obtained results indicated successful cloning and production of GST-EtpA$^{67-930}$ fusion protein. However, most of the protein was being incorporated into inclusion bodies, a clear stress response to when too much protein is being produced, the protein is unable to fold timeously, or if the protein poses a physical or functional threat to the cell. While the inclusion bodies could be solubilized, exploratory attempts to refold the protein proved unsuccessful. The failure to refold the longer fragments of EtpA may be due to factors such as its high molecular weight, producing it the cytoplasm

rather than the extracellular milieu, and the absence of its secretion partner which presumably functions as both a pore and a folding-chaperone. Producing EtpA$^{67-930}$ in the cytoplasm without its secretion partner may thus have presented the main stumbling block in allowing it to fold. By contrast, EtpA$^{67-447}$ was successfully produced in the cytoplasm. In fact, inclusion bodies for EtpA$^{67-447}$ could be solubilized and successfully refolded (urea denaturation studies). Similarly, EtpA$^{1-606}$ could be chemically denatured and refolded. The only difference between these EtpA constructs is the extended C-terminal region in EtpA$^{67-930}$. It would seem that the extended C-terminal domain blocked the folding path either by slowing the process, interfering with the folding of the N-terminal domain, or by preventing the folded N-terminal domain to function as a folding nucleus for the C-terminal domain. While EtpA$^{1-606}$ was natively secreted by EtpB, the potential chaperone function of EtpB would be absent during the refolding process after urea denaturation *in vitro*. Oddly, the production of GST-EtpA$^{67-606}$ within the cytoplasm using pGEX-6P-2 vector also resulted in quantitatively insoluble protein, implying that its folding path may already be impaired or that its physical size or unknown metabolic functions induced the *E. coli* cells to mount a regulatory response that sequestered the protein into the inclusion bodies. Full-length EtpA has successfully been produced by native secretion [150], though the protein yield was low, possibly implying inefficiencies as a result of the length of the polypeptide chain.

# 6 Conclusion

In summary, we used CD spectroscopy to show that the EtpA N-terminal TPS domain forms thermostable and urea-stable folds that efficiently refold after denaturation. We further used X-ray crystallography and structure modelling by AlphaFold to demonstrate that the β-helical structure of the EtpA and related N-terminal domains provide a possible template upon which the C-terminal domain can efficiently fold. Accumulating structural data on EtpA and related virulence factors provide an increasingly clearer understanding of this important family of proteins. These concepts could inform the design of novel microproteins for vaccine development. While β-helical fragments remain challenging to produce, their inherent stability once formed could offer highly attractive vaccine candidates against a range of pathogens. While N- and C-terminal capping domains would inevitably be required to stabilize such mini-proteins the combination with an N-terminal fragment of a readily folding protein could help to successfully nucleate the β-helical folding process. Functional studies between EtpA N-terminal domain and flagellin with size exclusion chromatography and molecular pulldown assays failed to show any interaction implying that other factors may be involved. This opens the platform for further future investigation. Though we have successfully used AlphaFold model to show that there are no major differences between the X-ray crystal structure of $EtpA^{67-447}$ and its corresponding AlphaFold model, the question that remains is whether X-ray crystallography will still be required to determine protein structures. Part of the answer to this question is to look at proteins beyond their apo-structures where some are required to interact with other molecules or with other proteins to perform vital roles in pathogenesis in the case of ETEC. For instance the interaction of EtpA with FliC. For the past decades, X-ray crystallography have been the main experimental techniques used to determine how proteins interaction with non-proteins molecules and other proteins. Can we also use AlphaFold to obtain such vital information? Certainly not at this moment, though it is obvious that such platforms are undergoing investigation. While we await such mega platforms to be put in place, X-ray crystallography will continue to play an impact in our understanding of protein interactions. Even if we do manage to put such platforms in place, it is possible that a combination with X-ray crystallography might provide some validation.

# 7 References


1. Escherich T: **Die Darmbacterien des Neugeborenen und Säuglings**. 1885.
2. Escobar-Páramo P, Giudicelli C, Parsot C, Denamur E: **The evolutionary history of *Shigella* and enteroinvasive *Escherichia coli* revised**. *J Mol Evol* 2003, **57**(2):140-148.
3. Kaper JB, Nataro JP, Mobley HL: **Pathogenic *Escherichia coli***. *Nat Rev Microbiol* 2004, **2**(2):123-140.
4. Lindstedt BA, Finton MD, Porcellato D, Brandal LT: **High frequency of hybrid *Escherichia coli* strains with combined Intestinal pathogenic *Escherichia coli* (IPEC) and extraintestinal pathogenic *Escherichia coli* (ExPEC) virulence factors isolated from human faecal samples**. *BMC Infect Dis* 2018, **18**(1):544.
5. Kaniuk NA, Vinogradov E, Li J, Monteiro MA, Whitfield C: **Chromosomal and plasmid-encoded enzymes are required for assembly of the R3-type core oligosaccharide in the lipopolysaccharide of *Escherichia coli* O157:H7**. *J Biol Chem* 2004, **279**(30):31237-31250.
6. Nataro JP, Kaper JB: **Diarrheagenic *Escherichia coli***. *Clin Microbiol Rev* 1998, **11**(1):142-201.
7. Paton AW, Paton JC: **Direct detection of Shiga toxigenic *Escherichia coli* strains belonging to serogroups O111, O157, and O113 by multiplex PCR**. *J Clin Microbiol* 1999, **37**(10):3362-3365.
8. Melton-Celsa AR, Darnell SC, O'Brien AD: **Activation of Shiga-like toxins by mouse and human intestinal mucus correlates with virulence of enterohemorrhagic *Escherichia coli* O91:H21 isolates in orally infected, streptomycin-treated mice**. *Infect Immun* 1996, **64**(5):1569-1576.
9. Chahales P, Thanassi DG: **Structure, function, and assembly of adhesive organelles by uropathogenic bacteria**. *Microbiol Spectr* 2015, **3**(5).
10. Jubelin G, Desvaux M, Schüller S, Etienne-Mesmin L, Muniesa M, Blanquet-Diot S: **Modulation of enterohaemorrhagic *Escherichia coli* survival and virulence in the human gastrointestinal tract**. *Microorganisms* 2018, **6**(4):115.
11. Read LT, Hahn RW, Thompson CC, Bauer DL, Norton EB, Clements JD: **Simultaneous exposure to *Escherichia coli* heat-labile and heat-stable enterotoxins increases fluid secretion and alters cyclic nucleotide and cytokine production by intestinal epithelial cells**. *Infect Immun* 2014, **82**(12):5308-5316.
12. Gaastra W, Svennerholm AM: **Colonization factors of human enterotoxigenic *Escherichia coli* (ETEC)**. *Trends Microbiol* 1996, **4**(11):444-452.
13. Roy K, Hilliard GM, Hamilton DJ, Luo J, Ostmann MM, Fleckenstein JM: **Enterotoxigenic *Escherichia coli* EtpA mediates adhesion between flagella and host cells**. *Nature* 2009, **457**(7229):594-598.
14. Amézquita-López B, Quiñones B, Lee B, Chaidez C: **Virulence profiling of Shiga toxin-producing *Escherichia coli* recovered from domestic farm animals in Northwestern Mexico**. *Frontiers in Cellular and Infection Microbiology* 2014, **4**.
15. Etcheverría AI, Padola NL: **Shiga toxin-producing *Escherichia coli*: factors involved in virulence and cattle colonization**. *Virulence* 2013, **4**(5):366-372.
16. Griffin PM, Tauxe RV: **The epidemiology of infections caused by *Escherichia coli* O157:H7, other enterohemorrhagic E. coli, and the associated hemolytic uremic syndrome**. *Epidemiol Rev* 1991, **13**:60-98.
17. Sandvig K: **Shiga toxins**. *Toxicon* 2001, **39**(11):1629-1635.
18. Gyles CL: **Shiga toxin-producing *Escherichia coli*: an overview**. *J Anim Sci* 2007, **85**(13 Suppl):E45-62.
19. Blanco M, Blanco JE, Mora A, Dahbi G, Alonso MP, González EA, Bernárdez MI, Blanco J: **Serotypes, virulence genes, and intimin types of Shiga toxin (verotoxin)-producing *Escherichia coli* isolates from cattle in Spain and identification of a new intimin variant gene (eae-xi)**. *J Clin Microbiol* 2004, **42**(2):645-651.
20. Guth BE, Prado V, Rivas M: **Shiga toxin-producing *Escherichia coli***. *Pathogenic Escherichia coli in Latin America, AG Torres (Ed)* 2010:65-83.

21. Trabulsi LR, Keller R, Tardelli Gomes TA: **Typical and atypical enteropathogenic *Escherichia coli***. *Emerg Infect Dis* 2002, **8**(5):508-513.
22. Girón JA, Ho AS, Schoolnik GK: **An inducible bundle-forming pilus of enteropathogenic *Escherichia coli***. *Science* 1991, **254**(5032):710-713.
23. Dean P, Kenny B: **The effector repertoire of enteropathogenic E. coli: ganging up on the host cell**. *Curr Opin Microbiol* 2009, **12**(1):101-109.
24. Litvak Y, Sharon S, Hyams M, Zhang L, Kobi S, Katsowich N, Dishon S, Nussbaum G, Dong N, Shao F *et al*: **Epithelial cells detect functional type III secretion system of enteropathogenic *Escherichia coli* through a novel NF-κB signaling pathway**. *PLOS Pathogens* 2017, **13**(7):e1006472.
25. Ochoa TJ, Contreras CA: **Enteropathogenic *Escherichia coli* infection in children**. *Curr Opin Infect Dis* 2011, **24**(5):478-483.
26. Cepeda-Molero M, Berger CN, Walsham ADS, Ellis SJ, Wemyss-Holden S, Schüller S, Frankel G, Fernández L: **Attaching and effacing (A/E) lesion formation by enteropathogenic E. coli on human intestinal mucosa is dependent on non-LEE effectors**. *PLoS Pathog* 2017, **13**(10):e1006706.
27. Henderson IR, Czeczulin J, Eslava C, Noriega F, Nataro JP: **Characterization of pic, a secreted protease of Shigella flexneri and enteroaggregative *Escherichia coli***. *Infect Immun* 1999, **67**(11):5587-5596.
28. Nataro JP, Yikang D, Yingkang D, Walker K: **AggR, a transcriptional activator of aggregative adherence fimbria I expression in enteroaggregative *Escherichia coli***. *J Bacteriol* 1994, **176**(15):4691-4699.
29. Harris CL, Abbott RJ, Smith RA, Morgan BP, Lea SM: **Molecular dissection of interactions between components of the alternative pathway of complement and decay accelerating factor (CD55)**. *J Biol Chem* 2005, **280**(4):2569-2578.
30. Pangburn MK: **Differences between the binding sites of the complement regulatory proteins DAF, CR1, and factor H on C3 convertases**. *J Immunol* 1986, **136**(6):2216-2221.
31. Bernet-Camard MF, Coconnier MH, Hudault S, Servin AL: **Pathogenicity of the diffusely adhering strain *Escherichia coli* C1845: F1845 adhesin-decay accelerating factor interaction, brush border microvillus injury, and actin disassembly in cultured human intestinal epithelial cells**. *Infect Immun* 1996, **64**(6):1918-1928.
32. Bilge SS, Clausen CR, Lau W, Moseley SL: **Molecular characterization of a fimbrial adhesin, F1845, mediating diffuse adherence of diarrhea-associated *Escherichia coli* to HEp-2 cells**. *J Bacteriol* 1989, **171**(8):4281-4289.
33. Harris JR, Wachsmuth IK, Davis BR, Cohen ML: **High-molecular-weight plasmid correlates with *Escherichia coli* enteroinvasiveness**. *Infection and Immunity* 1982, **37**(3):1295-1298.
34. Sansonetti P, Kopecko D, Formal S: **Involvement of a plasmid in the invasive ability of Shigella flexneri**. *Infection and immunity* 1982, **35**(3):852-860.
35. Schaetti C: **Vaccines for enteric diseases: Update on recent developments**, vol. 8; 2009.
36. Giddings SL, Stevens AM, Leung DT: **Traveler's diarrhea**. *Med Clin North Am* 2016, **100**(2):317-330.
37. Dubreuil JD, Isaacson RE, Schifferli DM: **Animal enterotoxigenic *Escherichia coli***. *EcoSal Plus* 2016, **7**(1).
38. Gupta SK, Keck J, Ram PK, Crump JA, Miller MA, Mintz ED: **Part III. Analysis of data gaps pertaining to enterotoxigenic *Escherichia coli* infections in low and medium human development index countries, 1984-2005**. *Epidemiol Infect* 2008, **136**(6):721-738.
39. Levine MM, Nalin DR, Hoover DL, Bergquist EJ, Hornick RB, Young CR: **Immunity to enterotoxigenic *Escherichia coli***. *Infect Immun* 1979, **23**(3):729-736.
40. Qadri F, Svennerholm AM, Faruque AS, Sack RB: **Enterotoxigenic *Escherichia coli* in developing countries: epidemiology, microbiology, clinical features, treatment, and prevention**. *Clin Microbiol Rev* 2005, **18**(3):465-483.

