REGULATION OF VIRULENCE DETERMINANTS IN PATHOGENIC *ESCHERICHIA COLI*

by

MUNIRAH MESHARI T. ALHAMMADI

A THESIS

SUBMITTED TO THE UNIVERSITY OF BIRMINGHAM FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

INSTITUTE OF MICROBIOLOGY AND INFECTION
SCHOOL OF BIOCSIENCES
UNIVERSITY OF BIRMINGHAM
NOVEMBER 2022

Abstract

Enteroaggregative and uropathogenic *Escherichia coli* are both pathogenic strains that likely evolved from commensal *E. coli* by acquisition of virulence genes. These pathogens, EAEC and UPEC, can cause problems in the intestines and the bladder, of the host. To understand how pathogens, initiate their infection, it is important to consider how they behave in nutritionally challenging conditions, such as low glucose levels in the environments and how they co-exist with other pathogens and microbiota. Many bacteria can respond to environmental changes using small effector molecules that activate the regulator proteins that are needed to cope with the circumstances. One of the best studied transcription regulators in *E. coli* is cyclic AMP receptor protein (CRP) which can regulate more than 100 genes/operon in *E. coli,* hence, comes the name a global regulator.

The regulation of virulence in EAEC has been intensively studied, and AggR is known to be the master regulator for the expression of most chromosomally-encoded and plasmid-encoded of virulence genes. However, for some genes that both analysis of promoter sequences, and experimental data, show that they are not involved in the AggR regulon. My project focuses on the regulation of such EAEC virulence determinants, during starvation condition. One of these is Pic (Protein involved in colonisation), encoded by chromosomal *pic* gene, which encodes an important enzyme, needed in the early stage of infection. The architecture of *pic* promoter was compared between EAEC and UPEC using different methodologies, using site-direct mutagenesis assays, the 5`rapid amplification of cDNA ends (5`RACE) method, and promoter activity measurements. My results show that *pic* is a CRP-dependent promoter with an unusual promoter architecture. In EAEC strain 042: it is the first example of an 'ambiguous' CRP-activiated promoter.

My unexpected results from studying *pic* regulation leads me to ask if there are more virulence genes involves in the EAEC CRP regulon. Hence, my second objective was to use chromatin immunoprecipitation with high- throughput sequencing (ChIP-seq) with the wild type EAEC 042 to locate the full complement of DNA site for CRP. My results showed more than 300 CRP targets in the EAEC 042 genome. Comparison of my data with the situation in commensal *E. coli* K-12 strains revealed 31 targets that were specific genes for EAEC 042, and I focussed further work on these. The results of these new CRP targets were validated using *in vivo* and *in vitro* experiments. Results showed involvement of CRP at some of these virulence genes

targets, whereas, at other, CRP appeared to play a role in chromosomal organisation. One outcome from the CRP ChIP-seq data was discovery of a very strong DNA site for CRP in the regulatory region upstream of the genes encoding polysialic acid transport proteins, KpsMII, located in EAEC 042 chromosome. In my investigations, I used *in vivo* and *in vitro* assays to demonstrate the involvement of CRP and the possibility of indirect regulation by CRP of *kpsMII* expression. My data argue that bacterial pathogens use global transcription regulators to overcome challenging conditions like starvation hence and assure the production of virulence factors in the early stages of infection.

Dedications

This work is dedicated to my parents, Jawzaa Alhammadi and Meshari Alhammadi, and to my lovely son, Ibrahim Alotaibi, for their endless love, support, and inspiration.

Acknowledgments.

I am deeply indebted to Professor Stephen Busby for guiding me through this project and for his kindness and encouragement all these years, especially during the COVID -19 crisis. I am also grateful to my second supervisor, Dr. Jessica Blair, for her support during my assessments and for answering all my questions. I also would like to express my deepest appreciation to Dr. Douglas Browning for teaching me and directing me through the project, as well as being my friend during the rough study years.

I also would like to thank and be grateful for all Busby Lab members during my studies in 2018-2022: Rita Godfrey, Joanne Hothersall, Radwa Wahab, Georgina Lloyd, and Gabrielle Christie; and all Master students who helped me in this project with their amazing effort. I could not have completed this journey without the support and collaboration with Dr. James Haycocks; his patience and kindness in teaching me the ChIP-seq and the analysis were outstanding. This work would not be a great story without Busby Lab`s support. I had the pleasure of working and attending in Grainger's Lab, and I am grateful for their advice and assistance. And above all, I would like to thank my sponsors, Princess Nourah Bint Abdulrahman University (PNU, Saudi Arabia); the Saudi Arabia Cultural Bureau (SACB, United Kingdom); and the Minister of Education in Saudi Arabia, for giving me this opportunity and scholarship to accomplish my goals.

I would like to extend my sincere thanks to my family, my parents, and their patience during all these years. I would also like to thank my lovely son, Ibrahim, for his patience and support throughout my life. This work is dedicated in honour of all my loves ones and relatives who could not be here today to celebrate with me, my uncles, and my grandparents. Lastly, I would like to acknowledge my gratitude to my amazing brothers, Mashal, Yasir, Saud, Khaled, Muhanned, and Waleed, for their support and encouragement over all these years. I also would like to thank all my friends in the U.K. and Saudi Arabia for helping me during these years and for their support. I am also thankful for all the people who helped me and supported me during the COVID -19 crisis in the UK, the Emirates and Saudi Arabia.

Table of Contents

List of Figures.

List of Tables.

List of abbreviations.

ABC transporters	ATP-binding cassette transporters
AggR	Aggregative regulator
BKO	Both promoter knockouts
cAMP	Cyclic adenosine monophosphate
ChIP-seq	Chromatin immunoprecipitation- sequence
CRP	Cyclic AMP receptor protein
CRP AR1 AR2	CRP active region 1 active region 2
DKO	Dowenstream promoter kockout
DNA	Deoxyribonucleic acid
E. coli	*Escherichia coli*
EAEC	Enteroaggregative *E. coli*
EAST-1	Enteroaggregative Heat Stable Toxin-1
Fis	Factor for inversion stimulation protein
HlyA	Haemolysin A
H-NS	Histone-like nucleoid-structuring protein
IBCs	Intracellular bacterial communities
IHF	Integration host factor
KpsMII	Polysialic acid transport proteins; group 2
LB	Luria-Bertani media
OC	Open promoter complex
Ops	Operon polarity suppressor
Pet	Plasmid encoded toxin
Pic	Protein involved in colonisation
(p)ppGpp	Guanosine pentaphosphate or tetraphosphate
PTS	Phosphotransferase system
QIRs	Quiescent intracellular reservoirs
5` RACE	5` Rapid amplification of cDNA ends
RNA	Ribonucleic acid
RNAP	RNA polymerase
RNA-seq	RNA-sequencing
SD	Shine–Dalgarno
SDW	Sterile distilled water

sRNA	Small RNA
T5SS	Type 5 secreation system
TF	Transcription factor
TSS	Transript start site
UKO	Upstream promoter knockout
UPEC	Uropathogenic *E. coli*
UTI	Urinary tract infection
VFGs	Virulance factor genes
WT	Wild type

Chapter 1

Introduction

1.1 *Escherichia coli*

In the tree of life kingdom, bacteria and Archaea are classified as prokaryotic cells that lack a nucleus, and organelles, in their structure. These organisms can be found in all environments, and, in some situations, they live in symbiotic relationships with other organisms, such as plants or animals. Bacteria are microscopic organisms that can be classified into five categories according to their shapes: rod (bacilli), spherical (cocci), comma (vibrios), spiral (spirilla), corkscrew (spirochaetes), and they can be found alone as a single cell, or in pairs or chains, dependent on the morphology of the bacterial species. One of the best-studied species is *Escherichia coli* due to its heterologous gene expression in it (Goeddel, 1990), which was named after Theodor Escherich, who discovered this microorganism in 1885 (Hacker and Blum-Oehler, 2007).

Escherichia coli is a Gram-negative rod-shaped facultative anaerobic bacterium that has been classified within the *Enterobacteriaceae* family. *Escherichia coli* can thrive in a variety of environments, including the gut of vertebrates, where it is an important component of the host microbiome. In fact, *E. coli* colonises the gastrointestinal tract of humans within the first hours of human infant life. This species and its human host have been co-existing in good health for mutual benefits, like producing vitamin K and vitamin B12 (Lawrence and Roth, 1996; Kaper *et al.*, 2004). Another key benefit to the host intestines from intestinal microbiota (including commensal *E. coli*) is protection against colonization by external pathogens and balancing the gut microbiota (Chow *et al.,* 2011). Some *E. coli* strains have evolved to increase their ability to colonise the human gut, and many of these strains cause diseases. Three common clinical diseases can result from an infection by pathogenic *E. coli*: urinary tract infection, diarrhoea, and sepsis/meningitis (Nataro and Kaper, 1998).

1.2 Pathogenic *E. coli*

Many commensal *E. coli* strains remain harmless in the human gut because they lack genes encoding virulence determinants such as enterotoxins, invasion proteins, and adhesion factors. When strains acquire these genetic elements, they can evolve by locking them into their genome (Kaper *et al.*, 2004). However, only the successful combination of certain virulence determinants results in *E. coli* pathotypes that cause illness to healthy humans. Six such pathogenic *E. coli* types have been described in the last decade, that differ according to their pathogenicity mechanisms: enterotoxigenic *E. coli* (ETEC), enteropathogenic *E. coli* (EPEC), enteroinvasive *E. coli* (EIEC), diffusely adherent *E. coli* (DAEC), enterohaemorrhagic *E. coli* (EHEC) and enteroaggregative *E. coli* (EAEC) (Nataro and

Kaper, 1998). Additionally, other pathogenic *E. coli* can cause extraintestinal infections in other parts of the human body. Examples are uropathogenic *E. coli* (UPEC), which causes bladder and urinary tract infections, and neonatal meningitis-associated *E. coli* (NMEC), which can cause cardiovascular or brain infection.

1.2.1 EAEC Pathogenesis

EAEC causes watery, mucoid, and bloody diarrhoea in children and travellers, worldwide. EAEC is the second most common cause of travellers' diarrhoea, second only to ETEC, and, over the past decades, outbreaks of EAEC have flared up in the United Kingdom (Smith *et al.*, 1997), Italy (Scavia *et al.*, 2008), and Japan (Itoh *et al.*, 1997). EAEC was first identified by Jim Nataro in 1987 who recognised aggregative and diffuse strains as different phenotypes, both associated with diarrhoea. The classical phenotype of disease causing EAEC is 'stacked-brick' aggregative bacterial adherence on Hep-2 cells due to EAEC cells sticking to each other and adhering to Hep-2 cells in a process known as auto-aggregation (Nataro *et al.*, 1987). At the molecular level, EAEC can be defined by plasmid-encoded aggregative adherence (AA) protein and the *att* locus, which encodes an ABC transporter system (Baudry *et al.,* 1990; Okeke *et al.,* 2000). Importantly, a typical EAEC can be recognised by the presence of the transcription activator AggR, which is often plasmid-encoded and regulates the expression of most of the EAEC virulence factors. Due to the high ability of EAEC strains to form biofilms, biofilm assays can be used to screen large numbers of strains from clinical cases. Hence Iwanaga and his colleagues (2002) in Japan used biofilm assays and found that 89% of the strains tested were EAEC.

The hallmark of EAEC pathogenesis is the presence of aggregative adhesion fimbriae (AAF), the production of toxins, and the formation of biofilm, as host intestinal cells are colonised. One EAEC strains, EAEC 042, is taken as a paradigm, and has been studied in detail. Figure 1.1 illustrates the infection process of this strain. Which is initiated by the attachment of AAF to human intestinal cells. This can be reproduced with mammalian cultured cells, such as HeLa and Hep-2 (Navarro-Garcia and Elias, 2011). In EAEC strain 042, AAF/II is the most common AAFs, binding to fibronectin and type IV collagen on host intestinal cells. The interaction of EAEC AAF to host cells is supported the help of dispersin; a hydrophilic protein that attaches non-covalently to the EAEC outer membrane, thereby allowing AAF extension from the cell surface and subsequent interaction with host cells to form a robust biofilm (Farfan *et al*., 2008).

The second stage of EAEC pathogenesis is the secretion of toxins. Autotransporter proteins are associated with adhesion, and these are secreted by type V protein secretion systems (Wells *et al.*, 2007). EAEC 042 produces two autotransporter proteins during growth in poor environmental conditions, and these help it to survive and colonise in the host colon. One is an enterotoxin, named as Pet (plasmid encoded toxin) (Navarro-Garcia and Elias, 2011). The other autotransporter protein is Pic (protein involved in colonisation), which is involved in mucinolytic activity. Additionally, some EAEC strains can produces enterotoxins like *Shigella* enterotoxin 1 (ShET1), and enteroaggregative *E. coli* heat-stable toxin-1 (EAST-1), which are involved in secretory and watery diarrhoea (Kaper *et al.*, 2004). These stages lead to the third stage of EAEC infection, which is intestinal secretion due to the EAEC colonisation, and mucosal inflammation.

Molecular and genetic studies of EAEC pathogenesis have underscored the importance of the aggregation regulator protein (AggR) in triggering the expression of many of the virulence genes of EAEC 042 (Sarantuya *et al.*, 2004; Nataro, 2005). However, some virulence genes, such as *pic* and *pet,* are not regulated via AggR. Rossiter *et al.* (2011) showed that *pet* gene transcription initiates from a single promoter that is co-dependent on two global transcription factors, the cyclic AMP receptor protein (CRP) and the factor for inversion stimulation (Fis). CRP and Fis are global transcription factors that are involved in the regulations of many genes beside *pet* in *E. coli* (Santos-Zavaleta *et al.*, 2019).