41. Smith JL, Fratamico PM, Gunther NWt: **Shiga toxin-producing *Escherichia coli***. *Adv Appl Microbiol* 2014, **86**:145-197.
42. Connolly JPR, Finlay BB, Roe AJ: **From ingestion to colonization: the influence of the host environment on regulation of the LEE encoded type III secretion system in enterohaemorrhagic *Escherichia coli***. *Frontiers in microbiology* 2015, **6**:568-568.
43. DuPont HL, Ericsson CD, Mathewson JJ, DuPont MW: **Five versus three days of ofloxacin therapy for traveler's diarrhea: a placebo-controlled study**. *Antimicrob Agents Chemother* 1992, **36**(1):87-91.
44. Ahmed T, Bhuiyan TR, Zaman K, Sinclair D, Qadri F: **Vaccines for preventing enterotoxigenic *Escherichia coli* (ETEC) diarrhoea**. *Cochrane Database Syst Rev* 2013, **2013**(7):Cd009029.
45. Stintzing G, Möllby R: **Colonization of the upper jejunum by enteropathogenic and enterotoxigenic *Escherichia coli* in paediatric diarrhoea**. *Acta Paediatr Scand* 1982, **71**(3):457-465.
46. Allen KP, Randolph MM, Fleckenstein JM: **Importance of heat-labile enterotoxin in colonization of the adult mouse small intestine by human enterotoxigenic *Escherichia coli* strains**. *Infect Immun* 2006, **74**(2):869-875.
47. Zhao B, Houry WA: **Acid stress response in enteropathogenic gammaproteobacteria: an aptitude for survival**. *Biochem Cell Biol* 2010, **88**(2):301-314.
48. Gaastra W, Svennerholm A-M: **Colonization factors of human enterotoxigenic *Escherichia coli* (ETEC)**. *Trends in Microbiology* 1996, **4**(11):444-452.
49. Anantha RP, McVeigh AL, Lee LH, Agnew MK, Cassels FJ, Scott DA, Whittam TS, Savarino SJ: **Evolutionary and functional relationships of colonization factor antigen i and other class 5 adhesive fimbriae of enterotoxigenic *Escherichia coli***. *Infect Immun* 2004, **72**(12):7190-7201.
50. Jansson L, Tobias J, Lebens M, Svennerholm AM, Teneberg S: **The major subunit, CfaB, of colonization factor antigen i from enterotoxigenic *Escherichia coli* is a glycosphingolipid binding protein**. *Infect Immun* 2006, **74**(6):3488-3497.
51. Mottram L, Liu J, Chavan S, Tobias J, Svennerholm A-M, Holgersson J: **Glyco-engineered cell line and computational docking studies reveals enterotoxigenic *Escherichia coli* CFA/I fimbriae bind to Lewis a glycans**. *Scientific Reports* 2018, **8**(1):11250.
52. Duthy TG, Manning PA, Heuzenroeder MW: **Identification and characterization of assembly proteins of CS5 pili from enterotoxigenic *Escherichia coli***. *J Bacteriol* 2002, **184**(4):1065-1077.
53. Fleckenstein JM, Hardwidge PR, Munson GP, Rasko DA, Sommerfelt H, Steinsland H: **Molecular mechanisms of enterotoxigenic *Escherichia coli* infection**. *Microbes and Infection* 2010, **12**(2):89-98.
54. Mammarappallil JG, Elsinghorst EA: **Epithelial cell adherence mediated by the enterotoxigenic *Escherichia coli* tia protein**. *Infect Immun* 2000, **68**(12):6595-6601.
55. Brown E, Hardwidge P: **Biochemical characterization of the enterotoxigenic *Escherichia coli* LeoA protein**. *Microbiology (Reading, England)* 2007, **153**:3776-3784.
56. Luo Q, Kumar P, Vickers TJ, Sheikh A, Lewis WG, Rasko DA, Sistrunk J, Fleckenstein JM: **Enterotoxigenic *Escherichia coli* secretes a highly conserved mucin-degrading metalloprotease to effectively engage intestinal epithelial cells**. *Infection and immunity* 2014, **82**(2):509-521.
57. Duan Q, Xia P, Nandre R, Zhang W, Zhu G: **Review of newly identified functions associated with the heat-labile toxin of enterotoxigenic *Escherichia coli***. *Frontiers in Cellular and Infection Microbiology* 2019, **9**.
58. Dubreuil JD: **The whole Shebang: the gastrointestinal tract, *Escherichia coli* enterotoxins and secretion**. *Curr Issues Mol Biol* 2012, **14**(2):71-82.
59. Spangler BD: **Structure and function of cholera toxin and the related *Escherichia coli* heat-labile enterotoxin**. *Microbiol Rev* 1992, **56**(4):622-647.
60. Johnson WM, Lior H, Johnson KG: **Heat-stable enterotoxin from *Escherichia coli*: factors involved in growth and toxin production**. *Infect Immun* 1978, **20**(2):352-359.

61. Kunkel SL, Robertson DC: **Factors affecting release of heat-labile enterotoxin by enterotoxigenic *Escherichia coli***. *Infect Immun* 1979, **23**(3):652-659.
62. Hegde A, Bhat GK, Mallya S: **Effect of stress on production of heat labile enterotoxin by *Escherichia coli***. *Indian J Med Microbiol* 2009, **27**(4):325-328.
63. Gonzales L, Ali ZB, Nygren E, Wang Z, Karlsson S, Zhu B, Quiding-Järbrink M, Sjöling Å: **Alkaline pH Is a Signal for Optimal production and secretion of the heat labile toxin, LT in enterotoxigenic *Escherichia coli* (ETEC)**. *PLOS ONE* 2013, **8**(9):e74069.
64. Chatterjee A, Chowdhury R: **Bile and unsaturated fatty acids inhibit the binding of cholera toxin and *Escherichia coli* heat-labile enterotoxin to GM1 receptor**. *Antimicrob Agents Chemother* 2008, **52**(1):220-224.
65. Sahl JW, Rasko DA: **Analysis of global transcriptional profiles of enterotoxigenic *Escherichia coli* isolate E24377A**. *Infection and immunity* 2012, **80**(3):1232-1242.
66. Sjöling A, Wiklund G, Savarino SJ, Cohen DI, Svennerholm AM: **Comparative analyses of phenotypic and genotypic methods for detection of enterotoxigenic *Escherichia coli* toxins and colonization factors**. *J Clin Microbiol* 2007, **45**(10):3295-3301.
67. Grewal HM, Valvatne H, Bhan MK, van Dijk L, Gaastra W, Sommerfelt H: **A new putative fimbrial colonization factor, CS19, of human enterotoxigenic *Escherichia coli***. *Infection and immunity* 1997, **65**(2):507-513.
68. Haines S, Gautheron S, Nasser W, Renauld-Mongénie G: **Identification of novel components influencing colonization factor antigen I expression in enterotoxigenic *Escherichia coli***. *PLoS One* 2015, **10**(10):e0141469.
69. Rappaport RS, Sagin JF, Pierzchala WA, Bonde G, Rubin BA, Tint H: **Activation of Heat-labile *Escherichia coli* enterotoxin by trypsin**. *J Infect Dis* 1976, **133 Suppl**:41-54.
70. Aranda KR, Fabbricotti SH, Fagundes-Neto U, Scaletsky IC: **Single multiplex assay to identify simultaneously enteropathogenic, enteroaggregative, enterotoxigenic, enteroinvasive and Shiga toxin-producing *Escherichia coli* strains in Brazilian children**. *FEMS Microbiol Lett* 2007, **267**(2):145-150.
71. Youmans BP, Ajami NJ, Jiang ZD, Petrosino JF, DuPont HL, Highlander SK: **Development and accuracy of quantitative real-time polymerase chain reaction assays for detection and quantification of enterotoxigenic *Escherichia coli* (ETEC) heat labile and heat stable toxin genes in travelers' diarrhea samples**. *Am J Trop Med Hyg* 2014, **90**(1):124-132.
72. Evans DG, Evans DJ, Pierce NF: **Differences in the Response of Rabbit Small Intestine to Heat-Labile and Heat-Stable Enterotoxins of *Escherichia coli***. *Infection and Immunity* 1973, **7**(6):873-880.
73. Dean AG, Ching Y-C, Williams RG, Harden LB: **Test for *Escherichia coli* enterotoxin using infant mice: Application in a study of diarrhea in children in Honolulu**. *The Journal of Infectious Diseases* 1972, **125**(4):407-411.
74. Guerrant RL, Brunton LL, Schnaitman TC, Rebhun LI, Gilman AG: **Cyclic adenosine monophosphate and alteration of chinese hamster ovary cell morphology: a rapid, sensitive *in vitro* assay for the enterotoxins of *Vibrio cholerae* and *Escherichia coli***. *Infection and Immunity* 1974, **10**(2):320-327.
75. Giannella RA, Drake KW, Luttrell M: **Development of a radioimmunoassay for *Escherichia coli* heat-stable enterotoxin: comparison with the suckling mouse bioassay**. *Infection and Immunity* 1981, **33**(1):186-192.
76. Cryan B: **Comparison of three assay systems for detection of enterotoxigenic *Escherichia coli* heat-stable enterotoxin**. *Journal of Clinical Microbiology* 1990, **28**(4):792-794.
77. Bäck E, Svennerholm AM, Holmgren J, Möllby R: **Evaluation of a ganglioside immunosorbent assay for detection of *Escherichia coli* heat-labile enterotoxin**. *Journal of Clinical Microbiology* 1979, **10**(6):791-795.
78. Boedeker EC: **Vaccines for enterotoxigenic *Escherichia coli*: current status**. *Curr Opin Gastroenterol* 2005, **21**(1):15-19.

79. Svennerholm AM, Tobias J: **Vaccines against enterotoxigenic *Escherichia coli***. *Expert Rev Vaccines* 2008, **7**(6):795-804.
80. Roy K, Bartels S, Qadri F, Fleckenstein JM: **Enterotoxigenic *Escherichia coli* elicits immune responses to multiple surface proteins**. *Infection and Immunity* 2010, **78**(7):3027-3035.
81. Zhang W, Sack DA: **Progress and hurdles in the development of vaccines against enterotoxigenic *Escherichia coli* in humans**. *Expert Rev Vaccines* 2012, **11**(6):677-694.
82. Zhang W, Sack DA: **Current progress in developing subunit vaccines against enterotoxigenic *Escherichia coli*-associated diarrhea**. *Clin Vaccine Immunol* 2015, **22**(9):983-991.
83. Mentzer Av, Tobias J, Wiklund G, Nordqvist S, Aslett M, Dougan G, Sjöling Å, Svennerholm A-M: **Identification and characterization of the novel colonization factor CS30 based on whole genome sequencing in enterotoxigenic *Escherichia coli* (ETEC)**. *Scientific Reports* 2017, **7**(1):12514.
84. Walker RI: **An assessment of enterotoxigenic *Escherichia coli* and Shigella vaccine candidates for infants and children**. *Vaccine* 2015, **33**(8):954-965.
85. Green ER, Mecsas J: **Bacterial secretion systems: An overview**. *Microbiol Spectr* 2016, **4**(1).
86. Natale P, Brüser T, Driessen AJ: **Sec- and Tat-mediated protein secretion across the bacterial cytoplasmic membrane--distinct translocases and mechanisms**. *Biochim Biophys Acta* 2008, **1778**(9):1735-1756.
87. Junker M, Schuster CC, McDonnell AV, Sorg KA, Finn MC, Berger B, Clark PL: **Pertactin β-helix folding mechanism suggests common themes for the secretion and folding of autotransporter proteins**. *Proc Natl Acad Sci U S A* 2006, **103**(13):4918-4923.
88. Meuskens I, Saragliadis A, Leo JC, Linke D: **Type V secretion systems: An overview of passenger domain functions**. *Frontiers in Microbiology* 2019, **10**(1163).
89. Guérin J, Bigot S, Schneider R, Buchanan SK, Jacob-Dubuisson F: **Two-partner secretion: combining efficiency and simplicity in the secretion of large proteins for bacteria-host and bacteria-bacteria interactions**. *Front Cell Infect Microbiol* 2017, **7**:148.
90. Junker M, Clark PL: **Slow formation of aggregation-resistant β-sheet folding intermediates**. *Proteins* 2010, **78**(4):812-824.
91. Yeo HJ, Yokoyama T, Walkiewicz K, Kim Y, Grass S, Geme JW, 3rd: **The structure of the *Haemophilus influenzae* HMW1 pro-piece reveals a structural domain essential for bacterial two-partner secretion**. *J Biol Chem* 2007, **282**(42):31076-31084.
92. Zambolin S, Clantin B, Chami M, Hoos S, Haouz A, Villeret V, Delepelaire P: **Structural basis for haem piracy from host haemopexin by *Haemophilus influenzae***. *Nature Communications* 2016, **7**(1):11590.
93. Weaver TM, Hocking JM, Bailey LJ, Wawrzyn GT, Howard DR, Sikkink LA, Ramirez-Alvarado M, Thompson JR: **Structural and functional studies of truncated hemolysin A from *Proteus mirabilis***. *J Biol Chem* 2009, **284**(33):22297-22309.
94. Clantin B, Hodak H, Willery E, Locht C, Jacob-Dubuisson F, Villeret V: **The crystal structure of filamentous hemagglutinin secretion domain and its implications for the two-partner secretion pathway**. *Proc Natl Acad Sci U S A* 2004, **101**(16):6194-6199.
95. ur Rahman S, van Ulsen P: **System specificity of the TpsB transporters of coexpressed two-partner secretion systems of *Neisseria meningitidis***. *J Bacteriol* 2013, **195**(4):788-797.
96. Braun V, Hobbie S, Ondraczek R: ***Serratia marcescens* forms a new type of cytolysin**. *FEMS Microbiol Lett* 1992, **100**(1-3):299-305.
97. Hertle R, Hilger M, Weingardt-Kocher S, Walev I: **Cytotoxic action of *Serratia marcescens* hemolysin on human epithelial cells**. *Infect Immun* 1999, **67**(2):817-825.
98. Könninger UW, Hobbie S, Benz R, Braun V: **The haemolysin-secreting ShlB protein of the outer membrane of Serratia marcescens: determination of surface-exposed residues and formation of ion-permeable pores by ShlB mutants in artificial lipid bilayer membranes**. *Molecular microbiology* 1999, **32**(6):1212-1225.