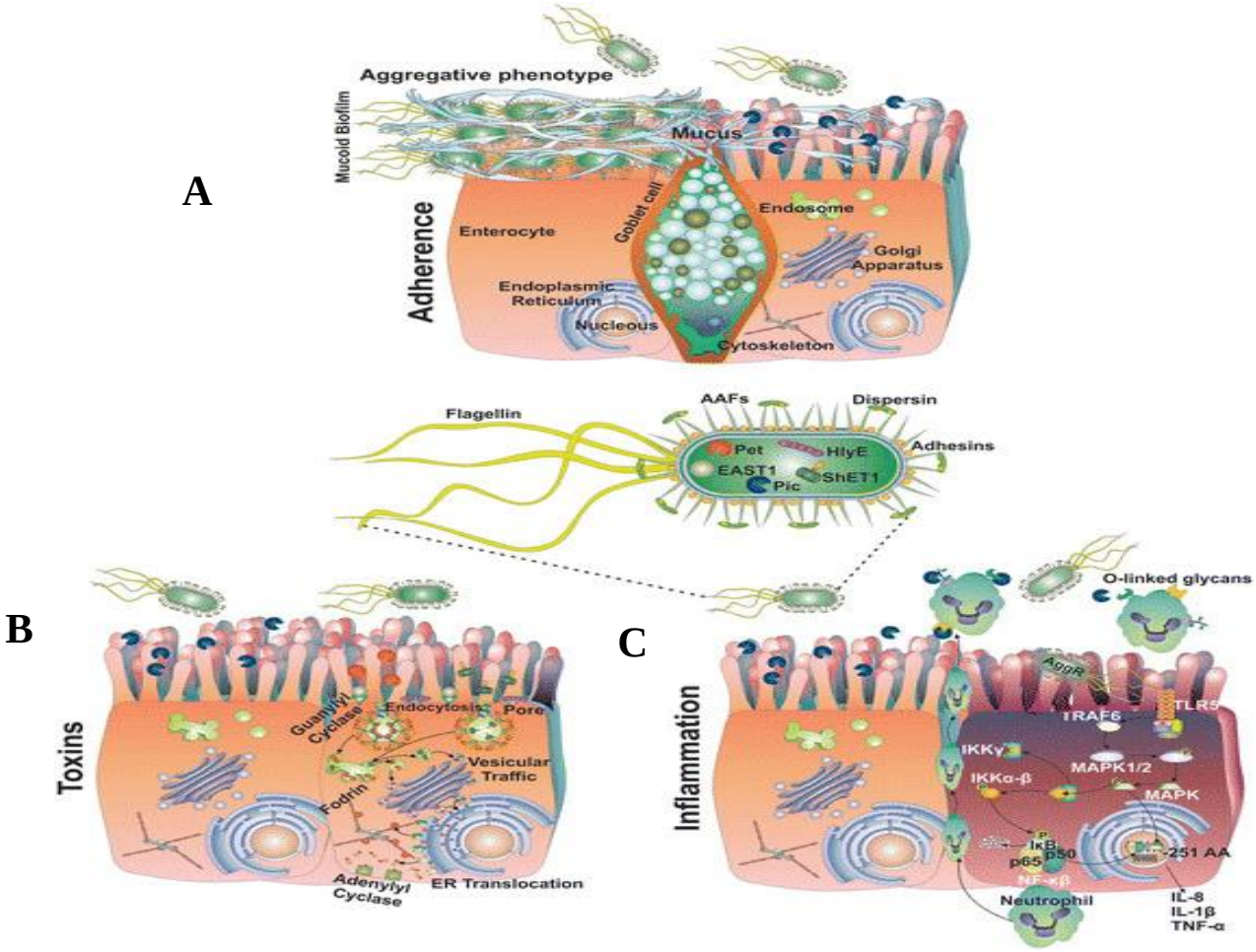


Figure 1.1: EAEC pathogenesis.

The panel shows the disease process that starts with (**A**) EAEC 042 adherence by AAF to human intestinal cells. (**B**) Toxin productions such as Pet (orange), EAST1 (white), Pic (blue) and ShET1(green). (**C**) Mucosal inflammation resulting from damage to the intestine and mucosa (taken from Elias and Navarro-Garcia, 2016).

1.2.2 UPEC pathogenesis

UPEC is responsible for more than 80% of urinary tract infection (UTI) cases (Klein and Hultgren, 2020). This pathogen can cause cystitis in the bladder and pyelonephritis in the kidneys. UTI symptomatically can be classified into three stages depending on the severity of the infection: to urosepsis syndrome (urine), pyelonephritis (kidneys), and cystitis (bladder) (Foxman, 2014).

The pathogenesis of UPEC, shown in Figure 1.2, starts with colonisation of the urethra. UPEC then ascends into the bladder where it appears in urine as planktonic cells. When the pathogen enters the bladder, it starts the colonisation there with adhesion to the uroepithelium layer, via the pathogen fimbria adhesin H (FimH), which binds to uroplakin receptors on the superficial facet cells (in uroepithelium layer) (Croxen and Finlay, 2010; Klein and Hultgren, 2020). Then UPEC initiates biofilm formation, leading to invasion, replication and the formation of biofilm-like complexes known as intracellular bacterial colonies (IBCs) that help to protect the pathogen from the host's immune systems. Secretion of haemolysin A (HlyA) toxin then triggers apoptosis of host cells, including peeling of the uroepithelium, which allows access to transitional cells for more invasion. UPEC can live inside these transitional cells as quiescent intracellular reservoirs (QIRs). Motile UPEC then moves to colonise the kidneys and futher damage to the host tissue generates a high risk of septicemia (Bien *et al.*, 2012; Terlizzi *et al*., 2017). In the kidney infection stage, the host's immune system responds to UPEC invasion by activating proinflammatory mediators that protect the kidney from pathogens, but until this response becomes excessive, pathogens can cause chronic pyelonephritis, which can cause damage to kidney cells and renal failure (Bien *et al.*, 2012).

Multiple virulence factors contribute to UPEC infection of UTI cells, including flagella, adhesins, iron-acquisition systems and polysaccharide capsules (Bien *et al.*, 2012). Moreover, UPEC, similar to EAEC, expresses an autotransporter mucinase (Pic), which targets the mucus, and polysialic acid transport proteins (KpsMII). However, UPEC toxins differ from EAEC, hence, UPEC strain CFT073 encodes α-haemolysin (HlyA) and vacuolating autotransporter toxin (Vat) (Wiles *et al.*, 2008).

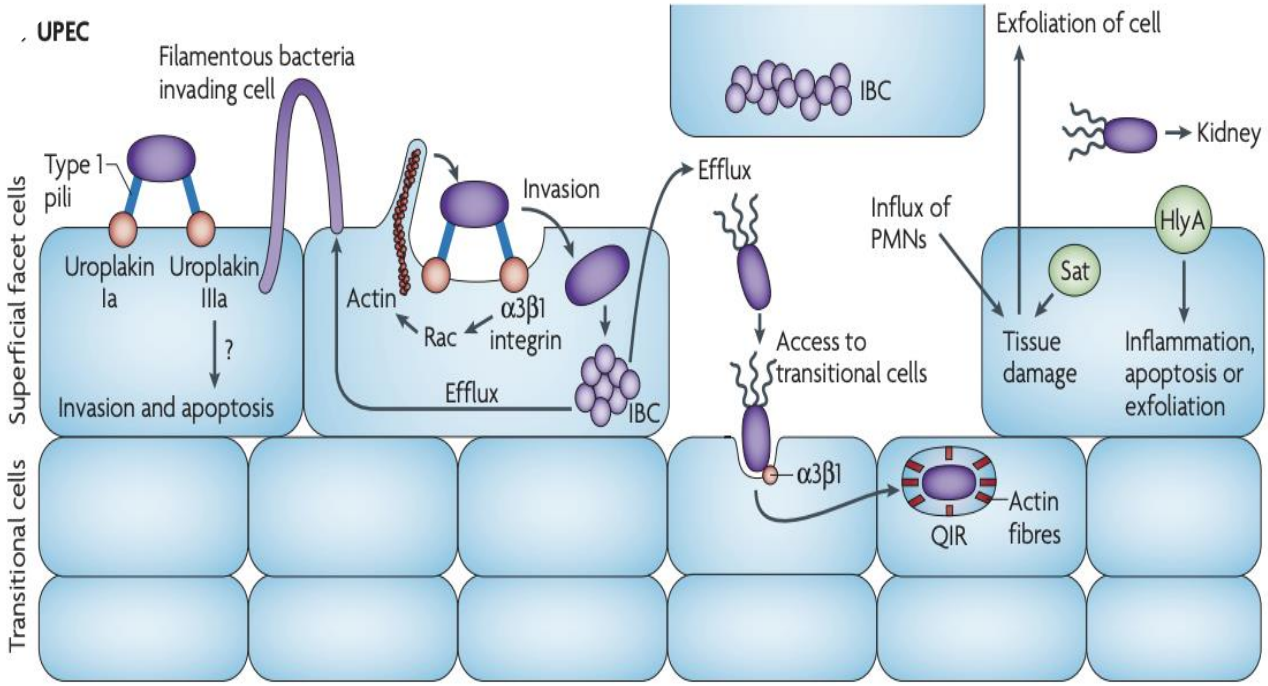


Figure 1.2: UPEC pathogenesis in the bladder stage.

The figure illustrates the pathogenesis of UPEC in the bladder. Starting from the left side, UPEC binds to uroplakin receptor (Ia and IIIa) on the uroepithelium to initiate invasion. The binding of type 1 pili on the pathogen to α3β1 integrins on host cells then leads to the formation of intracellular bacterial communities (IBCs). Filamentous bacteria fluxing out from infected cells and invading the surrounding superficial cells. The secretion of toxins like haemolysin A (HlyA) prevents Akt proteins activation that are crucial in cell survival and proliferation. This results in apoptosis, which exposes transitional cells and extends the invasion to these cells, with the formation of quiescent intracellular reservoirs (QIRs). Motile UPEC then continues to infect the kidneys. Figure taken from Croxen and Finlay (2010).

1.3 Autotransporters in pathogenic *E. coli*

Autotransporter proteins are frequently involved in pathogenesis. They have the property of transporting themselves from their cytoplasmic site of synthesis to the extracellular milieu (Henderson *et al.*, 1998). This is known as type 5 secretion, mediated by type V in the protein secretion systems (T5SS) (Henderson *et al.*, 2004). The T5SS consists of three subfamilies: the autotransporter system (Va), the two-partner secretion pathway (Vb); and the trimeric autotransporter adhesin system (Vc) (Henderson *et al*., 2004; Wells *et al*., 2010).

Autotransporter proteins are composed of three functional domains: a signal sequence, a passenger domain which carries the virulence function, and a translocation β -barrel unit (Figure 1.3). The signal sequence is found in the N-terminal end of the protein, and its function is to target the protein to the inner membrane, so it can be transferred to the periplasm (Henderson *et al.*, 1998). The passenger domain is destined to be the secreted extracellular protein. The translocation unit (named as the β-domain) is present in the C-terminal domain at the end of protein. The translocation here follows the secretion across the outer membrane (Henderson *et al.*, 2004). In type Vb pathways, the passenger domain and the translocation unit are two separate proteins, TpsA and TpsB, each of them carrying their own signal sequence to target them to the inner membrane. For the type Vc pathway, the translocation unit trimerizes, with the C-terminal domains of each monomer contributing four β-strands to form a complete pore (Henderson *et al*., 1998).

Type Va autotransporter proteins share similar structures. Interestingly, the EAEC and UPEC autotransporter protein Pic and EAEC Pet autotransporter protein all use the type Va pathway (AT-1 pathway). The autotransporter translocation domain consists of 250-300 amino acids and is assumed to form 14 antiparallel β-strands as it inserts into the outer membrane (Yen *et al.,* 2002). However, passenger domains can differ greatly, from one type Va autotransporter to another. This allows passenger domains to be involved in highly diverse functions, such as adhesion, serum resistance, haemagglutination, protease activity, biofilm formation and toxin activities. Basically, in the Va autotransporters, the passenger domain fits into a common framework formed by the rest of the protein (Emsley *et al.*, 1994).

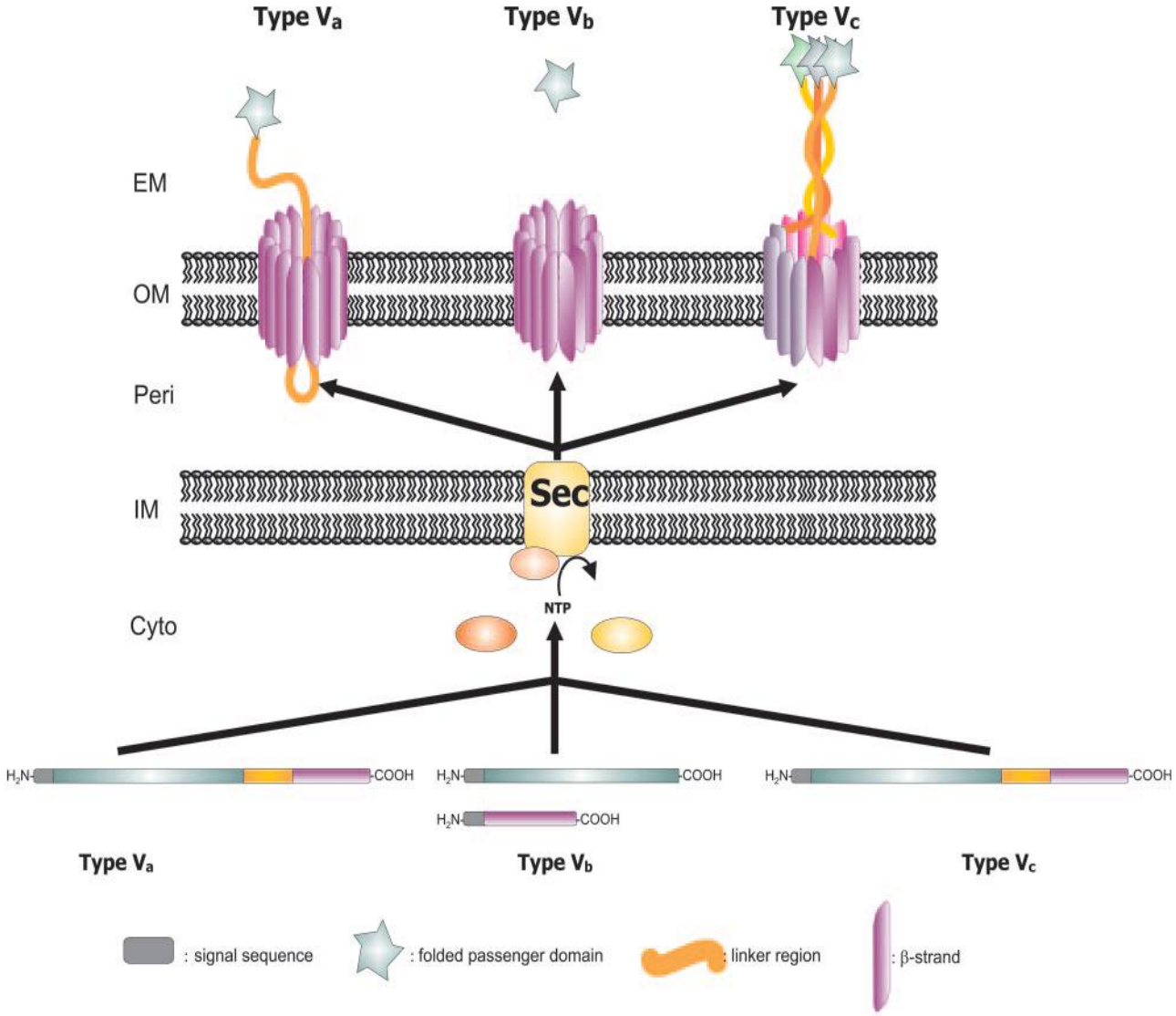


Figure 1.3: Schematic summary of type V secretion.

The diagram illustrates type V secretion with a type Va autotransporter protein type (left), a two-partner type Vb system (middle), and a type Vc system (right). Different parts of the autotransporter are colour coded as indicated in the figure. These autotransporters are exported to the periplasm membrane by the Sec machinery. The signal sequence is cleaved in the periplasmic space and the β-domain ejects the passenger domain after insertion into the bacterial outer membrane. The diagram is taken from Henderson *et al.*, 2004.

1.4 Transport of capsular polysaccharides via ABC transporter system

The bacterial capsule is defined as the extracellular structure that surrounds bacterial cells, and it can be viewed via a microscope as a thick layer (Kröncke *et al*., 1990). These extracellular materials include O antigens, lipopolysaccharide (LPS) and capsular (K) antigens (Whitfield and Roberts, 1999). Capsule are made as long polysaccharide chains which have negative charge to provide a highly hydrated capsular layer to the cell. It is believed that the role of these polysaccharides is to mediate the interactions between the bacteria and their surroundings, and to protect bacteria from the innate and acquired immunity of their host (Vimr *et al.*, 2004; Whitfield, 2006).

The many K antigens of *E. coli* are classified into 4 groups, based on genetic, biochemical, and serological data (group 1, group 2, group 3, and group 4 capsule; Whitfield, 2006) Non-protein molecules such as capsular polysaccharides cannot use the Gram-negative Sec pathway for export to the cell surface. Hence, capsule molecules are exported to the cell surface using two systems: ABC transporter-dependent (required for group 2 and group 3 capsules) or Wzy-dependent systems (required for groups 1 and 4 capsules) (Cuthbertson *et al*., 2009; Cuthbertson *et al*., 2010). Note that the ABC transporter dependent pathway uses an ATP-binding cassette (ATP) to provide the energy to facilitate the translocation of the polysaccharide chains.

In mucosal pathogens, such as pathogenic *E. coli* strains, *Haemophilus influenzae* and *Neisseria meningitidis*, the capsular (K) polysaccharides were defined as major virulence factors that use ABC transporter-dependent pathways to export the polymer to the cell surface (Willis and Whitfield, 2013). Capsules that are exported by this system have a unique structure at the end of the polysaccharides chain. This structure consists of a phospholipid named lyso-phosphatidylglycerol and it is linked to the polysaccharide via a "β-linked poly-3-deoxy-D-*manno*-oct-2-ulosonic acid (Kdo) linker" (Willis *et al.,* 2013). It has been suggested that the role of this terminal linker is to attach the polysaccharide chain to the cell surface (Fischer *et al*., 1982). The organisation of an ABC transporter-dependent system is shown in Figure 1.4. After the polymer is synthesised in the cytoplasm, it is exported across the inner membrane via KpsT, which consists of two identical nucleotide binding domains (NBD), and KpsM, which consists of two integral membrane domains (TMD). These proteins only export the mature polysaccharide chains to the periplasm, and the process is then continued by two other proteins: KpsE (polysaccharide co-polymerase (PCP)) and KpsD, which are located at the outer membrane (outer membrane polysaccharide (OPX)) (Cuthbertson, *et al*., 2009). Structural studies

show that the 258-aa KpsM protein forms a spanning pore in the inner membrane, whereas the 223-aa KpsT faces the cytoplasmic membrane and contains the ATP binding cassette (Figure 1.4) (Chaudhuri *et al.*, 2010).

E. coli group 2 K antigen is expressed at 37˚C, when it is inside a host, but not in external environments where there are lower temperatures (Whitfield, 2006). The essential genes for group 2 K antigen are located in the chromosomal *kps* cluster. This consists of three regions: region 1 (encoding *kpsFEDUCS*), region 2, that is strain serotype-specific (encoding the enzymes for polymer synthesis) and region 3 (*kpsMT*) (Roberts, 1996). Together, these regions encoded the proteins that involved in capsule polysaccharide biosynthesis, maturation, and location (Steenbergen and Vimr, 2008)

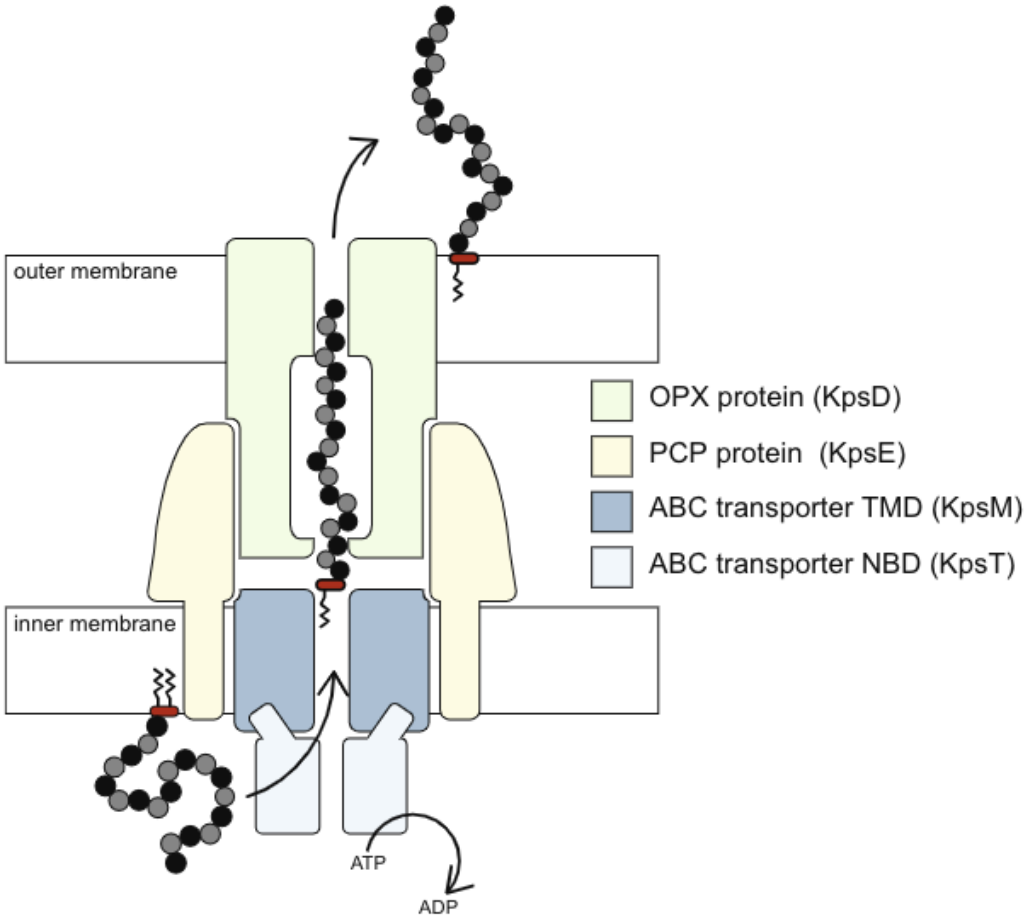


Figure 1.4: Model for capsular polysaccharide export by the ABC system in *E. coli*.

The cartoon illustrates the four main proteins of the Kps system that are involved in exporting polysaccharide chains from their site of synthesis (the cytoplasm) to the cell surface. The mature polysaccharide chains are shown here as black and grey chains that identify via the terminal (lyso) phosphatidylglycerol-poly-Kdo, illustrated here as the red shape, to attach the polysaccharide chain to the cell surface. On the inner membrane are shown KpsT and KpsM which form an ABC transporter. The outer membrane consists of KpsE and KpsD that are involved in export of the polysaccharide chain to the cell surface. The model is adapted from Willis and Whitfield (2013).

1.5 Virulence determinants in pathogenic *E. coli* strains

The chromosome and pAA plasmid of EAEC strain 042 encode many virulence factor genes (VFGs) that are involved in pathogenesis (the chromosomal VFGs of EAEC 042 are listed in Table 1.1A, and plasmid encoded VFGs are in Table 1.1B). Expression of the majority of these VFGs, such as *aaf*, *aat* and *aap*, is controlled by AggR (Yasir *et al.*, 2019). This transcription activator belongs to the AraC family of transcription regulators and is encoded on the pAA plasmid (Elias *et al*., 1999; Morin *et al.*, 2013). AggR is also involved in the regulation of some chromosomally encoded VFGs, such as a T6SS (Dudley *et al.*, 2006). Previous studies have argued that AggR is the global regulator of expression of most virulence determinants in EAEC. However, not all VFGs such as *pet*, belong to the regulon controlled by CRP, a well-characterised global transcription factor (Rossiter *et al*., 2011). In the case of UPEC strain CFT073, all its virulence factors are chromosomally encoded, as this strain especially does not carry a plasmid (Rodriguez-Siek *et al.*, 2005). VFGs in the UPEC strain CFT073 are listed in Table 1.2, for comparison purposes with EAEC strain 042.

1.5.1 Plasmid –encoded Toxin (Pet)

The full-length Pet protein is a serine protease autotransporter of the *Enterobacteriaceae* SPATE family. It has a molecular weight of 104 kDa and is encoded on the EAEC strain 042 plasmid pAA2 (Navarro-Garcia *et al.*, 2001; Navarro-Garcia *et al*., 2010). Interestingly, it has been shown that Pet toxin has both cytotoxic and enterotoxic effects (Navarro-Garcia *et al.*, 2001). These effects are dependent on the activity of the serine protease and on its internalization into target cells (Navarro-Garcia *et al.*, 2007).

Studies with cultured epithelial cells revealed that Pet enters host cells, causing cytopathic effects (Navarro-Garcia *et al.*, 2001). The role of Pet is to damage the host cell cytoskeleton by interfering with actin stress fibres, thereby causing mucus release (Navarro-Garcia and Elias, 2011). Furthermore, Pet targets α-fodrin, which are cytoskeleton proteins and play a crucial role in preserving the membranes' structural stability. The target of Pet on α-fodrin is by binding and cleaving between residues M1198 and V1199. This results in redistribution of fodrin within cells, the formation of intercellular aggregates, and membrane blebs (Bennett and Gilligan, 1993). These studies show that Pet is an important virulence element in EAEC pathogenesis.

Table 1.1: A- Chromosomally-encoded virulence factor genes in EAEC strain 042.