99. Uphoff TS, Welch RA: **Nucleotide sequencing of the *Proteus mirabilis* calcium-independent hemolysin genes (hpmA and hpmB) reveals sequence similarity with the Serratia marcescens hemolysin genes (shlA and shlB)**. *J Bacteriol* 1990, **172**(3):1206-1216.
100. Swihart KG, Welch RA: **Cytotoxic activity of the Proteus hemolysin HpmA**. *Infect Immun* 1990, **58**(6):1861-1869.
101. Hirono I, Tange N, Aoki T: **Iron-regulated haemolysin gene from *Edwardsiella tarda***. *Mol Microbiol* 1997, **24**(4):851-856.
102. Strauss EJ, Ghori N, Falkow S: **An *Edwardsiella tarda* strain containing a mutation in a gene with homology to shlB and hpmB is defective for entry into epithelial cells in culture**. *Infect Immun* 1997, **65**(9):3924-3932.
103. Wang X, Wang Q, Xiao J, Liu Q, Wu H, Zhang Y: **Hemolysin EthA in *Edwardsiella tarda* is essential for fish invasion *in vivo* and *in vitro* and regulated by two-component system EsrA-EsrB and nucleoid protein HhaEt**. *Fish Shellfish Immunol* 2010, **29**(6):1082-1091.
104. Palmer KL, Munson RS, Jr.: **Cloning and characterization of the genes encoding the hemolysin of *Haemophilus ducreyi***. *Mol Microbiol* 1995, **18**(5):821-830.
105. Brillard J, Duchaud E, Boemare N, Kunst F, Givaudan A: **The PhlA hemolysin from the entomopathogenic bacterium *Photorhabdus luminescens* belongs to the two-partner secretion family of hemolysins**. *J Bacteriol* 2002, **184**(14):3871-3878.
106. Elsen S, Huber P, Bouillot S, Couté Y, Fournier P, Dubois Y, Timsit JF, Maurin M, Attrée I: **A type III secretion negative clinical strain of *Pseudomonas aeruginosa* employs a two-partner secreted exolysin to induce hemorrhagic pneumonia**. *Cell Host Microbe* 2014, **15**(2):164-176.
107. Basso P, Ragno M, Elsen S, Reboud E, Golovkine G, Bouillot S, Huber P, Lory S, Faudry E, Attrée I: ***Pseudomonas aeruginosa* pore-forming exolysin and type IV pili cooperate to induce host cell lysis**. *mBio* 2017, **8**(1).
108. Kida Y, Higashimoto Y, Inoue H, Shimizu T, Kuwano K: **A novel secreted protease from *Pseudomonas aeruginosa* activates NF-kappaB through protease-activated receptors**. *Cell Microbiol* 2008, **10**(7):1491-1504.
109. Kida Y, Shimizu T, Kuwano K: **Cooperation between LepA and PlcH contributes to the in vivo virulence and growth of *Pseudomonas aeruginosa* in mice**. *Infect Immun* 2011, **79**(1):211-219.
110. Cope LD, Thomas SE, Latimer JL, Slaughter CA, Müller-Eberhard U, Hansen EJ: **The 100 kDa haem:haemopexin-binding protein of *Haemophilus influenzae*: structure and localization**. *Mol Microbiol* 1994, **13**(5):863-873.
111. Cope LD, Yogev R, Muller-Eberhard U, Hansen EJ: **A gene cluster involved in the utilization of both free heme and heme:hemopexin by *Haemophilus influenzae* type b**. *J Bacteriol* 1995, **177**(10):2644-2653.
112. Locht C, Bertin P, Menozzi FD, Renauld G: **The filamentous haemagglutinin, a multifaceted adhesion produced by virulent *Bordetella* spp**. *Mol Microbiol* 1993, **9**(4):653-660.
113. Anderson MS, Garcia EC, Cotter PA: **The *Burkholderia bcpAIOB* genes define unique classes of two-partner secretion and contact dependent growth inhibition systems**. *PLoS Genet* 2012, **8**(8):e1002877.
114. Serra DO, Conover MS, Arnal L, Sloan GP, Rodriguez ME, Yantorno OM, Deora R: **FHA-mediated cell-substrate and cell-cell adhesions are critical for *Bordetella pertussis* biofilm formation on abiotic surfaces and in the mouse nose and the trachea**. *PLoS One* 2011, **6**(12):e28811.
115. St Geme JW, 3rd, Yeo HJ: **A prototype two-partner secretion pathway: the *Haemophilus influenzae* HMW1 and HMW2 adhesin systems**. *Trends Microbiol* 2009, **17**(8):355-360.
116. Fleckenstein JM, Roy K, Fischer JF, Burkitt M: **Identification of a two-partner secretion locus of enterotoxigenic *Escherichia coli***. *Infect Immun* 2006, **74**(4):2245-2258.

117. Borlee BR, Goldman AD, Murakami K, Samudrala R, Wozniak DJ, Parsek MR: **Pseudomonas aeruginosa uses a cyclic-di-GMP-regulated adhesin to reinforce the biofilm extracellular matrix**. *Mol Microbiol* 2010, **75**(4):827-842.
118. Monteiro-Neto V, Bando SY, Moreira-Filho CA, Girón JA: **Characterization of an outer membrane protein associated with haemagglutination and adhesive properties of enteroaggregative *Escherichia coli* O111:H12**. *Cell Microbiol* 2003, **5**(8):533-547.
119. Relman DA, Domenighini M, Tuomanen E, Rappuoli R, Falkow S: **Filamentous hemagglutinin of Bordetella pertussis: nucleotide sequence and crucial role in adherence**. *Proc Natl Acad Sci U S A* 1989, **86**(8):2637-2641.
120. Vakevainen M, Greenberg S, Hansen E: **Inhibition of Phagocytosis by Haemophilus ducreyi Requires Expression of the LspA1 and LspA2 Proteins**. *Infection and immunity* 2003, **71**:5994-6003.
121. Gentle IE, Burri L, Lithgow T: **Molecular architecture and function of the Omp85 family of proteins**. *Molecular Microbiology* 2005, **58**(5):1216-1225.
122. Han L, Zheng J, Wang Y, Yang X, Liu Y, Sun C, Cao B, Zhou H, Ni D, Lou J *et al*: **Structure of the BAM complex and its implications for biogenesis of outer-membrane proteins**. *Nature Structural & Molecular Biology* 2016, **23**(3):192-196.
123. Soll J, Schleiff E: **Protein import into chloroplasts**. *Nature Reviews Molecular Cell Biology* 2004, **5**(3):198-208.
124. Sánchez-Pulido L, Devos D, Genevrois S, Vicente M, Valencia A: **POTRA: a conserved domain in the FtsQ family and a class of β-barrel outer membrane proteins**. *Trends in Biochemical Sciences* 2003, **28**(10):523-526.
125. Noinaj N, Kuszak AJ, Gumbart JC, Lukacik P, Chang H, Easley NC, Lithgow T, Buchanan SK: **Structural insight into the biogenesis of β-barrel membrane proteins**. *Nature* 2013, **501**(7467):385-390.
126. Bakelar J, Buchanan SK, Noinaj N: **The structure of the β-barrel assembly machinery complex**. *Science* 2016, **351**(6269):180-186.
127. Maier T, Clantin B, Gruss F, Dewitte F, Delattre A-S, Jacob-Dubuisson F, Hiller S, Villeret V: **Conserved Omp85 lid-lock structure and substrate recognition in FhaC**. *Nature Communications* 2015, **6**(1):7452.
128. Clantin B, Delattre AS, Rucktooa P, Saint N, Méli AC, Locht C, Jacob-Dubuisson F, Villeret V: **Structure of the membrane protein FhaC: a member of the Omp85-TpsB transporter superfamily**. *Science* 2007, **317**(5840):957-961.
129. Baud C, Guérin J, Petit E, Lesne E, Dupré E, Locht C, Jacob-Dubuisson F: **Translocation path of a substrate protein through its Omp85 transporter**. *Nat Commun* 2014, **5**:5271.
130. Bos MP, Robert V, Tommassen J: **Biogenesis of the gram-negative bacterial outer membrane**. *Annu Rev Microbiol* 2007, **61**:191-214.
131. Knowles TJ, Scott-Tucker A, Overduin M, Henderson IR: **Membrane protein architects: the role of the BAM complex in outer membrane protein assembly**. *Nat Rev Microbiol* 2009, **7**(3):206-214.
132. Ricci DP, Silhavy TJ: **The Bam machine: a molecular cooper**. *Biochim Biophys Acta* 2012, **1818**(4):1067-1084.
133. Hodak H, Clantin B, Willery E, Villeret V, Locht C, Jacob-Dubuisson F: **Secretion signal of the filamentous haemagglutinin, a model two-partner secretion substrate**. *Mol Microbiol* 2006, **61**(2):368-382.
134. Delattre AS, Saint N, Clantin B, Willery E, Lippens G, Locht C, Villeret V, Jacob-Dubuisson F: **Substrate recognition by the POTRA domains of TpsB transporter FhaC**. *Mol Microbiol* 2011, **81**(1):99-112.
135. Surana NK, Grass S, Hardy GG, Li H, Thanassi DG, Geme JW, 3rd: **Evidence for conservation of architecture and physical properties of Omp85-like proteins throughout evolution**. *Proc Natl Acad Sci U S A* 2004, **101**(40):14497-14502.

136. Grass S, Rempe KA, St Geme JW, 3rd: **Structural determinants of the interaction between the TpsA and TpsB proteins in the Haemophilus influenzae HMW1 two-partner secretion system**. *J Bacteriol* 2015, **197**(10):1769-1780.
137. Grass S, St. Geme III JW: **Maturation and secretion of the non-typable Haemophilus influenzae HMW1 adhesin: roles of the N-terminal and C-terminal domains**. *Molecular Microbiology* 2000, **36**(1):55-67.
138. Chevalier N, Moser M, Koch HG, Schimz KL, Willery E, Locht C, Jacob-Dubuisson F, Müller M: **Membrane targeting of a bacterial virulence factor harbouring an extended signal peptide**. *J Mol Microbiol Biotechnol* 2004, **8**(1):7-18.
139. Hodak H, Jacob-Dubuisson F: **Current challenges in autotransport and two-partner protein secretion pathways**. *Res Microbiol* 2007, **158**(8-9):631-637.
140. Mazar J, Cotter PA: **Topology and maturation of filamentous haemagglutinin suggest a new model for two-partner secretion**. *Mol Microbiol* 2006, **62**(3):641-654.
141. Guérin J, Baud C, Touati N, Saint N, Willery E, Locht C, Vezin H, Jacob-Dubuisson F: **Conformational dynamics of protein transporter FhaC: large-scale motions of plug helix**. *Mol Microbiol* 2014, **92**(6):1164-1176.
142. Jacob-Dubuisson F, Fernandez R, Coutte L: **Protein secretion through autotransporter and two-partner pathways**. *Biochim Biophys Acta* 2004, **1694**(1-3):235-257.
143. Walker G, Hertle R, Braun V: **Activation of Serratia marcescens hemolysin through a conformational change**. *Infection and immunity* 2004, **72**(1):611-614.
144. Jacob-Dubuisson F, Guérin J, Baelen S, Clantin B: **Two-partner secretion: as simple as it sounds?** *Res Microbiol* 2013, **164**(6):583-595.
145. Julio SM, Cotter PA: **Characterization of the filamentous hemagglutinin-like protein FhaS in Bordetella bronchiseptica**. *Infect Immun* 2005, **73**(8):4960-4971.
146. St Geme I, Joseph W., Grass S: **Secretion of the Haemophilus influenzae HMW1 and HMW2 adhesins involves a periplasmic intermediate and requires the HMWB and HMWC proteins**. *Molecular Microbiology* 1998, **27**(3):617-630.
147. Grass S, Lichti CF, Townsend RR, Gross J, St Geme JW, 3rd: **The Haemophilus influenzae HMW1C protein is a glycosyltransferase that transfers hexose residues to asparagine sites in the HMW1 adhesin**. *PLoS Pathog* 2010, **6**(5):e1000919.
148. Stones DH, Krachler AM: **Against the tide: the role of bacterial adhesion in host colonization**. *Biochem Soc Trans* 2016, **44**(6):1571-1580.
149. Berne C, Ducret A, Hardy GG, Brun YV: **Adhesins Involved in Attachment to Abiotic Surfaces by Gram-Negative Bacteria**. *Microbiology Spectrum* 2015, **3**(4).
150. Kumar P, Kuhlmann FM, Chakraborty S, Bourgeois AL, Foulke-Abel J, Tumala B, Vickers TJ, Sack DA, DeNearing B, Harro CD *et al*: **Enterotoxigenic *Escherichia coli*-blood group A interactions intensify diarrheal severity**. *J Clin Invest* 2018, **128**(8):3298-3311.
151. Fleckenstein JM, Roy K, Fischer JF, Burkitt M: **Identification of a two-partner secretion locus of enterotoxigenic *Escherichia coli***. *Infection and immunity* 2006, **74**(4):2245-2258.
152. Kumar P, Kuhlmann FM, Bhullar K, Yang H, Vallance BA, Xia L, Luo Q, Fleckenstein JM: **Dynamic Interactions of a Conserved Enterotoxigenic *Escherichia coli* Adhesin with Intestinal Mucins Govern Epithelium Engagement and Toxin Delivery**. *Infect Immun* 2016, **84**(12):3608-3617.
153. Farkas E, Patko D, Khanh NQ, Toth E, Vonderviszt F, Horvath R: **Self-assembly and structure of flagellin–polyelectrolyte composite layers: polyelectrolyte induced flagellar filament formation during the alternating deposition process**. *RSC Advances* 2016, **6**(95):92159-92167.
154. Chevance FF, Hughes KT: **Coordinating assembly of a bacterial macromolecular machine**. *Nat Rev Microbiol* 2008, **6**(6):455-465.
155. Prüss BM, Besemann C, Denton A, Wolfe AJ: **A complex transcription network controls the early stages of biofilm development by *Escherichia coli***. *J Bacteriol* 2006, **188**(11):3731-3739.