Virulence factor	Position in contig	Protein function	Accession number
aaiC	4895043..4895549	Type VI secretion protein	FN554766
air	4259271..4270691	Enteroaggregative immunoglobulin repeat protein	FN554766
capU	5117126..5118214	Hexosyltransferase homolog	CU928145
chuA	4024131..4026113	Outer membrane hemin receptor	FN554766
eilA	4257522..4259219	Salmonella HilA homolog	FN554766
fyuA	2301968..2303989	Siderophore receptor	FN554766
gad	4042972..4044372	Glutamate decarboxylase	CU928163
gad	1699700..1701100	Glutamate decarboxylase	FN554766
hra	3390633..3391424	Heat-resistant agglutinin	CP043942
irp2	2282673..2288780	High molecular weight protein 2 non-ribosomal peptide synthetase	NZ_BGNW01000020
kpsE	3443946..3445094	Capsule polysaccharide export inner-membrane protein	FN554766
kpsMII	3457805..3458581	Polysialic acid transport protein; Group 2 capsule	FN554766
lpfA	4351536..4352108	Long polar fimbriae	AB198066
mchB	3416292..3416585	Microcin H47 part of colicin H	AJ009631
mchC	3414471..3416021	MchC protein	FN554766
mchF	3410477..3412573	ABC transporter protein MchF	FN554766
mcmA	3410003..3410224	Microcin M part of colicin H	FN554766
pic	4924840..4928958	Serine protease autotransporters of Enterobacteriaceae (SPATE)	FN554766
sitA	1582859..1583773	Iron transport protein	FN554766
terC	3232455..3233168	Tellurium ion resistance protein	CP007491
terC	3609332..3610290	Tellurium ion resistance protein	MG591698

B- Plasmid-encoded virulence factor genes in EAEC strain 042.

Virulence factor	Position in contig	Protein function	Accession number
ORF3	1397..2425	Isoprenoid Biosynthesis	FN554767
ORF4	2429..2968	Putative isopentenyl-diphosphate delta-isomerase	FN554767
aafA	39652..40134	AAF/II major fimbrial subunit	FN554767
aafB	22554..22994	AAF/II minor adhesin. Enterobacteria AfaD invasin protein	FN554767
aafC	23011..25581	Usher, AAF/II assembly unit	AF114828
aafD	38499..39248	Chaperone, AAF/II assembly unit	FN554767
aap	42684..43034	Dispersin, antiaggregation protein	FN554767
aar	46246..46437	AggR-activated regulator	042

aatA	5647..6885	Dispersin transporter protein	FN554767
aggR	41080..41877	AraC- family transcriptional activator	FN554767
astA	32628..32744	EAST-1 heat-stable toxin	AB042002
capU	16404..17492	Hexosyltransferase homolog	FN554767
pet	28073..31960	Plasmid-encoded toxin	AF056581
traT	60523..61299	Outer membrane protein complement resistance	FN554767

*****The tables were produced by genomicepidemiology.org.

Table 1.2: Chromosomally encoded virulence factor genes in UPEC strain CFT073.

Virulence factor	Position in contig	Protein function	Accession number
cea	1168463..1168702	Colicin E1	GG772217
chuA	4086073..4088055	Outer membrane hemin receptor	UFZT01000001
clbB	2304280..2313901	Hybrid non-ribosomal peptide / polyketide megasynthase	NZ_CP028714
clbB	2304280..2313901	Hybrid non-ribosomal peptide / polyketide megasynthase	NZ_QLAC01000024
focCsfaE	1189406..1190101	-	CP002211
focG	1193363..1193866	F1C adhesion	DQ301498
foci	1188817..1189365	S fimbrial/F1C minor subunit	AM690762
fyuA	2247129..2249150	Siderophore receptor	ADTO01000058
gad	4104924..4106324	Glutamate decarboxylase	CP001671
gad	1756318..1757718	Glutamate decarboxylase	CP001671
iha	3448776..3450866	Adherence protein	AE014075
ireA	4935990..4938038	Siderophore receptor	AE014075
iroN	1198162..1200339	Enterobactin siderophore receptor protein	AE014075
irp2	2227128..2231359	High molecular weight protein 2 non-ribosomal peptide synthetase	NZ_PXVW01000053
irp2	2227128..2231359	High molecular weight protein 2 non-ribosomal peptide synthetase	NZ_QLAC01000001
irp2	2227128..2231359	High molecular weight protein 2 non-ribosomal peptide synthetase	NZ_UIHY01000010
irp2	2227128..2231359	High molecular weight protein 2 non-ribosomal peptide synthetase	NZ_UIKF01000007
iss	1422880..1423221	Increased serum survival	AE014075
iucC	3464728..3466470	Aerobactin synthetase	AAJW02000016
iutA	3461193..3463391	Ferric aerobactin receptor	UFZT01000001
kpsE	3513147..3514295	Capsule polysaccharide export inner-membrane protein	AE014075
kpsMII	3528516..3529292	Polysialic acid transport protein; Group 2 capsule	MG736912
mchB	1176756..1177049	Microcin H47 part of colicin H	AE014075
mchC	1177321..1178871	MchC protein	AE014075
mchF	1180751..1182865	ABC transporter protein MchF	AE014075
mcmA	1183119..1183397	Microcin M part of colicin H	AJ586887
ompT	629100..630053	Outer membrane protease (protein protease 7)	AE014075
papA_F7-1	4948203..4948787	Major pilin subunit F7-1	AF447814
papA_F7-2	3436965..3437531	Major pilin subunit F7-2	ECOFIMAA
papC	4944977..4947487	Outer membrane usher P fimbriae	ADTJ01000356
papC	3433739..3436249	Outer membrane usher P fimbriae	JNRA01000049
papC	4944977..4947487	Outer membrane usher P fimbriae	JOSS01000102
pic	326209..330324	Serine protease autotransporters of Enterobacteriaceae (SPATE)	AE014075
sat	3456362..3460249	Secreted autotransporter toxin	AE014075
sitA	1449293..1450150	Iron transport protein	AM072350

tcpC	2202763..2203686	Tir domain-containing protein	GQ903008
terC	3277741..3278454	Tellurium ion resistance protein	CP000468
terC	3678652..3679609	Tellurium ion resistance protein	MG591698
usp	127216..128997	Uropathogenic specific protein	FQSQ01000009
vat	371877..376007	Vacuolating autotransporter toxin	AE014075
yfcV	2737155..2737721	Fimbrial protein	UFZT01000001

*****The table was produced by genomicepidemiology.org.

1.5.2 Protein involved in colonisation (Pic)

Pic identified as the second SPATE member in EAEC strain 042, has a molecular weight of 114 kDa. Pic, in contrast to Pet, is chromosomally encoded in EAEC, and also in UPEC (Navarro-Garcia *et al.*, 2001; Navarro-Garcia *et al.*, 2010). It has been suggested that Pic is involved in the early stage of EAEC pathogenesis by stimulating colonisation of the intestines due to its mucinolytic activity (Henderson *et al.*, 1999). Pic activates the pathways that secrete mucus rapidly, which causes a large fusion of mucin granules with the goblet cells apical membrane and complete cavitation of these cells (Specian and Oliver, 1991). Gutierrez-Jimenez *et al.* (2008) reported that Pic mucinolytic activity was reduced when sialic acid was removed from mucin. In order to generate carbon sources and penetrate the mucus layer, the dual function of Pic as a mucinase and a mucus secretagogue is important. This enables EAEC to dwell within the mucus layer while also having access to epithelial cells.

Pic is also found in UPEC, where it has a similar function in urinary tract infections. It is known as PicU and has 96% identity at the level of amino acid sequence to EAEC Pic (Parham *et al.*, 2004). Navarro-Garcia *et al*. (2010) observed that EAEC Δ*pic* mutants cannot provoke mucus hypersecretion compared to EAEC wild type, indicating an important role of Pic in pathogenesis. Interestingly, purified Pic protease is capable of causing hypersecretion of intestinal mucus, whereas no effect was found observed when Pic protein was heat-denatured (Navarro-Garcia *et al.*, 2010). Hence, Pic mucinase is considered to be an important infection determinant in both EAEC and UPEC.

1.5.3 Enteroaggregative Heat Stable Toxin-1 (EAST-1)

In contrast to Pic and Pet, EAST-1 toxin is found in both pathogenic and non-pathogenic commensal *E. coli.* In EAEC strain 042, this toxin is encoded by the pAA plasmid via *astA*. The mature toxin is 38 amino acid polypeptides, that contains four cysteines, at positions 17, 20, 24 and 27 which form two disulfide bridges (Yamamoto and Nakazawa, 1997). The EAST-1 toxin acts by interacting with the guanylyl cyclase C receptor of host intestinal cells, thereby causing an increase in intracellular cGMP levels, that trigger anion secretion and diarrhoea (Bertin *et al.*, 1998). Studies with pigs and cattle (Bertin *et al.*, 1998; Osek, 2003; Yamamoto and Nakazawa, 1997) concluded that EAST1 alone

did not cause diarrhoea and found a strong association between EAST-1 and CS31A adhesin. Thus, EAST-1 is a co-virulence factor in diarrheagenic *E. coli*.

1.5.4 Polysialic acid transport protein: Group 2 capsule (KpsMII)

Group 2 capsule is considered as a virulence factor for infection (Russo *et al*., 1994). As outlined in Section 1.4, the genetic locus responsible for group 2 capsule in EAEC 042 consists of three regions: region 1 (*kpsFEDUCS*), region 2 (EC042_3236, EC042_3237 and EC042_3238) and region 3 (*kpsMT*) (Figure 1.5). The role of KpsMII is to translocate the polymer across the cytoplasmic membrane, and it also acts as a membrane anchor. Hence, mutations in region three (*kpsMT*) lead to the accumulation of polysaccharides inside the cell. For example, Pigeon and Silver (1994) showed that the GE93 substitution (at position 93; glycine residue→glutamic acid) in the KpsMII protein reduces polymer export through the inner membrane. Zong *et al.* (2016) showed that a *kpsM* knock out in strain PCN033 results in a significant decrease in colonisation ability. Additionally, they found that the knockout led to reductions in adhesion, resistance to complement-mediated serum bactericidal activity, and resistance to phagocytosis.

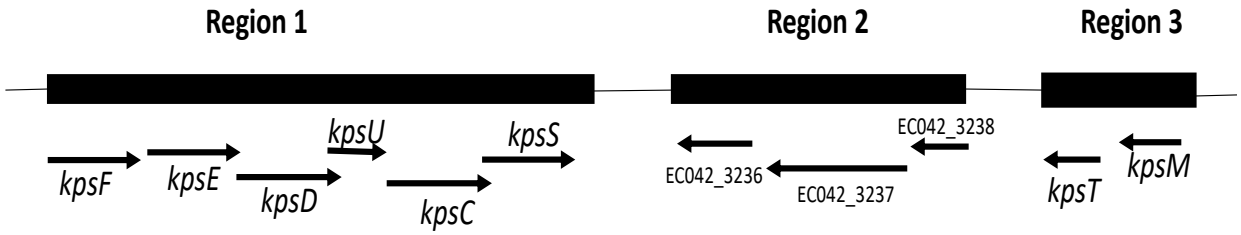


Figure 1.5: Organization of the genetic locus responsible for the group 2 capsule in EAEC strain 042.

The Figure illustrates the organisation of group 2 capsule genes in the EAEC 042 chromosome. Region 1 consists of *kpsFEDUCS* that encode proteins involved in the assembly of capsule components, which include the K antigen of *E. coli*. Region 2 consists of *EC042_3236, EC042_3237* and *EC042_3238* that encode enzymes for capsular biosynthesis. Region 3 consists of *kpsMT*, which encodes an ABC transport protein that is responsible for polymer translocation through cell inner membranes.

1.6 Bacterial Gene Expression

1.6.1 Transcription in bacteria

The term transcription refers to the process when one strand of DNA is used as a template for RNA synthesis by a DNA-dependent RNA polymerase (RNAP). Transcription produces various RNA molecules such as mRNA (messenger RNA), tRNA (transfer RNA), non-coding RNA and rRNA (ribosomal RNA); each of them has a role in transcription or protein synthesis. Messenger RNAs are used as a template for the synthesis of proteins in a process named translation. In *E. coli*, and likely most bacteria, the translation of mRNA begins before its transcription is completed (Hardin *et al.*, 2018).

In most bacteria for most of the time, only a subset of genes are being transcribed. Specific transcripts are switched on or switched off, depending on the environmental conditions, and much (though not all) of this regulation occurs at the level of transcription initiation which plays a crucial role in the control of gene expression in bacteria. The process of transcription occurs in four phases: initiation, promoter escape, elongation, and termination. When the mRNA is made, it sometimes goes through a processing step before it is translated into protein. In most bacteria, transcription initiates when RNAP recognises a promoter. The 'players' in this process are the multi-subunits DNA -dependent RNA polymerase (RNAP), the base sequence of the promoter, and various transcription factors which can help to hinder initiation, and often regulate the process in response according to environmental conditions (Browning and Busby ,2016; Griesenbeck *et al.*, 2017).

1.6.2 RNA Polymerase (RNAP)

Transcription in *E. coli* is done by a multi-subunit enzyme, RNAP, which has a molecular mass of 450 kDa. The "core" of this polymerase is five subunits: beta β, beta prime β', alphas α (two copies) and omega ω. The core enzyme itself is able to transcribe DNA alone, however, it is incapable of locating promoter regions and is not competent to initiate transcription with double stranded DNA (Browning and Busby, 2004). For these reasons, the σ subunit is the key factor for specific transcript initiation, and is essential for promoter recognition (Lee *et al.*, 2012). Hence, to initiate bacterial transcription, the $\beta\beta\alpha_2\omega\sigma$ assembly is required, and this is named as "holoenzyme".

Structural studies have shown that RNAP has the shape of a crab claw, with two pincers surrounding a cleft, that contains the RNAP active centre, which is where the template DNA strand is threaded (Chakraborty *et al.*, 2012). The β subunit forms one claw whilst the β' subunit forms the other. The β and β' subunits are assembled by the α dimer that acts as a hinge for the claw. The small ω subunit attaches to the β' subunit and so is in the lower claw (Mathew and Chatterji, 2006). The β and β' subunits provide the catalytic centre for RNA synthesis in addition to the binding site for flanking DNA (Chakraborty *et al.*, 2012). The schematic diagram of RNAP is illustrated in Figure 1.6.

Each α subunit, consist of two domains: the amino terminal domain (αNTD) and the carboxyl terminal domain (αCTD). αNTD (residues 8-235) is important in RNAP assembly, and it binds β or β' in the core enzyme (Murakami, 2015). αCTD (residues 250-329) binds to upstream sequences (UP element) at many promoters and interacts with many transcription factors and hence is important for regulation (Browning and Busby, 2004). αNTD and αCTD are linked by a 13 amino acid linker that plays a role in αCTD flexibility, allowing it to bind at different functional locations (Ebright and Busby, 1995).

1.6.3 Sigma factors

σ factors have a crucial role in promoter recognition by RNAP and in the formation of the open complex on a DNA template, which supports RNAP in site-specific transcript initiation. As the job of sigma factors is to lead the RNAP to target promoters, it dissociates from the RNAP as it begins to elongate its transcript during the promoter escape phase (Murakami, 2015). *E. coli* strains possess several σ factors, each of which plays a regulatory role. In most conditions, most transcript depend on the housekeeping σ_{70} factor, it is replaced by other σ factors in certain conditions. For example, σ_{32} is mobilised in response to temperature and stress conditions; and σ_{54} in limited nitrogen conditions (Russell, 2002).

All housekeeping σ factors, and many alternative σ factors, consist of four distinct independently folding domains. σD1, σD2, σD3 and σD4, that are connected to each other by flexible linkers. Each domain has a function and σD2, σD3 and σD4 are involved in the recognition of different promoter elements and subsequent DNA melting and the transcriptionally competent open complex formation (Lee *et al*, 2012) (Figure 1.6).

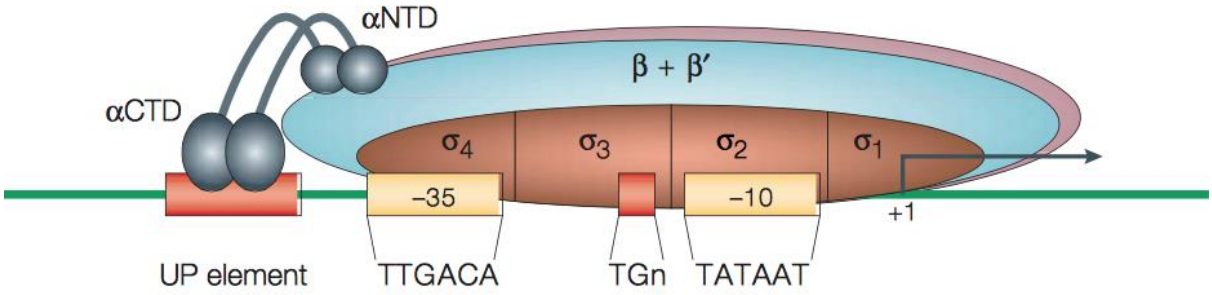


Figure 1.6: Schematic drawing of RNAP holoenzyme structure.

The Figure shows the RNAP binding site at a promoter region. The holoenzyme 'RNAP' shows in the cartoon as follows: beta β 'light blue oval', beta prime β' 'pink oval', alphas α (two copies; grey sphere), and sigma σ ' brown oval'. The following RNAP domains attach to promoter elements: σD2 to the -10 element and the discriminator, σD3 to the extended -10 element, σD4 to the -35 element, and αCTD to the UP element. The Figure is adapted from Browning and Busby (2004).

1.6.4 Promoter Regions

Promoters are segments of DNA that are located upstream of genes, and they are where RNAP binds to start transcription (Russell, 2002). Recognition of these regions by RNAP is the responsibility of the σ subunit. When the RNAP holoenzyme attaches to a promoter, the DNA is still in the double-stranded helical form and this stage is named the "closed promoter complex". Therefore, this attachment then leads to unwinding of the double-stranded DNA helix around the transcription initiation site and this structure is termed the "open complex" (Rivetti *et al*, 1999).

All bacterial promoters contain key sequence elements: two of them are the -10-hexamer element (consensus 5`-TATAAT-3`) and the -35-hexamer element (consensus 5`-TTGACA-3`). These are found upstream of the transcript start point, which by convention is denoted as +1. At many *E. coli* promoters, the -10-hexamer element and the -35-hexamer element are separated by 17 bp, and this distance is crucial in the efficiency of promoters (Feklistov, 2013). Other functional elements at bacterial promoters include the upstream element (UP-element) which is an A/T rich sequence and located upstream of the -35 hexamer element, the extended -10 element (ex-10, consensus 5`-TG-3`) located one bp upstream from the -10 hexamer element, and the discriminator elements, downstream and adjacent to the -10 hexamer, involved in stabilising open complexes (Lee *et al.*, 2012; Feklistov, 2013). In the case of holo-RNAP binding to promoters, the σ domains 2, 3 and 4 are responsible for recognition of specific promoter elements (Figure 1.6).

1.6.5 Transcription initiation in *E. coli*

The initiation stage of transcription can be subdivided into promoter recognition, isomerisation "forming of transcription bubble", DNA scrunching and promoter escape (Figure 1.7) (Browning and Busby, 2004; Murakami, 2015). The process begins when the core RNAP binds a σ factor. This is crucial for RNAP to recognise a promoter. The holoenzyme often binds first to the promoter -35 element, with the DNA still as a double helix. Attachment of holo-RNAP to promoter DNA becomes tighter and this to leads to untwisting of ±17 bp in the region starting from the -10 element, shaping the open promoter complex (OC). The richness of A/T bases in the -10-hexamer element helps to break DNA helices more easily than GC bases (Hardin *et al.*, 2018). In the open complex, RNAP becomes oriented properly to start the transcription of the DNA strand usually starting 12 bases downstream from the highly conserved non-template strand A_2 in of the -10 hexamer (consensus 5`-T<u>A</u>TAAT-3`; Russell, 2002).

The synthesis of RNA starts at the transcription bubble where the DNA region is melted and the bubble is maintained by the β and β' subunits (Borukhov and Nudler, 2008). When 9 to 15 RNA nucleotides have been linked together, σ_{70} detaches from the RNAP which enters the promoter escape phase. After that, RNAP completes gene transcription and the elongation become productive. RNAP continues to untwist the double helix of DNA resulting in torsional stress in the DNA. However, the DNA helix recovers by untwisting behind the transcription bubble. Around 17 bp of DNA remains untwisted as the RNAP continues transcription at an average of speed of 30 to 50 bases per second (Russell, 2002). When RNAP reaches the end of the transcript, it terminates transcription either with or without the help of the Rho terminator protein (Russell, 2002).

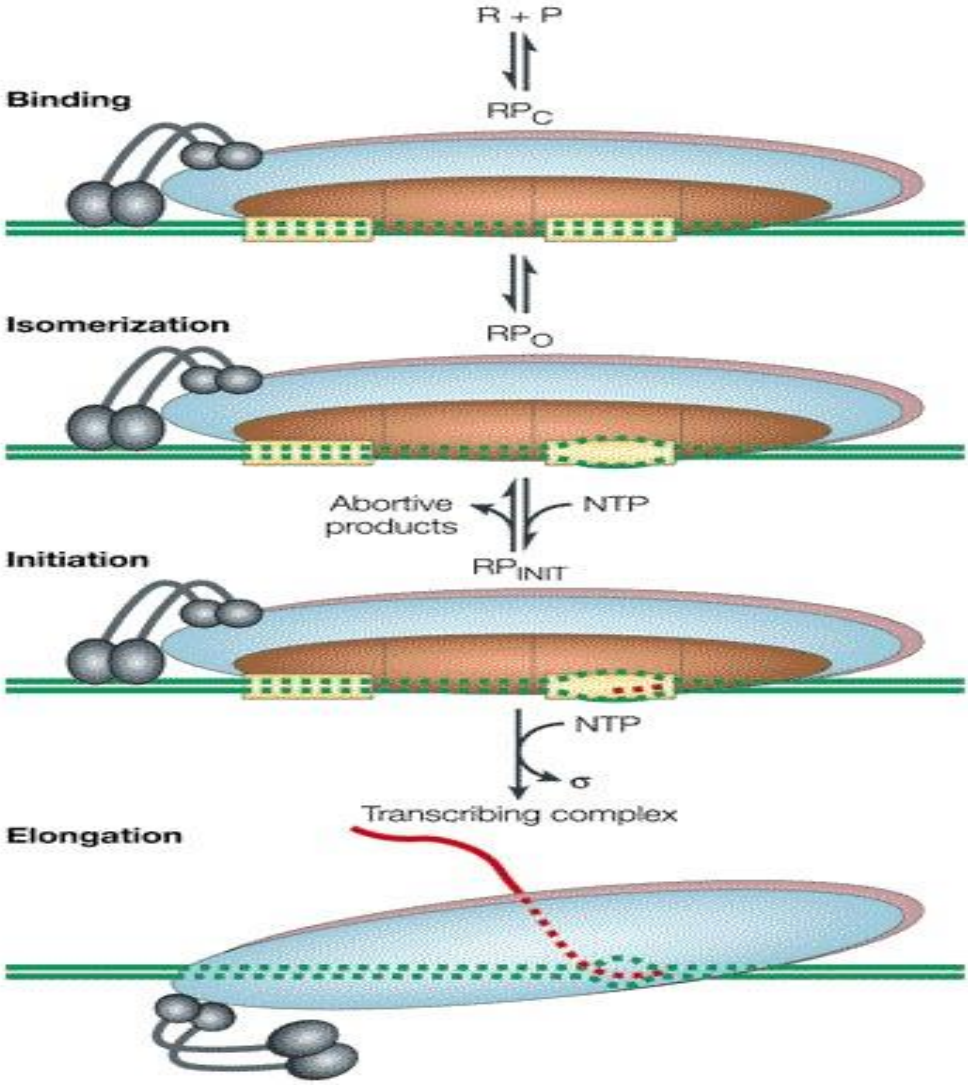


Figure 1.7: Transcription initiation stages in *E. coli*.

The diagram shows the different stages of transcript initiation. **Binding** represents the phase of RNAP-σ (R) binding to promoter DNA(P) to form a closed complex (RPC). **Isomerization** is the stage where the DNA promoter starts to untwist to form the open complex (RPO). **Initiation** shows the initiation complex (RPINIT) and synthesis of RNA (shown as a dashed red line). **Elongation** represents the final stage, where the RNA chain, shown as a red line, grows. Adapted from Browning and Busby (2004).

1.7 Regulation of transcript initiation by transcription factors

Much of the regulation of transcription in bacteria is due to transcription factor (TF) proteins. These play an important role by activating or repressing gene transcription in response to environmental signals. In most cases, transcription factors have a specific DNA-binding 'operator', located either near or a little upstream from the target promoter sequence, dependent on their mode of actions. It was once believed that TFs would have to break the hydrogen bonds between base pairs to recognise target sequences. However, because the major and minor grooves of double-stranded DNA are studded with base sequence information, TFs can recognise their targets without the need to open up the double helix (Alberts *et al.*, 2002).

Regulation at any promoter can be simple, with just one transcription factor involved, or complex, when one or more transcription factors work together to control promoter activity. Transcription factors can be sorted into two categories: bespoke transcription factors that regulate a small number of promoters, such as *E. coli* strain K-12, AraC, or global transcription factors that regulate more than scores (or even hundreds) genes such as *E. coli* cAMP receptor protein (CRP) (Dhiman and Schleif, 2000; Shimada *et al.,* 2011). Transcription factor roles can be divided into two groups, depending on their influence on the promoter activity: activators or repressors. Some transcription factors (such as CRP) can function as either an activator or a repressor depending on the architecture of the target promoter.

1.7.1 Transcription activation

Activation of transcription can be simple at some promoters, where a single activator is required to recruit RNAP. In most of these cases, one of three activation mechanism operates, depending on the location of activator binding with respect to key promoter elements. Hence, Class I activation occurs when the activator binding site is located upstream of the promoter -35 element. Here the activator helps recruit RNAP to the target promoter via direct interaction with RNAP αCTD (Figure 1.8a). The flexibility of the linker that joins αNTD and αCTD in RNAP permits the activator to bind at certain different locations upstream of the promoter. One of the best studied examples of Class I activation is the activation of the *lac* promoter by CRP (Ebright, 1993).

In Class II activation, the activator binds to an operator that overlaps the target promoter -35 element and interacts with domain 4 of the holoenzyme σ subunit, again, helping to recruit RNAP to the promoter to initiate the transcription (Figure 1.8b). In some cases, the activator interacts with other parts in RNAP, such as αNTD and αCTD (Browning and Busby, 2004). An example of this is the action of bacteriophage λ CI protein at the bacteriophage λ PRM promoter (Nickels *et al*., 2002).

The third commonly occurring mechanism of activation is when the activator binds the spacer sequence in the target promoter, between the -35 and -10 elements. Here, the activator changes the conformation of the target promoter and allows RNAP to bind the promoter -35 and -10 elements efficiently (Figure 1.8c) (Browning and Busby ,2004). Good examples of this are found with members of the MerR family of activators, where the space between the -35 and -10 elements in the target promoter is greater than the optimal ±17 bp for binding of the σ subunit in holo-RNAP so the MerR-like regulator binds this space and twists the DNA to make a good binding site for the σ subunit (Brown *et al*., 2003).