156. Duan Q, Zhou M, Zhu L, Zhu G: **Flagella and bacterial pathogenicity**. *J Basic Microbiol* 2013, **53**(1):1-8.
157. O'Neil HS, Marquis H: **Listeria monocytogenes Flagella Are Used for Motility, Not as Adhesins, To Increase Host Cell Invasion**. *Infection and Immunity* 2006, **74**(12):6675-6681.
158. Ormonde P, Hörstedt P, O'Toole R, Milton DL: **Role of Motility in Adherence to and Invasion of a Fish Cell Line by Vibrio anguillarum**. *J Bacteriol* 2000, **182**(8):2326-2328.
159. Eaton KA, Suerbaum S, Josenhans C, Krakowka S: **Colonization of gnotobiotic piglets by Helicobacter pylori deficient in two flagellin genes**. *Infection and Immunity* 1996, **64**(7):2445-2448.
160. Haiko J, Westerlund-Wikström B: **The role of the bacterial flagellum in adhesion and virulence**. *Biology (Basel)* 2013, **2**(4):1242-1267.
161. Wang G, Geisbrecht BV, Rueter C, Hardwidge PR: **Enterotoxigenic *Escherichia coli* Flagellin Inhibits TNF-Induced NF-κB Activation in Intestinal Epithelial Cells**. *Pathogens* 2017, **6**(2).
162. Zhou M, Duan Q, Zhu X, Guo Z, Li Y, Hardwidge PR, Zhu G: **Both flagella and F4 fimbriae from F4ac+ enterotoxigenic *Escherichia coli* contribute to attachment to IPEC-J2 cells in vitro**. *Vet Res* 2013, **44**(1):30.
163. Zhou M, Guo Z, Yang Y, Duan Q, Zhang Q, Yao F, Zhu J, Zhang X, Hardwidge P, Zhu G: **Flagellin and F4 fimbriae have opposite effects on biofilm formation and quorum sensing in F4ac+ enterotoxigenic *Escherichia coli***. *Veterinary microbiology* 2013, **168**.
164. Garibyan L, Avashia N: **Polymerase chain reaction**. *J Invest Dermatol* 2013, **133**(3):1-4.
165. Delidow BC, Lynch JP, Peluso JJ, White BA: **Polymerase chain reaction : basic protocols**. *Methods Mol Biol* 1993, **15**:1-29.
166. Motohashi K: **Development of highly sensitive and low-cost DNA agarose gel electrophoresis detection systems, and evaluation of non-mutagenic and loading dye-type DNA-staining reagents**. *PLoS One* 2019, **14**(9):e0222209.
167. Buckhout-White S, Person C, Medintz IL, Goldman ER: **Restriction Enzymes as a Target for DNA-Based Sensing and Structural Rearrangement**. *ACS Omega* 2018, **3**(1):495-502.
168. Takita E, Kohda K, Tomatsu H, Hanano S, Moriya K, Hosouchi T, Sakurai N, Suzuki H, Shinmyo A, Shibata D: **Precise sequential DNA ligation on a solid substrate: solid-based rapid sequential ligation of multiple DNA molecules**. *DNA Res* 2013, **20**(6):583-592.
169. Froger A, Hall JE: **Transformation of plasmid DNA into E. coli using the heat shock method**. *J Vis Exp* 2007(6):253.
170. Ren J, Karna S, Lee HM, Yoo SM, Na D: **Artificial transformation methodologies for improving the efficiency of plasmid DNA transformation and simplifying its use**. *Appl Microbiol Biotechnol* 2019, **103**(23-24):9205-9215.
171. Bachman J: **Site-directed mutagenesis**. *Methods Enzymol* 2013, **529**:241-248.
172. Heather JM, Chain B: **The sequence of sequencers: The history of sequencing DNA**. *Genomics* 2016, **107**(1):1-8.
173. Sanger F, Nicklen S, Coulson AR: **DNA sequencing with chain-terminating inhibitors**. *Proceedings of the national academy of sciences* 1977, **74**(12):5463-5467.
174. Palmer I, Wingfield PT: **Preparation and extraction of insoluble (inclusion-body) proteins from *Escherichia coli***. *Curr Protoc Protein Sci* 2004, **Chapter 6**:6.3.1-6.3.18.
175. Rudolph R, Lilie H: **In vitro folding of inclusion body proteins**. *The FASEB journal* 1996, **10**(1):49-56.
176. Burgess RR: **Refolding solubilized inclusion body proteins**. *Methods in enzymology* 2009, **463**:259-282.
177. Clark EDB: **Protein refolding for industrial processes**. *Current opinion in biotechnology* 2001, **12**(2):202-207.
178. Lu S-C, Lin S-C: **Recovery of active N-acetyl-D-glucosamine 2-epimerase from inclusion bodies by solubilization with non-denaturing buffers**. *Enzyme and microbial technology* 2012, **50**(1):65-70.

179. Singh A, Upadhyay V, Upadhyay AK, Singh SM, Panda AK: **Protein recovery from inclusion bodies of *Escherichia coli* using mild solubilization process**. *Microbial Cell Factories* 2015, **14**(1):41.
180. Upadhyay AK, Murmu A, Singh A, Panda AK: **Kinetics of inclusion body formation and its correlation with the characteristics of protein aggregates in *Escherichia coli***. *PloS one* 2012, **7**(3):e33951.
181. Fischer B, Perry B, Sumner I, Goodenough P: **A novel sequential procedure to enhance the renaturation of recombinant protein from Escherichia coil inclusion bodies**. *Protein Engineering, Design and Selection* 1992, **5**(6):593-596.
182. Singh SM, Panda AK: **Solubilization and refolding of bacterial inclusion body proteins**. *Journal of bioscience and bioengineering* 2005, **99**(4):303-310.
183. Khan R, Rao KA, Eshwari A, Totey S, Panda A: **Solubilization of recombinant ovine growth hormone with retention of native-like secondary structure and its refolding from the inclusion bodies of *Escherichia coli***. *Biotechnology progress* 1998, **14**(5):722-728.
184. Kabsch W: **XDS**. *Acta Crystallogr D Biol Crystallogr* 2010, **66**(Pt 2):125-132.
185. Liebschner D, Afonine PV, Baker ML, Bunkoczi G, Chen VB, Croll TI, Hintze B, Hung L-W, Jain S, McCoy AJ *et al*: **Macromolecular structure determination using X-rays, neutrons and electrons: recent developments in Phenix**. *Acta Crystallographica Section D* 2019, **75**(10):861-877.
186. Emsley P, Lohkamp B, Scott WG, Cowtan K: **Features and development of Coot**. *Acta Crystallographica Section D* 2010, **66**(4):486-501.
187. Delisse-Gathoye AM, Locht C, Jacob F, Raaschou-Nielsen M, Heron I, Ruelle JL, de Wilde M, Cabezon T: **Cloning, partial sequence, expression, and antigenic analysis of the filamentous hemagglutinin gene of Bordetella pertussis**. *Infect Immun* 1990, **58**(9):2895-2905.
188. Chen L, Yang J, Yu J, Yao Z, Sun L, Shen Y, Jin Q: **VFDB: a reference database for bacterial virulence factors**. *Nucleic Acids Res* 2005, **33**(Database issue):D325-328.
189. Barenkamp SJ, Leininger E: **Cloning, expression, and DNA sequence analysis of genes encoding nontypeable Haemophilus influenzae high-molecular-weight surface-exposed proteins related to filamentous hemagglutinin of Bordetella pertussis**. *Infect Immun* 1992, **60**(4):1302-1313.
190. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A *et al*: **Highly accurate protein structure prediction with AlphaFold**. *Nature* 2021, **596**(7873):583-589.
191. Wiedemann C, Bellstedt P, Görlach M: **CAPITO--a web server-based analysis and plotting tool for circular dichroism data**. *Bioinformatics* 2013, **29**(14):1750-1757.
192. Perez-Iratxeta C, Andrade-Navarro MA: **K2D2: estimation of protein secondary structure from circular dichroism spectra**. *BMC Struct Biol* 2008, **8**:25.
193. Jacob-Dubuisson F, Buisine C, Willery E, Renauld-Mongénie G, Locht C: **Lack of functional complementation between Bordetella pertussis filamentous hemagglutinin and Proteus mirabilis HpmA hemolysin secretion machineries**. *J Bacteriol* 1997, **179**(3):775-783.
194. Mariani V, Biasini M, Barbato A, Schwede T: **lDDT: a local superposition-free score for comparing protein structures and models using distance difference tests**. *Bioinformatics* 2013, **29**(21):2722-2728.
195. Pang YT, Hazel AJ, Gumbart JC: **Uncovering the folding mechanism of pertactin: A comparative study of isolated and vectorial folding**. *Biophysical Journal* 2023.
196. Xiao Z, Bergeron H, Grosse S, Beauchemin M, Garron ML, Shaya D, Sulea T, Cygler M, Lau PC: **Improvement of the thermostability and activity of a pectate lyase by single amino acid substitutions, using a strategy based on melting-temperature-guided sequence alignment**. *Appl Environ Microbiol* 2008, **74**(4):1183-1189.

197. Buth SA, Menin L, Shneider MM, Engel J, Boudko SP, Leiman PG: **Structure and Biophysical Properties of a Triple-Stranded Beta-Helix Comprising the Central Spike of Bacteriophage T4**. *Viruses* 2015, **7**(8):4676-4706.
198. Ruhe ZC, Townsley L, Wallace AB, King A, Van der Woude MW, Low DA, Yildiz FH, Hayes CS: **CdiA promotes receptor-independent intercellular adhesion**. *Mol Microbiol* 2015, **98**(1):175-192.
199. Schönherr R, Tsolis R, Focareta T, Braun V: **Amino acid replacements in the Serratia marcescens haemolysin ShlA define sites involved in activation and secretion**. *Mol Microbiol* 1993, **9**(6):1229-1237.
200. Baek M, DiMaio F, Anishchenko I, Dauparas J, Ovchinnikov S, Lee GR, Wang J, Cong Q, Kinch LN, Schaeffer RD *et al*: **Accurate prediction of protein structures and interactions using a three-track neural network**. *Science* 2021, **373**(6557):871-876.
201. Menozzi FD, Boucher PE, Riveau G, Gantiez C, Locht C: **Surface-associated filamentous hemagglutinin induces autoagglutination of Bordetella pertussis**. *Infect Immun* 1994, **62**(10):4261-4269.
202. Colombi D, Horton DSPQ, Oliveira MLS, Sakauchi MA, Ho PL: **Antibodies produced against a fragment of filamentous haemagglutinin (FHA) of Bordetella pertussis are able to inhibit hemagglutination induced by the whole adhesin**. *FEMS Microbiology Letters* 2004, **240**(1):41-47.
203. Andrade MA, Perez-Iratxeta C, Ponting CP: **Protein repeats: structures, functions, and evolution**. *J Struct Biol* 2001, **134**(2-3):117-131.
204. Görner K, Holtorf E, Waak J, Pham T-T, Vogt-Weisenhorn DM, Wurst W, Haass C, Kahle PJ: **Structural Determinants of the C-terminal Helix-Kink-Helix Motif Essential for Protein Stability and Survival Promoting Activity of DJ-1***. *Journal of Biological Chemistry* 2007, **282**(18):13680-13691.
205. Heras B, Totsika M, Peters KM, Paxman JJ, Gee CL, Jarrott RJ, Perugini MA, Whitten AE, Schembri MA: **The antigen 43 structure reveals a molecular Velcro-like mechanism of autotransporter-mediated bacterial clumping**. *Proc Natl Acad Sci U S A* 2014, **111**(1):457-462.
206. Wilman HR, Shi J, Deane CM: **Helix kinks are equally prevalent in soluble and membrane proteins**. *Proteins* 2014, **82**(9):1960-1970.
207. Kobe B, Kajava AV: **The leucine-rich repeat as a protein recognition motif**. *Curr Opin Struct Biol* 2001, **11**(6):725-732.
208. Maldonado-Contreras A, Birtley JR, Boll E, Zhao Y, Mumy KL, Toscano J, Ayehunie S, Reinecker HC, Stern LJ, McCormick BA: **Shigella depends on SepA to destabilize the intestinal epithelial integrity via cofilin activation**. *Gut Microbes* 2017, **8**(6):544-560.
209. Barbirz S, Müller JJ, Uetrecht C, Clark AJ, Heinemann U, Seckler R: **Crystal structure of *Escherichia coli* phage HK620 tailspike: podoviral tailspike endoglycosidase modules are evolutionarily related**. *Mol Microbiol* 2008, **69**(2):303-316.
210. Muñoz PA, Márquez SL, González-Nilo FD, Márquez-Miranda V, Blamey JM: **Structure and application of antifreeze proteins from Antarctic bacteria**. *Microbial Cell Factories* 2017, **16**(1):138.
211. Alsteens D, Martinez N, Jamin M, Jacob-Dubuisson F: **Sequential Unfolding of Beta Helical Protein by Single-Molecule Atomic Force Microscopy**. *PLOS ONE* 2013, **8**(8):e73572.
212. Deryusheva EI, Machulin AV, Galzitskaya OV: **Structural, Functional, and Evolutionary Characteristics of Proteins with Repeats**. *Molecular Biology* 2021, **55**(5):683-704.
213. Finn RD, Mistry J, Tate J, Coggill P, Heger A, Pollington JE, Gavin OL, Gunasekaran P, Ceric G, Forslund K *et al*: **The Pfam protein families database**. *Nucleic Acids Res* 2010, **38**(Database issue):D211-222.
214. Marcotte EM, Pellegrini M, Yeates TO, Eisenberg D: **A census of protein repeats**. *J Mol Biol* 1999, **293**(1):151-160.