1.7.2 Transcription repression

At many bacterial promoters, transcription factors play a repressing role, thereby decreasing transcription of specific target genes. In most such cases, the repressing factor blocks RNAP from binding to the target promoters. Three simple mechanisms, involving a single repressor, have been identified: steric hindrance, DNA looping and anti-activator (Browning and Busby, 2004). In the steric hindrance mechanism, the repressor blocks promoter elements from RNAP by binding over key promoter elements (Figure 1.9a), an example being repression by LacI at the *lac* promoter (Müller *et al*, 1996). The second looping mechanism usually involves more than one operator: here, when repressors bind, they cause the DNA to loop, which closes off promoter elements and prevents RNAP binding (Figure 1.9b). An example is the repression by GalR at the *galETK* promoter region (Choy and Adhya, 1992). The third repression mechanism involves disabling an activator protein. Usually, the repressor interacts directly with the activator, and a good example is the interaction of CytR with CRP at promoters such as *deoP2* (Figure 1.9c) (Shin *et al.*, 2001).

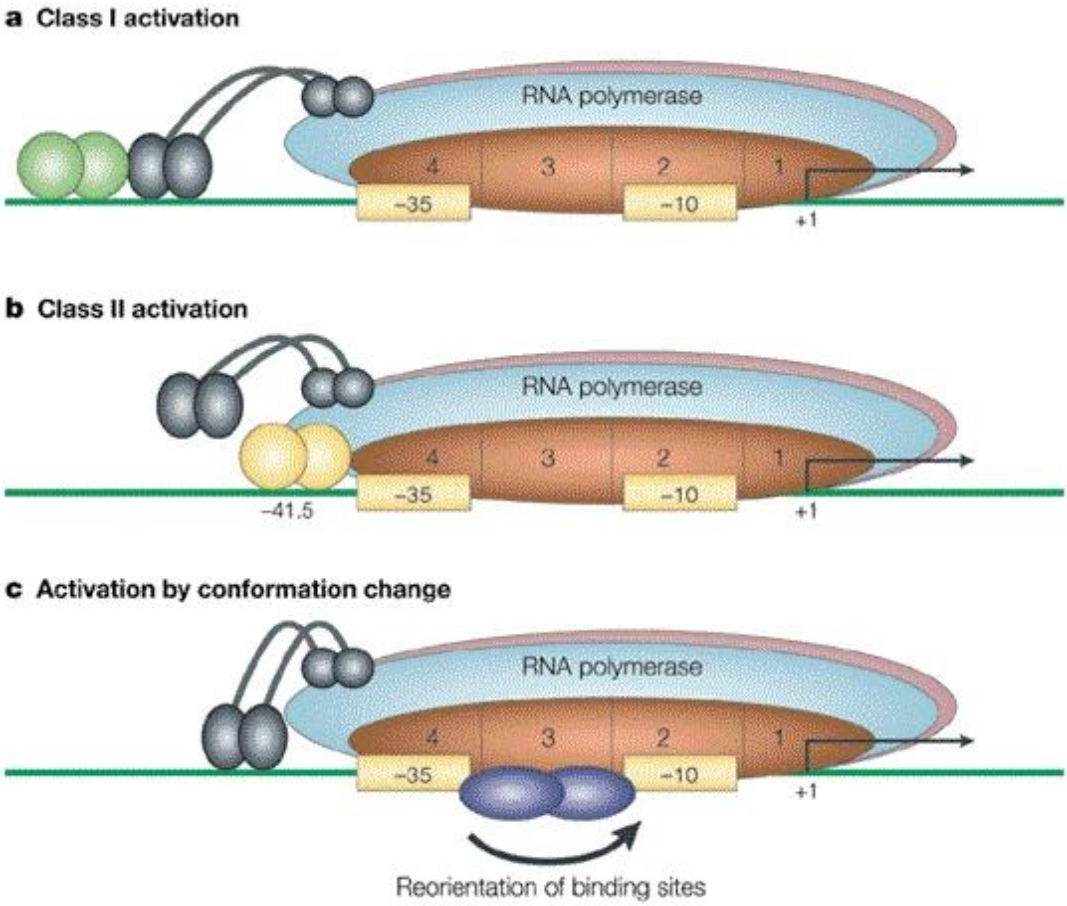


Figure 1.8: Simple promoter activation mechanisms

This diagram shows three ways of activation at simple promoters: **a**. Class I activation, an activator binds upstream of the promoter and interacts with αCTD in RNAP. **b**. Class II activation, here, the activator binding site overlaps the promoter -35 element and so the activator makes direct contact with the RNAP σ subunit. **c**. The activator binds to the spacer sequence between the promoter -35 and -10 elements and reorientates these elements, so they are competent to interact with RNAP. Figure taken from Browning and Busby (2004).

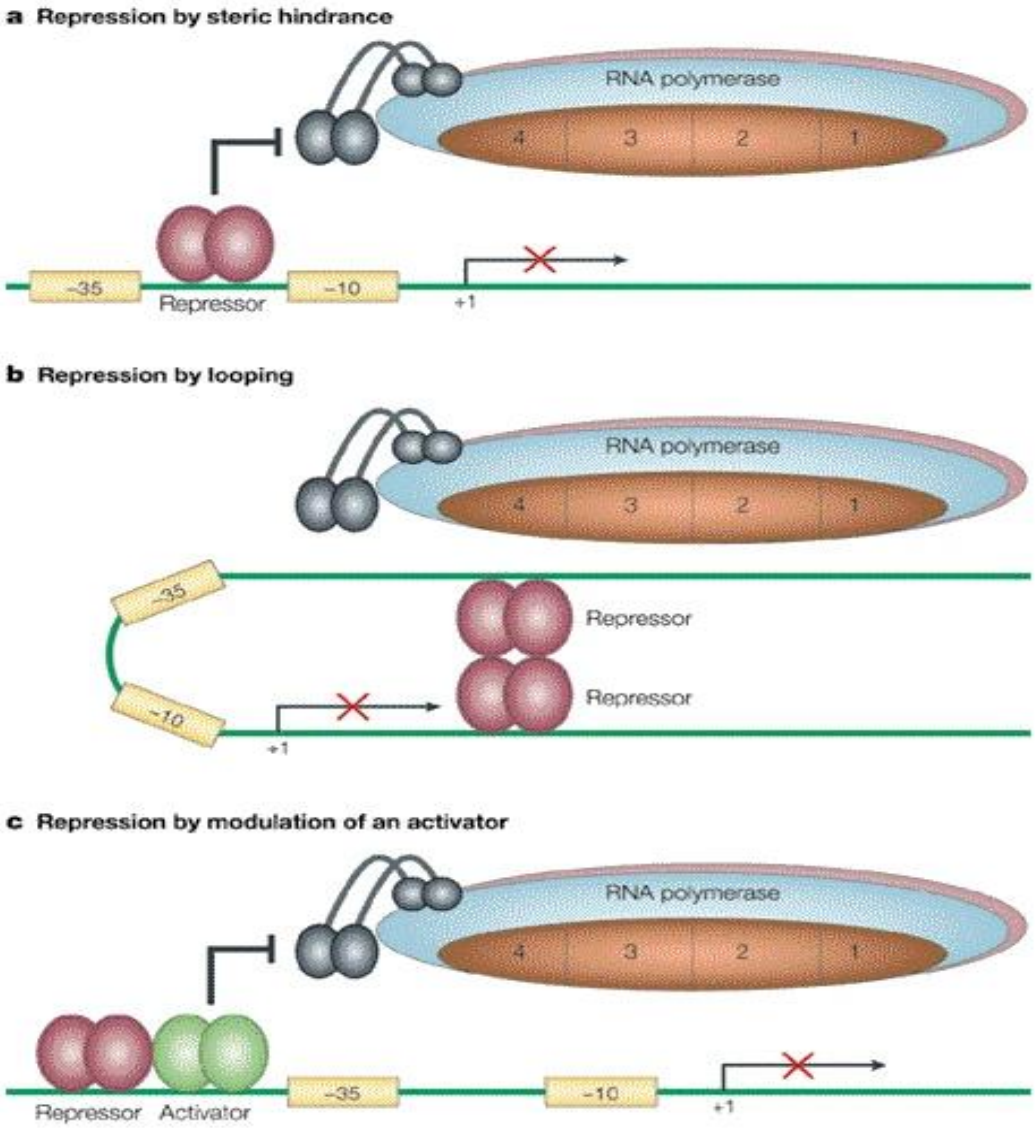


Figure 1.9: Simple promoter repression mechanisms

This panel shows three ways of inhibition of transcription initiation by a repressor. **a**. shows steric hindrance where the bound repressor (red dimer) overlaps with key promoter elements. **b**. shows looping of DNA by the interaction of bound repressor. **c**. shows the modulation of an activator by the repressor interacting with the activator (green), thereby stopping its function. Figure taken from Browning and Busby (2004).

1.8 Regulation of promoters by modifications

One mechanism whereby promoter activity can be regulated is promoter modification, involving covalent changes or modifications to specific bases in a gene regulatory region. These changes can modify the affinity of TF binding, or modify the efficiency of RNAP binding, to the promoter region. One of the most common modifications is DNA methylation which can result in global changes to the transcription pattern, due both to activation and repression at target promoters. This has been seen with the regulation of *agn43* expression in *E. coli*. In this case, the TF, OxyR, binds to GATC sequences at the *agn43* promoter site and blocks RNAP access. Following DNA replication, the DNA adenine methylase enzyme (Dam) will methylate the adenine in the GATC motif, which prevents OxyR from binding, and triggers *agn43* promoter activity (Wallecha *et al.*, 2002).

At some bacterial regulatory regions, an invertible cassette mediates an ON-OFF switch by toggling the ordination of a promoter. Hence, when the promoter is oriented to allow RNAP to bind and transcribe the target gene, it is in the ON phase. However, in the OFF phase, the inversion enzymes, together with other regulatory proteins, invert the promoter and prevent transcription of the target gene. A prime example of this is the regulation of *E. coli fim.* Here, the *fim* recombinases (*fimE* and *fimB),* with assistance from regulatory proteins, IHF and H-NS, invert the *fimA* promoter and orient it away from *fimA* expression (Blumer *et al.*, 2005).

Variation of the promoter base sequence can also influence promoter activity or regulation. Here, random insertion or a deletion in one base pair dinucleotides repeat can modulate promoter activity (Browning and Busby, 2016). For example, the *Haemophilus influenzae hifA* and *hifB* promoters, that drive expression of fimbrial components, contain a variable number of AT repeats between the -10 and -35 hexamer elements. Cell-to cell differences in the number of repeats results in variation of gene expression levels due to differences in activity between the different possible promoter sequences (Moxon *et al.*, 2006; Browning and Busby, 2016).

1.9 Regulation of transcription by small molecules

Bacteria use signals and effectors to communicate environmental changes to the machinery of transcription and translation. These changes, which can be either a limitation in nutrition or a change in growth conditions, often involve small molecules that regulate essential processes like stress

responses, nutrient acquisition, and other important adaptation (Mejía-Almonte *et al.*, 2020). Good examples of these small molecules are guanosine pentaphosphate or tetraphosphate (p)ppGpp and cyclic adenosine monophosphate (cAMP) (Kalia *et al.*, 2013).

(p)ppGpp is an important *E. coli* second messenger that reflects the shortage of amino acids in the cell. Hence, (p)ppGpp is involved in the regulation of genes involved in translation, amino acid biosynthesis and nucleotide metabolism (Potrykus and Cashel, 2008; Gallant and Cashel, 1971). When the cell faces amino acid starvation, deacylated tRNAs accumulate and enter the ribosomal acceptor site (A-site) leading to an activation of RelA (ribosomal-associated protein) and synthesis of ppGpp from ATP and GTP/GDP. High levels of ppGpp trigger the expression of stress genes and lower the expression of genes that are responsible for translation, growth, and cell division. This is due to the action of ppGpp in altering the stability of the transcription bubble in conjunction with DksA; a crucial component for rRNA regulation, at certain promoters. Hence, (p)ppGpp/DksA can activate or inhibit of transcription, depending on the promoter (Barker *et al*., 2001; Paul *et al*., 2005). DksA does not bind to DNA, it rather binds to the RNAP secondary (NTP) channel, therefore, reducing the lifetime of competitor-resistant complexes and the ppGpp concentration required for rRNA transcription inhibition (Perederina *et al.,* 2004; Paul *et al.*, 2004)

cAMP is one of the most important second messengers, signalling carbon starvation. During carbon catabolite repression, the EIIA ^{Glu} domain in the glucose transporter is the main factor to induce the cAMP synthesis. The transfer of carbohydrates in *E. coli* are dependent on the phosphotransferase system (PTS). This system is a multi-compnent system that can be encoded separately or fused together in the bacterial genome. The PTS mainly consists of three distinct proteins: histidine protein (HPr), enzyme I (EI) and enzyme II (EII) (Figure 1.10) (Görke and Stulke, 2008). The role of the PTS is to transfer the carbohydrate substrates *e.g.,* glucose, across the cytoplasmic membrane with their simultaneous phosphorylation.

The regulation of carbon catabolite repression is dependent on the phosphorylation state of the EIIA^{Glu} unit. When the glucose is available in the media, the phosphate group in EIIA^{Glu} is consumed by glucose (Görke and Stulke, 2008) (Figure 1.10). This process starts when the glucose enters the cell via the EIIBC membrane bound complex by phosphorylating the glucose using the phosphate group from phosphorylated EIIA^{Glu} and converting the glucose to glucose-6-phosphate. Then EIIA^{Glu} accepts the phosphate group from phosphor-enol pyruvate which is converted to pyruvate after losing

the phosphate group. Then, glucose-6-phosphate loses phosphate group and enters the glycolysis cycle, and the phosphate group is transferred to enzyme I then to histidine protein (HPr) and then which eventually transferred to $EIIA^{Glu}$ to start the cycle again (Görke and Stulke, 2008).