215. Clarke S, Harris L, Richards R, Foster S: **Analysis of Ebh, a 1.1-Megadalton Cell Wall-Associated Fibronectin-Binding Protein of Staphylococcus aureus**. *Infection and immunity* 2003, **70**:6680-6687.
216. Gravekamp C, Horensky DS, Michel JL, Madoff LC: **Variation in repeat number within the alpha C protein of group B streptococci alters antigenicity and protective epitopes**. *Infection and immunity* 1996, **64**(9):3576-3583.
217. Xie Y, Karki CB, Du D, Li H, Wang J, Sobitan A, Teng S, Tang Q, Li L: **Spike Proteins of SARS-CoV and SARS-CoV-2 Utilize Different Mechanisms to Bind With Human ACE2**. *Frontiers in Molecular Biosciences* 2020, **7**(392).
218. Hagan MF, Zandi R: **Recent advances in coarse-grained modeling of virus assembly**. *Curr Opin Virol* 2016, **18**:36-43.

# 8 APPENDICES

## 8.1 Gene sequences

### 8.1.1 EtpA full-length nucleotide sequence

GenBank: AY920525.2

**Appendix 8-1: EtpA full-length nucleotide sequence**

ATGAACCGTATATATAAACTGAAGTTTGACAAACGCCGCAACGAACTGGTGGTGGTGAGTGAAATCACCACCGGCGTGGGTAATGC
AAAAGCCACGGGCAGCGTGGAGGGCGAAAAGTCCCCCCGTCGTGGCGTGCGCGCCATGGCGCTGAGCCTGCTGTCGGGTATGATG
ATAATGGCCCATCCGGCGATGTCAGCAAACCTGCCGACCGGTGGCCAGATTGTGGCAGGTTCAGGCAGTATCCAGACGCCTTCCGG
CAACCAGATGAATATTCATCAGAACAGCCAGAACATGGTGGCCAACTGGAACAGCTTTGACATTGGTAAAGGAAATACGGTGCAGT
TTGACCAGCCCAGCAGCAGTGCGGTGGCGCTGAACCGTGTTGTGGGTGGCGGTGAATCGCAGATTATGGGTAACCTGAAGGCGAA
TGGTCAGGTGTTCCTGGTTAACCCGAACGGCGTGCTGTTTGGTGAGGGGGCCAGTGTCAGCACGTCAGGTTTTGTGGCATCGACCC
GCGACATTAAAAACGACGACTTCATGAACCGTCGTTACACCTTCAGCGGCGGACAGAAAGCCGGGGCAGCGATTGTGAACCAGGG
GGAACTGACCACAAATGCCGGTGGCTATATTGTGCTGGCAGCAGACAGGGTCAGCAACAGTGGCACCATCCGTACGCCGGGCGGC
AAGACCGTCCTGGCGGCCAGCGAGCGCATCACGCTGCAGCTGGATAATGGTGGCCTGATGTCCGTGCAGGTGACAGGAGATGTGG
TTAATGCCCTGGTGGAAAACCGCGGTCTGGTCAGTGCCCGGGATGGTCAGGTGTACCTGACCGCACTTGGCCGGGGTATGCTGATG
AACACGGTACTGAACGTGAGCGGGGTGGTGGAAGCCAGCGGTATGCACCGTCAGGACGGTAACATTGTACTGGACGGTGGCGACA
GTGGTGTGGTGCACCTGAGTGGTACCCTGCAGGCGGACAATGCGTCCGGTCAGGGTGGTAAGGTTGTCGTGCAGGGTAAGAATAT
TCTGCTGGACAAGGGCAGCAACATCACAGCAACCGGTGGTCAGGGCGGCGGTGAAGTGTATGTCGGTGGCGGCTGGCAGGGTAA
GGACAGCAACATCCGTAATGCGGACAAGGTGGTGATGCAGGGCGGCGCCCGCATTGACGTTTCTGCAACGCAGCAGGGTAACGGC
GGTACGGCTGTGCTGTGGTCAGACAGCTACACCAACTTCCATGGTCAGATTAGCGCGAAGGGCGGTGAGACCGGCGGTAACGGTG
GTCGGGTGGAGACCTCTTCGCACGGTAACCTGCAGGCATTTGGTACGGTCAGTGCATCCGCGAAGAAAGGCAAGGCGGGTAACTG
GCTGCTGGACTCGGCGGATATCACCATTGTGAATGGTAGCAATGTTAGCAAAACTGAGACGACTCAATCGCCGCCGCACACGCAATT
TGCACCCACCGCTGCGGGCTCTGCGGTCAGCAATACCAGTATCAACAACAGGCTGAACAACGGGACCAGTGTCACTATTCTGACCCA
TCGCACAAGAACAGGCACAGCTCAGGGCGGGAATATTACCGTTAATGCGGCAATTAACAAAAGCAACGGAAGTGATGTCAACCTGA
CGCTGCAGGCTGGCGGCAACATCACGGTAAACAACAGCATCACGTCCACCGAGGGTAAGCTGAATGTTAATCTGTCGGGCGCCAGG
ACCAGCAATGGCAGTATCACCATTAGCAATAACGCCAATATAACGACCAATGGTGGGGATATAACTGTTGGGACGACAAATACTTCA
AACCGTGTGAATATATCTATTAATAACACTACCCTGAATGCGTCAAATGGCAACATCCAGTTGACCGGGACCGGGACCGATAGCGGG
ATTCTGTTTGCTGGCAACAACAGGCTGACGGCCAGTAACATTGCTCTTACCGGGAACAGTACGAGTGGGAATGCCATCAACCTTACA
GGCACTGCCACGCTGAATGCCACGAATAACATTACTCTTACCGGGAGCAGTACGAGTGGGAATGCCATCAACCTTAAAGGCAACAA
CACGCTGACGGCCAGTAACATTACTCTTACCGGGGAAAGTACGAGTGGGAATGCCATCAACCTTACAGACACTACAGGCACTACCAC
GCTGAATGCCACGAATAACATCACTATGCAGGGGACCCGTGTTCAGATTAAACACTCCAACATCACCGCGGGCAACTTTGCGCTGAA
TGCGACAGTGGCCGGCTCTGAAATCAGCAATACCACGCTGACGGCCACCAACAACATCAACCTGGCGGCTAAGACGAACAGTGCGA
GCTCTGGTGTTTACCTGAAAGATGCAAGAATTACATCCACCAATGGCAGTATCACGGCTAACGGTACTGCCACAGCAAACGGCAAGG
CCACGCATCTGGACGGCAACGTCACCCTGAATGCGTCAAATGGCAGAATCAAGTTGACCGGGAACGGGCACGGTAGCGCCTCCGG
GATTCTGTTTGCTGGCAACAACAGGCTGACGGCCAGTAACATTGCTCTTACCGGGAACAGTACGAGTGGGAATGCCATCAACCTTAC
AGGCACTGCCACGCTGAATGCCACGAATGACATTACTCTTACCGGGAGCAGTACGAGTGGGAATGCCATCAACCTTACAGGCACTG
CCACGCTGAATGCCACGAATAACATTACTCTTACCGGGAGCAGTACGAGTGGGAATGCCATCAACCTTAAAGGCAACAACACGCTGA
CGGCCAGTAACATTACTCTTACCGGGGAAAGTACGAGTGGGAATGCCATCAACCTTACAGACACTACAGGCACTACCACGCTGAATG
CCACGAATAACATCACTATGCAGGGGACCCGTGTTCAGATTAAACACTCCAACATCACCGCGGGCAACTTTGCGCTGAATGCGACAG
TGGCCGGCTCTGAAATCAGCAATACCACGCTGACGGCCACCAACAACATCAACCTGGCGGCTAAGACGAACAGTGCGAGCTCTGGT
GTTTACCTGAAAGATGCAAGAATTACATCCACCAATGGCAGTATCACGGCTAACGGTACTGCCACAGCAAACGGCAAGGCCACGCAT
CTGGACGGCAACGTCACCCTGAATGCGTCAAATGGCAGAATCAAGTTGACCGGGAACGGGCACGGTAGCGCCTCCGGGATTCTGTT
TGCTGGCAACAACAGGCTGACGGCCAGTAACATTGCTCTTACCGGGAACAGTACGAGTGGGAATGCCATCAACCTTACAGGCACTG
CCACGCTGAATGCCACGAATGACATTACTCTTACCGGGAGCAGTACGAGTGGGAATGCCATCAACCTTACAGGCACTGCCACGCTGA
ATGCCACGAATAACATTACTCTTACCGGGAGCAGTACGAGTGGGAATGCCATCAACCTTAAAGGCAACAACACGCTGACGGCCAGT
AACATTACTCTTACCGGGGAAAGTACGAGTGGGAATGCCATCAACCTTACAGACACTACAGGCACTACCACGCTGAATGCCACGAAT
AACATCACTATGCAGGGGACCCGTGTTCAGATTAAACACTCCAACATCACCGCGGGCAACTTTGCGCTGAATGCGACAGTGGCCGGC
TCTGAAATCAGCAATACCACGCTGACGGCCACCAACAACATCAACCTGGCGGCTAAGACGAACAGTGCGAGCTCTGGTGTTTACCTG
AAAGATGCAAGAATTACATCCACCAATGGCAGTATCACGGCTAACGGTACTGCCACAGCAAACGGCAAGGCCACGCATCTGGACGG
CAACGTCACCCTGAATGCGTCAAATGGCAGAATCAAGTTGACCGGGAACGGGCACGGTAGCGCCTCCGGGATTCTGTTTGCTGGCA
ACAACAGGCTGACGGCCAGTAACATTGCTCTTACCGGGAACAGTACGAGTGGGAATGCCATCAACCTTACAGGCACTGCCACGCTG
AATGCCACGAATGACATTACTCTTACCGGGAGCAGTACGAGTGGGAATGCCATCAACCTTACAGGCACTGCCACGCTGAATGCCACG
AATAACATTACTCTTACCGGGAGCAGTACGAGTGGGAATGCCATCAACCTTAAAGGCAACAACACGCTGACGGCCAGTAACATTACT
CTTACCGGGGAAAGTACGAGTGGGAATGCCATCAACCTTACAGACACTACAGGCACTACCACGCTGAATGCCACGAATAACATCACT
ATGCAGGGGACCCGTGTTCAGATTAAACACTCCAACATCACCGCGGGCAACTTTGCGCTGAATGCGACAGTGGCCGGCTCTGAAATC
AGCAATACCACGCTGACGGCCACCAACAACATCAACCTGGCGGCTAAGACGAACAGTGCGAGCTCTGGTGTTTACCTGAAAGATGC

AAGAATTACATCCACTAATGGCAGTATCACGACTAACGGTACTGCCACAGCAAACGGCAAGGCCACGCATCTGGACGGCAACGTCA
CCCTGAATGCGTCAAATGGCAGAATCAAGTTGACCGGGAACGGGCACGGTAGCGCCTCCGGGATTCTGTTTGCTGGCAACAACAGG
CTGACGGCCAGTAACATTGCTCTTACCGGGAACAGTACGAGTGGGAATGCCATCAACCTTACAGGCACTGCCACGCTGAATGCCACG
AATGACATTACTCTTACCGGGAGCAGTACGAGTGGGAATGCCATCAACCTTACAGGCACTGCCACGCTGAATGCCACGAATAACATT
ACTCTTACCGGGAGCAGTACGAGTGGGAATGCCATCAACCTTAAAGGCAACAACACGCTGACGGCCAGTAACATTACTCTTACCGGG
GAAAGTACGAGTGGGAATGCCATCAACCTTACAGACACTACAGGCACTACCACGCTGAATGCCACGAATAACATCACTATGCAGGG
GACCCGTGTTCAGATTAAACACTCCAACATCACCGCGGGCAACTTTGCGCTGAATGCGACAGTGGCCGGCTCTGAAATCAGCAATAC
CACGCTGACGGCCACCAACAACATCAACCTGGCGGCTAAGACGAACAGTGCGAGCTCTGGTGTTTACCTGAAAGATGCAAGAATTA
CATCCACCAATGGCAGTATCACGGCTAACGGTACTGCCCCAGCAAACGACAATGCCACGTATCTGGACGGCAACGTCACCCTGAATG
CGTCAAATGGCAGCATCAAGTTGACCGGGAACGGGAACGGTAGCACCTCCGGGATTCTGTTTGCTGGCAACAACACGCTGACGGCC
AGTAACATTACTCTTACCGGGAACAGTGAGGTGTACTGGCAATAG

## 8.1.2 EtpA$^{67-447}$ nucleotide sequence

Obtained from EtpA full-length nucleotide sequence.