However, in the absence of glucose in the bacterial environments, the phosphorylation of $EIIA^{Glu}$ for a long period activates adenylate cyclase (AC); a membrane bound enzyme; by binding the C-terminal domain of the enzyme itself and with the help of factor "X" (unknown factor) (Figure 1.10). The activation of adenylate cyclase leads to induction of the synthesis of cyclic adenosine monophosphate (cAMP) from ATP. The high level of cAMP then binds to its receptor protein to trigger the formation of the cAMP-CRP complex. This changes the conformation of CRP, so it becomes competent to bind to target sites and regulate the activity of target promoters. Thus cAMP-CRP binds and can activate the transcription of catabolic genes promoters to overcome glucose limitation in the surroundings (Görke and Stulke, 2008).

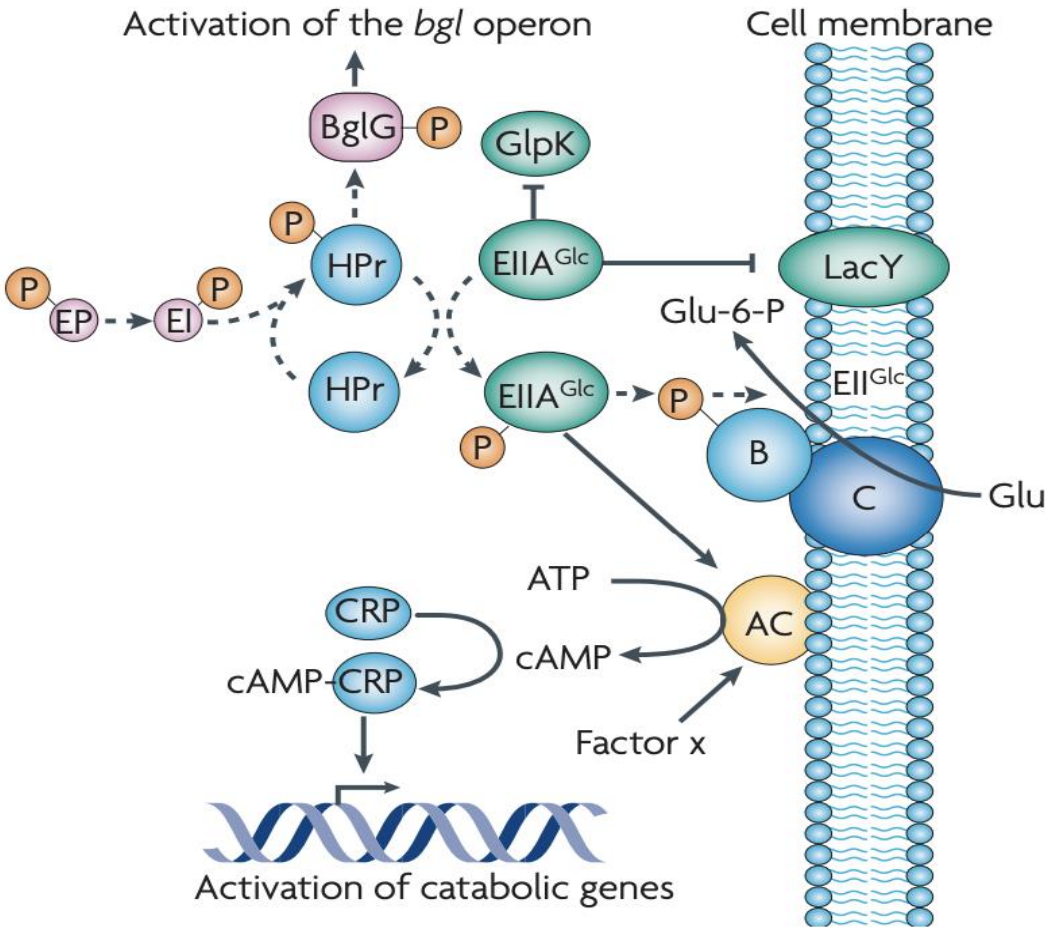


Figure 1.10: Regulation of Carbon catabolite repression in *E. coli*

The Figure illustrates the phosphotransferase system (PTS) in *E. coli*. Glucose transfers through the cell inner membrane using the EIIBC complex phosphorylation in phosphotransferase system (PTS). EIIBC uses the phosphate group from phosphorylated $EIIA^{Glu}$ to convert glucose to glucose-6-phosphate. Then $EIIA^{Glu}$ gains the phosphate group from histidine protein (HPr) and enzyme I (EI). The Histidine protein (HPr) phosphorylates the transcriptional antiterminator protein (BglG) that regulates the expression of the β-glucoside utilization genes. High levels of phosphorylated $EIIA^{Glu}$ will activate the carbon catabolite repression strategy. Phosphorylated $EIIA^{Glu}$ binds to adenylate cyclase (AC) with the help of an unknown factor (factor X). This binding stimulates the synthesis of cAMP from ATP which binds to its protein receptor (CRP) and induces the expression of catabolite genes. This Figure was taken from Görke and Stulke (2008).

1.10 The cyclic AMP receptor protein and its helpers: their role in health and sickness

1.10.1 Cyclic AMP receptor protein (CRP)

As described above, CRP is a global transcription factor involved in carbon source consumption. CRP also controls the expression of many genes that encode others transcription factors (Shimada *et al*., 2011). Among the 300 CRP targets genes, around 70 encode transcription factors. The molecular mass of CRP is 45 kDa and it is a dimer of two identical subunits, with each subunit containing two domains: The N-terminal domain (residues 1-133) that interacts with the allosteric effector cAMP and is responsible for dimerization, and C-terminal domain (residues 139-209) and that carries the determinants for DNA binding (Busby and Ebright, 1999; Won *et al.*, 2009). CRP recognises and binds to a 22 bp DNA sequence, composed of two inverted motifs (consensus 5`TGTGA`3), which are recognised by the helix-turn-helix of each C-terminal domain separated by six bp. Hence, the consensus DNA site for CRP is 5`-**TGTGA**TCTAGA**TCACA-**3` (Busby and Ebright, 1999; Zheng *et al.*, 2004).

Most CRP- dependent promoters fall into three classes, depending on the location of the DNA binding target and how CRP interacts with RNAP. At Class I promoters, CRP binds upstream of the RNA-binding promoter region and interacts with a target in RNAP αCTD via activation region 1 (AR1) (Busby and Ebright, 1999; Browning, and Busby, 2004). At Class II promoters, two distinct determents in CRP, AR1 and AR2, are required. Here, CRP binds to a target that overlaps the promoter -35 element and makes multiple interactions: AR1 in the upstream subunit of the CRP dimer binds αCTD, AR2 in the downstream subunit of the CRP dimer binds αNTD, and, in some cases, AR3 binds Domain 4 of the RNAP holoenzyme σ subunit (Busby and Ebright, 1999; Browning, and Busby, 2004). Class III activation occurs when two or more CRP dimers bind to separate DNA sites at the promoter region. Activation of transcription at these promoters involves both Class I and Class II interactions simultaneously (Busby and Ebright, 1999; Browning, and Busby, 2004).

1.10.2 Factor for inversion stimulation (Fis)

Fis is considered primarily as a nucleoid-associated protein (NAP) due to its ability to bend DNA and facilitate synapse formation, but it also has direct effects on the activity of many promoters in *E. coli.* Bradley *et al.* (2007) showed that Fis regulates 231 genes either directly or indirectly. Additionally,

in EAEC strain 042, at the *pet* gene promoter, Fis, together with CRP, is an essential co-activator (Rossiter *et al*., 2011). Fis regulation of gene expression is triggered by changes in its level in response to changes in growth rate due to the nutritional environment (Mallik *et al.*, 2006; Azam *et al.*, 1999).

1.10.3 Other bespoke transcription factors

Some strains of pathogenic *E. coli*, for instance EAEC and EHEC, have master transcription regulators that are dedicated to their virulence genes. In the case of EAEC, AggR is the specific virulence regulator that regulates dispersin, AAF and some plasmid encoded toxins (Morin *et al*., 2013). AggR is encoded on the EAEC pAA2 plasmid and is a member of the greater AraC family of transcription factors (Morin *et al*., 2013). In EHEC, regulation of the virulence genes is under control of Ler (LEE encoded regulator) and GrIA (global regulator of Ler activation), and both are encoded in the LEE chromosome: pathogenicity island (Yerushalmi *et al*., 2008).

1.10.4 The involvement of CRP in virulence gene regulation

Pathogens require physiological traits and virulence determinant expression to adapt to the changes in their environment, and, also, to promote infection, they need to resist the host immune defences. Therefore, pathogens have evolved different ways to cope, by exploiting signals, using effectors, and evolving regulatory circuits. The global transcriptional protein, CRP, has been reported to be involved directly or indirectly in such regulation. Here, I discuss some examples in species other than *E. coli.*

In *Yersinia pestis*, the knockout of the *crp* gene affects the expression of the virulence factor plasminogen activator (Pla), which is crucial for causing pneumonic plague and bubonic in infected cells (Kim *et al.*, 2007). Moreover, the ability of the *Y. pestis crp* deleted strain to cause infection by subcutaneous inoculation is compromised (Zhan *et al.*, 2008). Many attenuated *Y. pestis* strains contain lower levels of CRP and this is thought to be a factor in helping host evasion of pneumonic plague (Sun *et al.,* 2010). Type three secretion systems (TTSSs) in *Y. enterocolitica* involved in its CRP regulon include the flagellar type III secretion systems, YSA, and the YSC system that exports virulence proteins (Petersen and Young, 2002)

In *Salmonella enterica,* CRP indirectly post-transcriptionally impacts the level of the master regulator of invasion, HilD. This is due to CRP repression of Spot 42 sRNA transcription. This small RNA has a positive effect on HilD expression, and, hence, downregulation of Spot42 transcription leads to a

negative effect on HilD expression (El Mouali *et al.*, 2018). The regulator of *std*, chaperone-usher fimbriae, in *S. enterica* has been reported by Dufresne and Daigle (2021) to be significantly affected in *crp* mutants. CRP was also reported as a regulator of virulence genes, hlyE and taiA, in *S. Typhi* which are encoded in a pathogenicity island, SPI-18 (Jofré *et al*., 2014).

CRP is also involved in the expression of virulence genes in *Klebsiella pneumoniae*. A study carried out by Xue *et al.* (2016) revealed a binding motif for CRP in the promoter region of *allS* and the expression of this virulence gene was decreased in the absence of CRP. In *Aggregatibacter actinomycetemcomitans*, the lower expression of leukotoxin RNA levels in the absence of *crp* indicates that CRP is involved in the transcription of this virulence factor. However, as there is no CRP-binding site consensus at the leukotoxin promoter, Feuerbacher and his colleague (2011) suggested that either CRP binds to a non-consensus sequence or it acts indirectly by regulation of another transcription factor. In previous study on *Mycobacterium tuberculosis* suggestes that deleting *crp* gene in this pathogen caused significant diminishing the growth *in vitro* and lower the expression of virulence genes in model mouse (Rickman *et al.*, 2005).

Müller *et al.* (2009) suggest that cAMP-CRP controls the expression of type 1 fimbriae during the exponential phase of uropathogenic *E. coli* growth, and this control decreases when the pathogen reaches stationary phase, leading to type 1 fimbriae expression being coupled to environmental conditions. CRP is also involved in the modulation of multidrug resistance genes in *E. coli.* Nishino *et al.* (2008) report that CRP repressed the expression of the MdtEF multidrug efflux pump, introducing the connection between sugar metabolism and multidrug efflux pumps. In *Vibrio cholerae*, an investigation by Manneh-Roussel *et al.* (2018) proved that CRP is required for *V. cholerae* gene expressions in addition to the *rtxBDE* operon in order to respond to the intestinal environment; therefore, CRP is an important part of the *V. cholerae* regulatory network for lifestyle switching.

1.11 Hypotheses and aims of this project

My thesis focusses on one of the major causes of intestinal and UTI illness, caused by pathogenic EAEC and UPEC, respectively. The overall aim of my project is to discover other virulence determinants that are under CRP regulation beside the previous findings by Rossiter *et al*. (2011). My study follows up on a previous observation that the expression of certain of virulence genes is regulated by the global regulator, CRP. The research here focusses on molecular characterization of promoters that are regulated by CRP in two strains, EAEC strain 042 and UPEC strain CFT073. Also, this project aims to study and to identify any new virulence factor genes that could be involved in the CRP regulon and to understand the behaviour of CRP in EAEC strain 042.

The work on this thesis has been divided into sub-projects:

1. Investigation of the *pic* gene promoter region of in Enteroaggregative *E. coli* and Uropathogenic *E. coli* (Chapter Three).
2. Identification of CRP binding sites across the EAEC 042 genome using chromatin immunoprecipitation (ChIP) (Chapter Four).
3. Investigation of some EAEC 042-specific genes targeted by CRP *in vitro* and *in vivo* (Chapter Five).
4. CRP regulation at the EAEC strain 042 *kpsMII* promoter region (Chapter Six).

Within each of these sub-projects, there are other questions that arise, such as the mechanism of CRP action at target promoters, and whether CRP activates or represses. This leads to the issue of how to prevent and treat the EAEC infections, and whether, since it is highly conserved in Gram-negative bacteria, CRP might be engineered and used a ‘drug’.