**Appendix 8-2: EtpA$^{67-447}$ nucleotide sequence**

AACCTGCCGACCGGTGGCCAGATTGTGGCAGGTTCAGGCAGTATCCAGACGCCTTCCGGCAACCAGATGAATATTCATCAGAACAG
CCAGAACATGGTGGCCAACTGGAACAGCTTTGACATTGGTAAAGGAAATACGGTGCAGTTTGACCAGCCCAGCAGCAGTGCGGTGG
CGCTGAACCGTGTTGTGGGTGGCGGTGAATCGCAGATTATGGGTAACCTGAAGGCGAATGGTCAGGTGTTCCTGGTTAACCCGAAC
GGCGTGCTGTTTGGTGAGGGGGCCAGTGTCAGCACGTCAGGTTTTGTGGCATCGACCCGCGACATTAAAAACGACGACTTCATGAA
CCGTCGTTACACCTTCAGCGGCGGACAGAAAGCCGGGGCAGCGATTGTGAACCAGGGGGAACTGACCACAAATGCCGGTGGCTAT
ATTGTGCTGGCAGCAGACAGGGTCAGCAACAGTGGCACCATCCGTACGCCGGGCGGCAAGACCGTCCTGGCGGCCAGCGAGCGCA
TCACGCTGCAGCTGGATAATGGTGGCCTGATGTCCGTGCAGGTGACAGGAGATGTGGTTAATGCCCTGGTGGAAAACCGCGGTCTG
GTCAGTGCCCGGGATGGTCAGGTGTACCTGACCGCACTTGGCCGGGGTATGCTGATGAACACGGTACTGAACGTGAGCGGGGTGG
TGGAAGCCAGCGGTATGCACCGTCAGGACGGTAACATTGTACTGGACGGTGGCGACAGTGGTGTGGTGCACCTGAGTGGTACCCT
GCAGGCGGACAATGCGTCCGGTCAGGGTGGTAAGGTTGTCGTGCAGGGTAAGAATATTCTGCTGGACAAGGGCAGCAACATCACA
GCAACCGGTGGTCAGGGCGGCGGTGAAGTGTATGTCGGTGGCGGCTGGCAGGGTAAGGACAGCAACATCCGTAATGCGGACAAG
GTGGTGATGCAGGGCGGCGCCCGCATTGACGTTTCTGCAACGCAGCAGGGTAACGGCGGTACGGCTGTGCTGTGGTCAGACAGCT
ACACCAACTTCCATGGTCAGATTAGCGCGAAGGGCGGTGAGACCGGCGGTAACGGTGGTCGGGTGGAGACCTCTTCGCACGGTAA
CCTGCAGGCATTTGGTACGGTCAGTGCATCCGCG

## 8.1.3 EtpA$^{1-606}$ nucleotide sequence

Obtained from EtpA full-length nucleotide sequence.

**Appendix 8-3: EtpA$^{1-606}$ nucleotide sequence**

ATGAACCGTATATATAAACTGAAGTTTGACAAACGCCGCAACGAACTGGTGGTGGTGAGTGAAATCACCACCGGCGTGGGTAATGC
AAAAGCCACGGGCAGCGTGGAGGGCGAAAAGTCCCCCCGTCGTGGCGTGCGCGCCATGGCGCTGAGCCTGCTGTCGGGTATGATG
ATAATGGCCCATCCGGCGATGTCAGCAAACCTGCCGACCGGTGGCCAGATTGTGGCAGGTTCAGGCAGTATCCAGACGCCTTCCGG
CAACCAGATGAATATTCATCAGAACAGCCAGAACATGGTGGCCAACTGGAACAGCTTTGACATTGGTAAAGGAAATACGGTGCAGT
TTGACCAGCCCAGCAGCAGTGCGGTGGCGCTGAACCGTGTTGTGGGTGGCGGTGAATCGCAGATTATGGGTAACCTGAAGGCGAA
TGGTCAGGTGTTCCTGGTTAACCCGAACGGCGTGCTGTTTGGTGAGGGGGCCAGTGTCAGCACGTCAGGTTTTGTGGCATCGACCC
GCGACATTAAAAACGACGACTTCATGAACCGTCGTTACACCTTCAGCGGCGGACAGAAAGCCGGGGCAGCGATTGTGAACCAGGG
GGAACTGACCACAAATGCCGGTGGCTATATTGTGCTGGCAGCAGACAGGGTCAGCAACAGTGGCACCATCCGTACGCCGGGCGGC
AAGACCGTCCTGGCGGCCAGCGAGCGCATCACGCTGCAGCTGGATAATGGTGGCCTGATGTCCGTGCAGGTGACAGGAGATGTGG
TTAATGCCCTGGTGGAAAACCGCGGTCTGGTCAGTGCCCGGGATGGTCAGGTGTACCTGACCGCACTTGGCCGGGGTATGCTGATG
AACACGGTACTGAACGTGAGCGGGGTGGTGGAAGCCAGCGGTATGCACCGTCAGGACGGTAACATTGTACTGGACGGTGGCGACA
GTGGTGTGGTGCACCTGAGTGGTACCCTGCAGGCGGACAATGCGTCCGGTCAGGGTGGTAAGGTTGTCGTGCAGGGTAAGAATAT
TCTGCTGGACAAGGGCAGCAACATCACAGCAACCGGTGGTCAGGGCGGCGGTGAAGTGTATGTCGGTGGCGGCTGGCAGGGTAA
GGACAGCAACATCCGTAATGCGGACAAGGTGGTGATGCAGGGCGGCGCCCGCATTGACGTTTCTGCAACGCAGCAGGGTAACGGC
GGTACGGCTGTGCTGTGGTCAGACAGCTACACCAACTTCCATGGTCAGATTAGCGCGAAGGGCGGTGAGACCGGCGGTAACGGTG
GTCGGGTGGAGACCTCTTCGCACGGTAACCTGCAGGCATTTGGTACGGTCAGTGCATCCGCGAAGAAAGGCAAGGCGGGTAACTG
GCTGCTGGACTCGGCGGATATCACCATTGTGAATGGTAGCAATGTTAGCAAAACTGAGACGACTCAATCGCCGCCGCACACGCAATT
TGCACCCACCGCTGCGGGCTCTGCGGTCAGCAATACCAGTATCAACAACAGGCTGAACAACGGGACCAGTGTCACTATTCTGACCCA
TCGCACAAGAACAGGCACAGCTCAGGGCGGGAATATTACCGTTAATGCGGCAATTAACAAAAGCAACGGAAGTGATGTCAACCTGA
CGCTGCAGGCTGGCGGCAACATCACGGTAAACAACAGCATCACGTCCACCGAGGGTAAGCTGAATGTTAATCTGTCGGGCGCCAGG

ACCAGCAATGGCAGTATCACCATTAGCAATAACGCCAATATAACGACCAATGGTGGGGATATAACTGTTGGGACGACAAATACTTCA
AACCGTGTGAATATATCTATT

## 8.1.4 EtpA$^{448-930}$ nucleotide sequence

Obtained from EtpA full-length nucleotide sequence.

**Appendix 8-4: EtpA$^{448-930}$ nucleotide sequence**

AAGAAAGGCAAGGCGGGTAACTGGCTGCTGGACTCGGCGGATATCACCATTGTGAATGGTAGCAATGTTAGCAAAACTGAGACGA
CTCAATCGCCGCCGCACACGCAATTTGCACCCACCGCTGCGGGCTCTGCGGTCAGCAATACCAGTATCAACAACAGGCTGAACAACG
GGACCAGTGTCACTATTCTGACCCATCGCACAAGAACAGGCACAGCTCAGGGCGGGAATATTACCGTTAATGCGGCAATTAACAAA
AGCAACGGAAGTGATGTCAACCTGACGCTGCAGGCTGGCGGCAACATCACGGTAAACAACAGCATCACGTCCACCGAGGGTAAGCT
GAATGTTAATCTGTCGGGCGCCAGGACCAGCAATGGCAGTATCACCATTAGCAATAACGCCAATATAACGACCAATGGTGGGGATA
TAACTGTTGGGACGACAAATACTTCAAACCGTGTGAATATATCTATTAATAACACTACCCTGAATGCGTCAAATGGCAACATCCAGTT
GACCGGGACCGGGACCGATAGCGGGATTCTGTTTGCTGGCAACAACAGGCTGACGGCCAGTAACATTGCTCTTACCGGGAACAGTA
CGAGTGGGAATGCCATCAACCTTACAGGCACTGCCACGCTGAATGCCACGAATAACATTACTCTTACCGGGAGCAGTACGAGTGGG
AATGCCATCAACCTTAAAGGCAACAACACGCTGACGGCCAGTAACATTACTCTTACCGGGGAAAGTACGAGTGGGAATGCCATCAA
CCTTACAGACACTACAGGCACTACCACGCTGAATGCCACGAATAACATCACTATGCAGGGGACCCGTGTTCAGATTAAACACTCCAA
CATCACCGCGGGCAACTTTGCGCTGAATGCGACAGTGGCCGGCTCTGAAATCAGCAATACCACGCTGACGGCCACCAACAACATCAA
CCTGGCGGCTAAGACGAACAGTGCGAGCTCTGGTGTTTACCTGAAAGATGCAAGAATTACATCCACCAATGGCAGTATCACGGCTA
ACGGTACTGCCACAGCAAACGGCAAGGCCACGCATCTGGACGGCAACGTCACCCTGAATGCGTCAAATGGCAGAATCAAGTTGACC
GGGAACGGGCACGGTAGCGCCTCCGGGATTCTGTTTGCTGGCAACAACAGGCTGACGGCCAGTAACATTGCTCTTACCGGGAACAG
TACGAGTGGGAATGCCATCAACCTTACAGGCACTGCCACGCTGAATGCCACGAATGACATTACTCTTACCGGGAGCAGTACGAGTG
GGAATGCCATCAACCTTACAGGCACTGCCACGCTGAATGCCACGAATAACATTACTCTTACCGGGAGCAGTACGAGTGGGAATGCCA
TCAACCTTAAAGGCAACAACACGCTGACGGCCAGTAACATTACTCTTACCGGGGAAAGTACGAGTGGG

## 8.1.5 EtpA$^{67-930}$ nucleotide sequence

Obtained from EtpA full-length nucleotide sequence.

**Appendix 8-5: EtpA$^{67-930}$ nucleotide sequence**

AACCTGCCGACCGGTGGCCAGATTGTGGCAGGTTCAGGCAGTATCCAGACGCCTTCCGGCAACCAGATGAATATTCATCAGAACAG
CCAGAACATGGTGGCCAACTGGAACAGCTTTGACATTGGTAAAGGAAATACGGTGCAGTTTGACCAGCCCAGCAGCAGTGCGGTGG
CGCTGAACCGTGTTGTGGGTGGCGGTGAATCGCAGATTATGGGTAACCTGAAGGCGAATGGTCAGGTGTTCCTGGTTAACCCGAAC
GGCGTGCTGTTTGGTGAGGGGGCCAGTGTCAGCACGTCAGGTTTTGTGGCATCGACCCGCGACATTAAAAACGACGACTTCATGAA
CCGTCGTTACACCTTCAGCGGCGGACAGAAAGCCGGGGCAGCGATTGTGAACCAGGGGGAACTGACCACAAATGCCGGTGGCTAT
ATTGTGCTGGCAGCAGACAGGGTCAGCAACAGTGGCACCATCCGTACGCCGGGCGGCAAGACCGTCCTGGCGGCCAGCGAGCGCA
TCACGCTGCAGCTGGATAATGGTGGCCTGATGTCCGTGCAGGTGACAGGAGATGTGGTTAATGCCCTGGTGGAAAACCGCGGTCTG
GTCAGTGCCCGGGATGGTCAGGTGTACCTGACCGCACTTGGCCGGGGTATGCTGATGAACACGGTACTGAACGTGAGCGGGGTGG
TGGAAGCCAGCGGTATGCACCGTCAGGACGGTAACATTGTACTGGACGGTGGCGACAGTGGTGTGGTGCACCTGAGTGGTACCCT
GCAGGCGGACAATGCGTCCGGTCAGGGTGGTAAGGTTGTCGTGCAGGGTAAGAATATTCTGCTGGACAAGGGCAGCAACATCACA
GCAACCGGTGGTCAGGGCGGCGGTGAAGTGTATGTCGGTGGCGGCTGGCAGGGTAAGGACAGCAACATCCGTAATGCGGACAAG
GTGGTGATGCAGGGCGGCGCCCGCATTGACGTTTCTGCAACGCAGCAGGGTAACGGCGGTACGGCTGTGCTGTGGTCAGACAGCT
ACACCAACTTCCATGGTCAGATTAGCGCGAAGGGCGGTGAGACCGGCGGTAACGGTGGTCGGGTGGAGACCTCTTCGCACGGTAA
CCTGCAGGCATTTGGTACGGTCAGTGCATCCGCGAAGAAAGGCAAGGCGGGTAACTGGCTGCTGGACTCGGCGGATATCACCATTG
TGAATGGTAGCAATGTTAGCAAAACTGAGACGACTCAATCGCCGCCGCACACGCAATTTGCACCCACCGCTGCGGGCTCTGCGGTCA
GCAATACCAGTATCAACAACAGGCTGAACAACGGGACCAGTGTCACTATTCTGACCCATCGCACAAGAACAGGCACAGCTCAGGGC
GGGAATATTACCGTTAATGCGGCAATTAACAAAAGCAACGGAAGTGATGTCAACCTGACGCTGCAGGCTGGCGGCAACATCACGGT
AAACAACAGCATCACGTCCACCGAGGGTAAGCTGAATGTTAATCTGTCGGGCGCCAGGACCAGCAATGGCAGTATCACCATTAGCA
ATAACGCCAATATAACGACCAATGGTGGGGATATAACTGTTGGGACGACAAATACTTCAAACCGTGTGAATATATCTATTAATAACA
CTACCCTGAATGCGTCAAATGGCAACATCCAGTTGACCGGGACCGGGACCGATAGCGGGATTCTGTTTGCTGGCAACAACAGGCTG
ACGGCCAGTAACATTGCTCTTACCGGGAACAGTACGAGTGGGAATGCCATCAACCTTACAGGCACTGCCACGCTGAATGCCACGAAT
AACATTACTCTTACCGGGAGCAGTACGAGTGGGAATGCCATCAACCTTAAAGGCAACAACACGCTGACGGCCAGTAACATTACTCTT
ACCGGGGAAAGTACGAGTGGGAATGCCATCAACCTTACAGACACTACAGGCACTACCACGCTGAATGCCACGAATAACATCACTAT
GCAGGGGACCCGTGTTCAGATTAAACACTCCAACATCACCGCGGGCAACTTTGCGCTGAATGCGACAGTGGCCGGCTCTGAAATCA
GCAATACCACGCTGACGGCCACCAACAACATCAACCTGGCGGCTAAGACGAACAGTGCGAGCTCTGGTGTTTACCTGAAAGATGCA
AGAATTACATCCACCAATGGCAGTATCACGGCTAACGGTACTGCCACAGCAAACGGCAAGGCCACGCATCTGGACGGCAACGTCAC
CCTGAATGCGTCAAATGGCAGAATCAAGTTGACCGGGAACGGGCACGGTAGCGCCTCCGGGATTCTGTTTGCTGGCAACAACAGGC
TGACGGCCAGTAACATTGCTCTTACCGGGAACAGTACGAGTGGGAATGCCATCAACCTTACAGGCACTGCCACGCTGAATGCCACG

AATGACATTACTCTTACCGGGAGCAGTACGAGTGGGAATGCCATCAACCTTACAGGCACTGCCACGCTGAATGCCACGAATAACATT
ACTCTTACCGGGAGCAGTACGAGTGGGAATGCCATCAACCTTAAAGGCAACAACACGCTGACGGCCAGTAACATTACTCTTACCGGG
GAAAGTACGAGTGGG

### 8.1.6 [ETEC]FliC nucleotide sequence

ENA|STN14455|STN14455.1

**Appendix 8-6: [ETEC]FliC nucleotide sequence**

ATGGCACAAGTCATTAATACAAACAGCCTGTCGCTGTTGACCCAGAATAACCTGAACAAATCTCAGTCTTCTCTGAGCTCCGCCATTG
AACGTCTCTCTTCTGGCCTGCGTATTAACAGTGCTAAAGATGACGCAGCAGGTCAGGCGATTGCTAACCGTTTTACAGCAAATATTAA
AGGTCTGACTCAGGCTTCCCGTAACGCGAATGATGGTATTTCTGTTGCGCAGACCACTGAAGGTGCGCTGAATGAAATTAACAACAA
CCTGCAGCGTGTACGTGAACTGACTGTTCAGGCAACTAACGGTACTAACTCTGACAGCGATCTTTCTTCTATCCAGGCTGAAATTACT
CAACGTCTGGAAGAAATTGACCGTGTATCTGAGCAAACTCAGTTTAACGGCGTGAAAGTCCTTGCTGAAAATAATGAAATGAAAATT
CAGGTTGGTGCTAATGATGGTGAAACCATCACTATCAATCTGGCAAAAATTGATGCGAAAACTCTCGGCCTGGACGGTTTTAATATC
GATGGCGCGCAGAAAGCAACTGGCAGTGACCTGATTTCTAAATTTAAAGCGACAGGTACTGATAACTATGATGTTGGCGGTGATGC
TTATACTGTTAACGTAGATAGCGGAGCTGTTAAAGATACTACAGGGAATGATATTTTTGTTAGTGCAGCAGATGGTTCACTGACAAC
TAAATCTGACACAAACATAGCTGGTACAGGGATTGATGCTACAGCACTCGCAGCAGCGGCTAAGAATAAAGCACAGAATGATAAAT
TCACGTTTAATGGAGTTGAATTCACAACAACAACTGCAGCGGATGGCAATGGGAATGGTGTATATTCTGCAGAAATTGATGGTAAGT
CAGTGACATTTACTGTGACAGATGCTGACAAAAAAGCTTCTTTGATTACGAGTGAGACAGTTTACAAAAATAGCGCTGGCCTTTATAC
GACAACCAAAGTTGATAACAAGGCTGCCACACTTTCCGATCTTGATCTCAATGCAGCTAAGAAAACAGGAAGCACGTTAGTTGTTAA
CGGTGCAACTTACGATGTTAGTGCAGATGGTAAAACGATAACGGAGACTGCTTCTGGTAACAATAAAGTCATGTATCTGAGCAAATC
AGAAGGTGGTAGCCCGATTCTGGTAAACGAAGATGCAGCAAAATCGTTGCAATCTACCACCAACCCGCTCGAAACTATCGACAAAG
CATTGGCTAAAGTTGACAATCTGCGTTCTGACCTCGGTGCAGTACAAAACCGTTTCGACTCTGCTATCACCAACCTTGGCAACACCGT
AAACAACCTGTCTTCTGCCCGTAGCCGTATCGAAGATGCTGACTACGCGACCGAAGTGTCTAACATGTCTCGTGCGCAGATCCTGCA
ACAAGCGGGTACCTCTGTTCTGGCGCAGGCTAACCAGACCACGCAGAACGTACTGTCTCTGCTGCGTTAA

## 8.2 Protein sequences

### 8.2.1 EtpA full-length amino acid sequence

Uniprot: Q29XT7

**Appendix 8-7: EtpA full-length amino acid sequence**

MNRIYKLKFDKRRNELVVVSEITTGVGNAKATGSVEGEKSPRRGVRAMALSLLSGMMIMAHPAMSANLPTGGQIVAGSGSIQTPSGNQ
MNIHQNSQNMVANWNSFDIGKGNTVQFDQPSSSAVALNRVVGGGESQIMGNLKANGQVFLVNPNGVLFGEGASVSTSGFVASTRDIK
NDDFMNRRYTFSGGQKAGAAIVNQGELTTNAGGYIVLAADRVSNSGTIRTPGGKTVLAASERITLQLDNGGLMSVQVTGDVVNALVENR
GLVSARDGQVYLTALGRGMLMNTVLNVSGVVEASGMHRQDGNIVLDGGDSGVVHLSGTLQADNASGQGGKVVVQGKNILLDKGSNIT
ATGGQGGGEVYVGGGWQGKDSNIRNADKVVMQGGARIDVSATQQGNGGTAVLWSDSYTNFHGQISAKGGETGGNGGRVETSSHG
NLQAFGTVSASAKKGKAGNWLLDSADITIVNGSNVSKTETTQSPPHTQFAPTAAGSAVSNTSINNRLNNGTSVTILTHRTRTGTAQGGNIT
VNAAINKSNGSDVNLTLQAGGNITVNNSITSTEGKLNVNLSGARTSNGSITISNNANITTNGGDITVGTTNTSNRVNISINNTTLNASNGNI
QLTGTGTDSGILFAGNNRLTASNIALTGNSTSGNAINLTGTATLNATNNITLTGSSTSGNAINLKGNNTLTASNITLTGESTSGNAINLTDTTG
TTTLNATNNITMQGTRVQIKHSNITAGNFALNATVAGSEISNTTLTATNNINLAAKTNSASSGVYLKDARITSTNGSITANGTATANGKATH
LDGNVTLNASNGRIKLTGNGHGSASGILFAGNNRLTASNIALTGNSTSGNAINLTGTATLNATNDITLTGSSTSGNAINLTGTATLNATNNIT
LTGSSTSGNAINLKGNNTLTASNITLTGESTSGNAINLTDTTGTTTLNATNNITMQGTRVQIKHSNITAGNFALNATVAGSEISNTTLTATNN
INLAAKTNSASSGVYLKDARITSTNGSITANGTATANGKATHLDGNVTLNASNGRIKLTGNGHGSASGILFAGNNRLTASNIALTGNSTSGN
AINLTGTATLNATNDITLTGSSTSGNAINLTGTATLNATNNITLTGSSTSGNAINLKGNNTLTASNITLTGESTSGNAINLTDTTGTTTLNATN
NITMQGTRVQIKHSNITAGNFALNATVAGSEISNTTLTATNNINLAAKTNSASSGVYLKDARITSTNGSITANGTATANGKATHLDGNVTL
NASNGRIKLTGNGHGSASGILFAGNNRLTASNIALTGNSTSGNAINLTGTATLNATNDITLTGSSTSGNAINLTGTATLNATNNITLTGSSTS
GNAINLKGNNTLTASNITLTGESTSGNAINLTDTTGTTTLNATNNITMQGTRVQIKHSNITAGNFALNATVAGSEISNTTLTATNNINLAAKT
NSASSGVYLKDARITSTNGSITTNGTATANGKATHLDGNVTLNASNGRIKLTGNGHGSASGILFAGNNRLTASNIALTGNSTSGNAINLTGT
ATLNATNDITLTGSSTSGNAINLTGTATLNATNNITLTGSSTSGNAINLKGNNTLTASNITLTGESTSGNAINLTDTTGTTTLNATNNITMQG
TRVQIKHSNITAGNFALNATVAGSEISNTTLTATNNINLAAKTNSASSGVYLKDARITSTNGSITANGTAPANDNATYLDGNVTLNASNGSI
KLTGNGNGSTSGILFAGNNTLTASNITLTGNSEVYWQ

### 8.2.2 EtpA$^{67-447}$ amino acid sequence

**Appendix 8-8: EtpA$^{67-447}$ amino acid sequence**

| |
|---|
| PTGGQIVAGSGSIQTPSGNQMNIHQNSQNMVANWNSFDIGKGNTVQFDQPSSSAVALNRVVGGGESQIMGNLKANGQVFLVNPNG<br>VLFGEGASVSTSGFVASTRDIKNDDFMNRRYTFSGGQKAGAAIVNQGELTTNAGGYIVLAADRVSNSGTIRTPGGKTVLAASERITLQLDN<br>GGLMSVQVTGDVVNALVENRGLVSARDGQVYLTALGRGMLMNTVLNVSGVVEASGMHRQDGNIVLDGGDSGVVHLSGTLQADNAS<br>GQGGKVVVQGKNILLDKGSNITATGGQGGGEVYVGGGWQGKDSNIRNADKVVMQGGARIDVSATQQGNGGTAVLWSDSYTNFHGQ<br>ISAKGGETGGNGGRVETSSHGNLQAFGTVSASA |

### 8.2.3 EtpA$^{1-606}$ amino acid sequence

**Appendix 8-9: EtpA$^{1-606}$ amino acid sequence**

| |
|---|
| MNRIYKLKFDKRRNELVVVSEITTGVGNAKATGSVEGEKSPRRGVRAMALSLLSGMMIMAHPAMSANLPTGGQIVAGSGSIQTPSGNQ<br>MNIHQNSQNMVANWNSFDIGKGNTVQFDQPSSSAVALNRVVGGGESQIMGNLKANGQVFLVNPNGVLFGEGASVSTSGFVASTRDIK<br>NDDFMNRRYTFSGGQKAGAAIVNQGELTTNAGGYIVLAADRVSNSGTIRTPGGKTVLAASERITLQLDNGGLMSVQVTGDVVNALVENR<br>GLVSARDGQVYLTALGRGMLMNTVLNVSGVVEASGMHRQDGNIVLDGGDSGVVHLSGTLQADNASGQGGKVVVQGKNILLDKGSNIT<br>ATGGQGGGEVYVGGGWQGKDSNIRNADKVVMQGGARIDVSATQQGNGGTAVLWSDSYTNFHGQISAKGGETGGNGGRVETSSHG<br>NLQAFGTVSASAKKGKAGNWLLDSADITIVNGSNVSKTETTQSPPHTQFAPTAAGSAVSNTSINNRLNNGTSVTILTHRTRTGTAQGGNIT<br>VNAAINKSNGSDVNLTLQAGGNITVNNSITSTEGKLNVNLSGARTSNGSITISNNANITTNGGDITVGTTNTSNRVNISI |

### 8.2.4 EtpA$^{67-930}$ amino acid sequence

**Appendix 8-10: EtpA$^{67-930}$ amino acid sequence**

| |
|---|
| PTGGQIVAGSGSIQTPSGNQMNIHQNSQNMVANWNSFDIGKGNTVQFDQPSSSAVALNRVVGGGESQIMGNLKANGQVFLVNPNG<br>VLFGEGASVSTSGFVASTRDIKNDDFMNRRYTFSGGQKAGAAIVNQGELTTNAGGYIVLAADRVSNSGTIRTPGGKTVLAASERITLQLDN<br>GGLMSVQVTGDVVNALVENRGLVSARDGQVYLTALGRGMLMNTVLNVSGVVEASGMHRQDGNIVLDGGDSGVVHLSGTLQADNAS<br>GQGGKVVVQGKNILLDKGSNITATGGQGGGEVYVGGGWQGKDSNIRNADKVVMQGGARIDVSATQQGNGGTAVLWSDSYTNFHGQ<br>ISAKGGETGGNGGRVETSSHGNLQAFGTVSASAKKGKAGNWLLDSADITIVNGSNVSKTETTQSPPHTQFAPTAAGSAVSNTSINNRLNN<br>GTSVTILTHRTRTGTAQGGNITVNAAINKSNGSDVNLTLQAGGNITVNNSITSTEGKLNVNLSGARTSNGSITISNNANITTNGGDITVGTT<br>NTSNRVNISINNTTLNASNGNIQLTGTGTDSGILFAGNNRLTASNIALTGNSTSGNAINLTGTATLNATNNITLTGSSTSGNAINLKGNNTLT<br>ASNITLTGESTSGNAINLTDTTGTTTLNATNNITMQGTRVQIKHSNITAGNFALNATVAGSEISNTTLTATNNINLAAKTNSASSGVYLKDAR<br>ITSTNGSITANGTATANGKATHLDGNVTLNASNGRIKLTGNGHGSASGILFAGNNRLTASNIALTGNSTSGNAINLTGTATLNATNDITLTG<br>SSTSGNAINLTGTATLNATNNITLTGSSTSGNAINLKGNNTLTASNITLTGESTSG |

### 8.2.5 $^{ETEC}$FliC amino acid sequence

Uniprot: A0A0F3UY21

**Appendix 8-11: $^{ETEC}$FliC amino acid sequence**

| |
|---|
| MAQVINTNSLSLLTQNNLNKSQSSLSSAIERLSSGLRINSAKDDAAGQAIANRFTANIKGLTQASRNANDGISVAQTTEGALNEINNNLQRV<br>RELTVQATNGTNSDSDLSSIQAEITQRLEEIDRVSEQTQFNGVKVLAENNEMKIQVGANDGETITINLAKIDAKTLGLDGFNIDGAQKATGS<br>DLISKFKATGTDNYDVGGDAYTVNVDSGAVKDTTGNDIFVSAADGSLTTKSDTNIAGTGIDATALAAAAKNKAQNDKFTFNGVEFTTTTA<br>ADGNGNGVYSAEIDGKSVTFTVTDADKKASLITSETVYKNSAGLYTTTKVDNKAATLSDLDLNAAKKTGSTLVVNGATYDVSADGKTITETA<br>SGNNKVMYLSKSEGGSPILVNEDAAKSLQSTTNPLETIDKALAKVDNLRSDLGAVQNRFDSAITNLGNTVNNLSSARSRIEDADYATEVSN<br>MSRAQILQQAGTSVLAQANQTTQNVLSLLR |

### 8.2.6 HMW1A full-length amino acid sequence

Uniprot: WP_014550671.1

**Appendix 8-12: HMW1A full-length amino acid sequence**

MNKIYRLKFSKRLNALVAVSELARGCDHSTEKGSEKPARMKVRHLALKPLSAMLLSLGVTSIPQSVLASGLQGMDVVHGTATMQVDGNK
TIIRNSVDAIINWKQFNIDQNEMVQFLQENNNSAVFNRVTSNQISQLKGILDSNGQVFLINPNGITIGKDAIINTNGFTASTLDISNENIKAR
NFTFEQTKDKALAEIVNHGLITVGKDGSVNLIGGKVKNEGVISVNGGSISLLAGQKITISDIINPTITYSIAAPENEAVNLGDIFAKGGNINVRA
ATIRNQGKLSADSVSKDKSGNIVLSAKEGEAEIGGVISAQNQQAKGGKLMITGDKVTLKTGAVIDLSGKEGGETYLGGDERGEGKNGIQLA
KKTSLEKGSTINVSGKEKGGRAIVWGDIALIDGNINAQGSGDIAKTGGFVETSGHDLFIKDNAIVDAKEWLLDPDNVSINAETAGRSNTSED
DEYTGSGNSASTPKRNKEKTTLTNTTLESILKKGTFVNITANQRIYVNSSINLSNGSLTLWSEGRSGGGVEINNDITTGDDTRGANLTIYSGG
WVDVHKNISLGAQGNINITAKQDIAFEKGSNQVITGQGTITSGNQKGFRFNNVSLNGTGSGLQFTTKRTNKYAITNKFEGTLNISGKVNIS
MVLPKNESGYDKFKGRTYWNLTSLNVSESGEFNLTIDSRGSDSAGTLTQPYNLNGISFNKDTTFNVERNARVNFDIKAPIGINKYSSLNYAS
FNGNISVSGGGSVDFTLLASSSNVQTPGVVINSKYFNVSTGSSLRFKTSGSTKTGFSIEKDLTLNATGGNITLLQVEGTDGMIGKGIVAKKNIT
FEGGNITFGSRKAVTEIEGNVTINNNANVTLIGSDFDNHQKPLTIKKDVIINSGNLTAGGNIVNIAGNLTVESNANFKAITNFTFNVGGLFDN
KGNSNISIAKGGARFKDIDNSKNLSITTNSSSTYRTIISGNITNKNGDLNITNEGSDTEMQIGGDVSQKEGNLTISSDKINITKQITIKAGVDGE
NSDSDATNNANLTIKTKELKLTQDLNISGFNKAEITAKDGSDLTIGNTNSADGTNAKKVTFNQVKDSKISADGHKVTLHSKVETSGSNNNTE
DSSDNNAGLTIDAKNVTVNNNITSHKAVSISATSGEITTKTGTTINATTGNVEITAQTGSILGGIESSSGSVTLTATEGALAVSNISGNTVTVT
ANSGALTTLAGSTIKGTESVTTSSQSGDIGGTISGGTVEVKATESLTTQSNSKIKATTGEANVTSATGTIGGTISGNTVNVTANAGDLTVGNG
AEINATEGAATLTTSSGKLTTEASSHITSAKGQVNLSAQDGSVAGSINAANVTLNTTGKVTLTTVKGSNINATSGTLVINAKDAELNGAALG
NHTVVNATNANGSGSVIATTSSRVNITGDLITINGLNIISKNGINTVLLKGVKIDVKYIQPGIASVDEVIEAKRILEKVKDLSDEEREALAKLGVS
AVRFIEPNNTITVDTQNEFATRPLSRIVISEGRACFSNSDGATVCVNIADNGR

## 8.2.7 FHA full-length amino acid sequence

Uniprot: CPN83729.1

**Appendix 8-13: FHA full-length amino acid sequence**

MNTNLYRLVFSHVRGMLVPVSEHCTVGNTFCGRTRGQARSGARATSLSVAPNALAWALMLACTGLPLVTHAQGLVPQGQTQVLQGGN
KVPVVNIADPNSGGVSHNKFQQFNVANPGVVFNNGLTDGVSRIGGALTKNPNLTRQASAILAEVTDTSPSRLAGTLEVYGKGADLIIANP
NGISVNGLSTLNASNLTLTTGRPSVNGGRIGLDVQQGTVTIERGGVNATGLGYFDVVARLVKLQGAVSSKQGKPLADIAVVAGANRYDHA
TRRATPIAAGARGAAAGAYAIDGTAAGAMYGKHITLVSSDSGLGVRQLGSLSSPSAITVSSQGEIALGDATVQRGPLSLKGAGVVSAGKLA
SGGGAVNVAGGGAVKIASASSVGNLAVQGGGKVQATLLNAGGTLLVSGRQAVQLGAASSRQALSVNAGGALKADKLSATRRVDVDGK
QAVALGSASSNALSVRAGGALKAGKLSATGRLDVDGKQAVTLGSVASDGALSVSAGGNLRAKQLVSSAQLEVRGQREVALDDASSARG
MTVVAAGALAARNLQSKGAIGVQGGEAVSVANANSDAELRVRGRGQVDLHDLSAARGADISGEGRVNIGRARSDSDVKVSAHGALSID
SMTALGAIGVQAGGSVSAKDMRSRGAVTVSGGGAVNLGDVQSDGQVRATSAGAMTVRDVAAAADLALQAGDALQAGFLKSAGAMT
VNGRDAVRLDGAHAGGQLRVSSDGQAALGSLAAKGELTVSAARAATVAELKSLDNISVTGGERVSVQSVNSASRVAISAHGALDVGKVS
AKSGIGLEGWGAVGADSLGSDGAISVSGRDAVRVDQARSLADISLGAEGGATLGAVEAAGSIDVRGGSTVAANSLHANRDVRVSGKDAV
RVTAATSGGGLHVSSGRQLDLGAVQARGALALDGGAGVALQSAKASGTLHVQGGEHLDLGTLAAVGAVDVNGTGDVRVAKLVSDAGA
DLQAGRSMTLGIVDTTGDLQARAQQKLELGSVKSDGGLQAAAGGALSLAAAEVAGALELSGQGVTVDRASASRARIDSTGSVGIGALKA
GAVEAASPRRARRALRQDFFTPGSVVVRAQGNVTVGRGDPHQGVLAQGDIIMDAKGGTLLLRNDALTENGTVTISADSAVLEHSTIESKIS
QSVLAAKGDKGKPAVSVKVAKKLFLNGTLRAVNDNNETMSGRQIDVVDGRPQITDAVTGEARKDESVVSDAALVADGGPIVVEAGELVS
HAGGIGNGRNKENGASVTVRTTGNLVNKGYISAGKQGVL

## 8.2.8 HpmA full-length amino acid sequence

Uniprot: SUC39485.1

**Appendix 8-14: HpmA full-length amino acid sequence**

MKSKNFKLSPSGRLAASLAIIFVSLNAYGNGIVPDAGHQGPDVSAVNGGTQVINIVTPNNEGISHNQYQDFNVGKPGAVFNNALEAGQS
QLAGHLNANSNLNGQAASLILNEVVSRNPSFLLGQQEVFGIAAEYVLSNPNGITCDGCGFINTSRSSLVVGNPLFENGQLKGYSTLNNTNLL
SLGKNGLNTTGLLDLIAPRIDSRGKITAAEISAFTGQNTFSQHFDILSSQKPVSALDSYFFGSMQSGRIRIINTAEGSGVKLAGKFTADNDLSV
KADNIQTDSQVRYDSYDKDGSENYQNYRGGITVNNSGSSQTLTKTELKGKNITLVASSHNQIKASDLMGDDITLQGADLTIDGKQLQQKE
TDIDNRWFYSWKYDVTKEKEQIQQIGSQIDAKNNATLTATKGDVTLDAAKINAGNNLAINANKDIHINGLVEKESRSENGNKRNHTSRLES
GSWSNSHQTETLKASELTAGKDLGLDAQGSITAQGAKLHANENVLVNAKDNINLNVQKTNNDKTVTDNHVMWGGIGGGQNKNNNN
QQQVSHATQLTADGQLLLAADNNVNITGSQVKGNQGAFVKTTQGDVVIDNALSETISKIDERTGTAFNITKSSHKNETNKQTSTGSELISD
AQLTVVSGNDVNVIGSLIKSADKLGIHSLGDINVKSAQQVTKIDDEKTSLAITGHAKEVEDKQYSAGFHITHTTNKNTSTETEQANSTISGAN
VDLQANKNVTFAGSDLKTTAGNASITGDNVAFVSTENKKQTDNTDTTISGGFSYTGGVDKVGSKADFSI

### 8.2.9 HxuA full-length amino acid sequence

PDB: 4RT6

**Appendix 8-15: HxuA full-length amino acid sequence**

SARDLPQGSSVVVGEANVSTIGNKMTIDQKTPTTQIDWHSFDIGQNKEVEFKQPDANSVAYNRVTGGNASQIQGKLTANGKVYLANPN
GVIITQGAEINVAGLFATTKDLERISENGNGNGNKFTRKLKDGQVVKEGQVINKGKIKAKDFVVLNGDKVINEGEIDATNNGKVYLSSGYNF
TFTLSDSSISVALEDNAVQSIVQNEGIIKAGDITLNAKGRNQALDSLVMNNGVLEATKVSNKNGKVVLSADDVQLNNKSDIKGESEVVFTN
EPKNKIKITSQTGSKVTSPKINFTGKSVNINGDFGRDDSKAHYNEEHKRLDTEVNIDVPDNENIRIAEKDNTGTGTGTDSFIQTGALSSLLAN
NGKVNLKGKDVNISGRIHIDSFRGSDSLLKLTNQGHIKINHADIHSTGRLFFITSLQNEKDSQSDITITDSKINLGNGAMGLGRSLDKENCDN
QRWCRTETSQRKKFDVHMRNVVFDQVDDVVVAGGFKKVNLDNIVATGKTNFYIDGGVSRNNSRYEYGVLDLDKRTLLSELDQRRRRWK
YYNDLDLDMNKAYWHRFDMFATKNTGRSTIKDTEINISNSKINLKNGFVHLLAEKIKLDNSKIDITFDKDNSQDISTQINRLGMNGKVSMV
NSHIKIVGDEKSDISAKAPYATMFLIGELIGEKSSIFVKSHQGYTFRTDGDTKIAGKNSKDDLKITAINTGGRTGKEVIINGAPGSIDNDANIAN
MAFTIGDNANTKTTIENADITALAPNGGTAYLSSKGVEIEVNPNSNFTFFELPREKNFNQTKIKGDSTKLSERGFARLYDKINGVRASNLSAE
QLNVTDASEKIINTKLVSSLDVEKLVSVAVSDAGKGSEEQQFGDKGNNTKVSVGELETEQ