Chapter 2

Materials and Methods

2.1 Bacterial media

Luria-Bertani agar (LB, Sigma) agar was prepared by dissolving 35 g of product in 1L of deionised H_2O (dH_2O), and autoclaving. Luria-Bertani broth (LB broth, Sigma) was made by dissolving 20 g of product in 1L of dH_2O, and autoclaving. 20% (*w/v*) Arabinose (Sigma) stock was prepared by dissolving 10g of product in 50 ml of dH2O and filtered. MacConkey agar (BD Difco) was made by dissolving 50 g of product in 1L of dH_2O, and autoclaving. Minimal salt agar was prepared by dissolving 20 ml of autoclaved 5x M9 salts (Sigma), 2ml filtered 1M glucose (Fisher Chemical), 200 µl of autoclaved 1M $MgSo_4$ (Sigma), 100 µl 100 µM $CaCl_2$ (Fisher Scientific UK), in 100 ml of autoclaved technical agar (Oxoid). Dulbecco's modified Eagle medium (DMEM) (Sigma-Aldrich).

2.2 *E. coli* strains and plasmids, and growth conditions

For overnight cultures, a single bacterial colony from a streaked plate or transformation was inoculated into 5 ml of LB, with appropriate antibiotics (if needed). Then the bacterial culture was incubated at 37˚C with shaking, for overnight or for a given time, depending on the experiment. All *E. coli* strains and plasmids are shown in Table 2.1 and Table 2.2.

2.3 Antibiotic stock preparation

0.04 mg of tetracycline (Sigma) was added to 4 ml of 50% ethanol and stored at -20˚C. 200 mg of Ampicillin (Sigma) was added to 2 ml of dH_2O, filtered, and then stored at 4˚C. Final concentration of antibiotics in 100ml of LB broth was 15 μgml^{-1} for tetracycline, 100 μgml^{-1} for Ampicillin.

2.4 Chemicals, buffers, and reagent preparations

100 mM $CaCl_2$ was prepared by dissolving 1.1098 g of $CaCl_2$ in H_2O, making up to 100 ml, then autoclaving and storing at 4℃.

100 mM $CaCl_2$ + 15% (*v/v*) Glycerol was prepared by adding 18.9 g of glycerol (Fisher Chemical) into 100 ml of 100mM $CaCl_2$, followed by autoclaving and storage at 4℃.

Z buffer was prepared by dissolving the following chemicals in dH_2O: 0.75 g of KCl (BDH), 0.246 g of $MgSo_4 \cdot 7H_2O$ (Sigma), 8.53 g of Na_2PO_4 (Fisher Chemical), and 4.866 g of $NaH_2PO_4 \cdot 2H_2O$ (Fisher Chemical), making up to 1L and autoclaving. This buffer was stored at room temperature. On the assay day, 270 µl β-mercaptoethanol (Sigma) was added into 100 ml of Z buffer. Then, 20 ml of

the prepared Z buffer were used to dissolved 80 mg of *o*-nitrophenyl-β -D-galactopyranose (ONPG) (Sigma) and used as reagent for ß-galactosidase assays.

1% Sodium Deoxycholate (Sigma-Aldrich) was prepared by dissolved 0.25 g sodium deoxycholate in 25 ml of dH_2O.

1 M Na_2CO_3 (Fisher Chemical) was prepared by dissolving 52.995g of Na_2CO_3 in H_2O and making up to 500 ml.

100%Toluene (Fisher Chemical).

Ethanol stock (Fisher Chemical) 100 % (*v/v*) Ethanol, 70 % (*v/v*) Ethanol**.**

7.5% Acrylamide was prepared by mixing 125 ml 30% (*w/v*) acrylamide (National Diagnostics), 100 ml 5x TBE, 94.5 g glycerol, and 200 ml H_2O, with storage at 4℃.

10% Ammonium persulphate (APS) (Sigma) was prepared by dissolved 1g of APS in 10 ml of dH_2O.

5x TBE buffer (National Diagnostics) was prepared by dissolved 54 g of Tris base, 27.5 g of boric acid, and 20 ml of 0.5M EDTA (pH 8.0) in dH_2O, then made up to 1L.

Phenol/chloroform/isoamyl alcohol mixture pH 8 (Sigma).

N, N, N`, N`-Tetramethylethylenediamine (TEMED) (Sigma).

Agarose gel 0.8g of agarose (Bioline) was dissolved in 100 ml of 0.5X TBE.

Ethidium Bromide (Bio-Red).

3M Sodium acetate. was prepared by dissolving 408 g of Sodium acetate in 800 ml of dH_2O.

1 M $MgCl_2$ (BDH) was prepared by dissolving 203.3 g of $MgCl_2$ $6H_2O$ in 800 ml of H_2O.

2.5M Glycine was prepared by adding 18.79g of glycine (Fisher Bioreagents) into 100ml of dH_2O, autoclaved.

FA-2 buffer was prepared by adding 50mM HEPES-KOH (pH7.5), 150mM NaCl, 1mM EDTA, 1% Triton-x 100, and 0.1% Sodium Deoxycholate into dH_2O, made up to 500 ml.

FA-1 lysis buffer was prepared by adding 1 protease inhibitor cocktail tablet (Roche) per 50 ml of FA-2 buffer. Then store on ice until use.

FA-3 was prepared by adding 20mM Tris (pH8.0), 250mM LiCl, 0.5% IGEPAL CA-630, 0.5% Sodium Deoxycholate, and 1mM EDTA into dH_2O then made up to 500 ml.

TE buffer pH8 was prepared by adding 10 ml of 1M Tris (pH8), 1 ml of 0.5M EDTA (pH8).

TBS was prepared by adding 6 g of Tris and 8.76 of NaCl into 800 ml of dH2O, pH adject to 7.5 with 1M HCl, then make it up to 1 L with dH2O.

250 nM cAMP was provided from Dr. Doug Browning.

ChIP Elution buffer was prepared by adding 50mM Tris (pH7.5), 10mM EDTA and 1% SDS into 100 ml of dH_2O.

G-50 Sephadex beads 5g of Sephadex G-50 (Sigma) were washed 3 times with 50 ml dH_2O, with centrifugation at 4000 xg inbetween washes to collect the beads. Finally, beads were resuspended in 100 ml of 1x TBS buffer.

T4 polynucleotide kinase. Supplied by NEB, used with supplied 10x buffer.

10x Transcription buffer (TNSC buffer): 400 mM Tris acetate pH 7.9, 10 mM $MgCl_2$, 1 M KCl, 10 mM DTT.

STOP solution: 97.5 % (*v/v*) deionised formamide, 10 mM EDTA, 0.3 % (*v/v*) Bromophenol Blue/ Xylene Cyanol FF.

Denaturing PAGE 8% of denaturing PAGE was prepared following the manufacturer's instructions from Promega.

α-^{32} P labelled UTP (3000 Ci/mmol). Supplied by Perkin- Elmer.

G+A ladder was provided by Dr. James Haycocks.

DynaMag^{TM}-2 Magnet (Thermo Fisher Scientific).

Agencourt AMPure XP magnetic SPRI beads (Beckman Coulter).

Purified anti-*E. coli* CRP antibody (monoclonal anti-CRP, BioLegend).

***E. coli* RNA Polymerase, Holoenzyme** (New England BioLabs (NEB)).

Gel extraction kit (Qiagen).

PCR purification kit (Qiagen).

Table 2.1: *E. coli* strains

Strains	Characterisations	Source/ Reference
EAEC 042	Wild type, prototype strain, Sm^R, Tet^R, Cm^R, Diarrhoeagenic in volunteers, expresses AAF/II, biofilm positive, harbours pAA	(Nataro *et al.*, 1995)
EAEC 042 DFB	Wild type prototype strain that is Sm^R, Tet^S Cm^S, biofilm positive, but with no plasmid (pAA)	Douglas Browning
EAEC 042 DFB *Δcrp*	Wild type prototype strain that is Sm^R, Tet^S Cm^S, biofilm positive, but with no plasmid (pAA). The *crp* gene has been deleted.	Douglas Browning
EAEC DFB 042 *Δfis*	Wild type prototype strain that is Sm^R, Tet^S Cm^S, biofilm positive, pAAs lack and *fis* gene has been deleted.	Douglas Browning
E. coli M182	*Δ(codB-lacI)3, galK16, galE15(GalS), λ⁻, e14-, relA1, rpsL150(strR), spoT1*	(Casadaban and Cohen, 1980)
E. coli M182 *Δcrp*	*Δ(codB-lacI)3, galK16, galE15(GalS), λ⁻, e14-, relA1, rpsL150(strR), spoT1 and Δcrp*	(Busby *et al*., 1983)
E. coli BW25113	$lacI^q$, *rrnBT14,* $\Delta lacZ_{WJ16,}$ *hsdR514,* $\Delta araBAD_{AH33,}$ $\Delta rhaBAD_{LD78}$	(Baba *et al*., 2006)
E. coli BW25113 *Δcrp*	$lacI^q$, rrnBT14, $\Delta lacZ_{WJ16,}$ *hsdR514*, $\Delta araBAD_{AH33,}$ $\Delta rhaBAD_{LD78.}$ The *crp* gene has been deleted.	(Baba *et al.*, 2006)
E. coli S17`1	Gram negative bacteria use for conjugation purposes, *pro, res⁻ hsdR17 (rK⁻ mK⁺) recA⁻*, *RP4-2-Tc::Mu-Km::Tn7*, Tp^r	(Simon *et al.,* 1983)
E. coli BW25113 *Δfis*	$lacI^q$, *rrnBT14,* $\Delta lacZ_{WJ16,}$ *hsdR514,* $\Delta araBAD_{AH33,}$ $\Delta rhaBAD_{LD78.}$ The *fis* gene has been deleted.	(Baba *et al*., 2006)
UPEC CFT073	Wild type UPEC strain	Welch et al., 2002)

Table 2.2: Plasmids used in this study

Plasmids	Description	Source
pRW50	Low-copy- number *lacZ* expression vector *oriV*, *lacZYA*, tet^R	(Lodge *et al.*, 1992)
pRW224	A pRW50 derivative which allows promoter fragments to be cloned using EcoRI and HindIII sites as fusions to *lacZ* transcription.	(Islam *et al.*, 2011)
pRW225	pRW224 derivative which allows cloning of promoter fragments using EcoRI and HindIII sites as translational fusions to *lacZ*.	(Islam *et al*., 2011)
pSR	Supercoiled small plasmid with 2.6 kb with EcoRI and HindIII sites for inserting the desired promoter fragment which will be located upstream of a λ*oop* terminator site. Amp^R	Kolb *et al.* (1995)
pDCRP	pBR322 derivative which encodes the *crp* gene	(West *et al*., 1993)
pDU9	pDCRP derivative with the *crp* gene deleted	(Bell *et al*., 1990)
pDCRP AR1-AR2-	pDCRP derivative with inactivation of AR1 and AR2.	(Thiyagarajan, Unpublished)
pDCRP AR1-	pDCRP derivative with inactivation of active region 1 ($AR1^-$) of CRP	(Thiyagarajan, Unpublished)
pDCRP AR2-	pDCRP derivative with inactivation of active region 2 ($AR2^-$) of CRP	(Thiyagarajan, Unpublished)
pRW224/*pic* p042 WT	A derivative of pRW224 with *pic042* from EAEC 042	(Chaudhuri *et al*., 2010)

pRW224/*pic* p042 DKO	A derivative of pRW224 carrying *pic042* with the downstream promoter knockout.	Rita Godfrey
pRW224/*pic* p042 UKO	A derivative of pRW224 carrying *pic042* with the upstream promoter knockout.	This study
pRW224/*pic* p042 BKO	A derivative of pRW224 carrying *pic042* with both promoter knockouts.	This study
pRW224/*pic* p073 WT	A derivative of pRW224 with *pic p073* from UPEC CFT073	Douglas Browning
pRW224/*pic* p073 DKO	A derivative of pRW224 carrying *pic073* with the downstream promoter knockout.	Rita Godfrey
pRW224/*pic* p073 UKO	A derivative of pRW224 carrying *pic073* with the upstream promoter knockout.	This study
pRW224/*pic* p073 BKO	A derivative of pRW224 carrying *pic073* with both promoter knockouts.	This study
pJET1.2	Cloning vector, Amp^R	CloneJET PCR Cloning Kit
pJET1.2/ *pic* p04 TSS	TSS of *pic*042 cloned into pJET1.2	This study
pJET1.2/ *pic* p073 TSS	TSS of *pic073* cloned into pJET1.2	This study
pRW224/*pic p042 U* WT	A derivative of pRW224 carrying *pic042* with part of the sequence of *pic p073*.	This study
pRW224/*pic p042 U* UKO	A derivative of pRW224 carrying *pic042* with the upstream promoter knockout and with part of the sequence of *pic p073*.	This study
pRW224/*pic p042 U* DKO	A derivative of pRW224 carrying *pic042* with the downstream promoter	This study

	knockout and with part of the sequence of *pic p073*.	
pRW224/*pic p042 U* BKO	A derivative of pRW224 carrying *pic042* with both promoter knockouts and with part of the sequence of *pic p073.*	This study
pRW224/*pic p042 28* WT	A derivative of pRW224 carrying *pic042* with a 28bp deletion in *pic p042* sequence.	This study
pRW224/*pic p042 28* UKO	A derivative of pRW224 carrying *pic042* with the upstream promoter knockout and with a 28bp deletion in *pic p042* sequence.	This study
pRW224/*pic p042 28* DKO	A derivative of pRW224 carrying *pic042* with the downstream promoter knockout and with a 28bp deletion in *pic p042* sequence.	This study
pRW224/*pic p042 28* BKO	A derivative of pRW224 carrying *pic042* with both promoter knockouts and with a 28bp deletions in *pic p042* sequence.	This study
pRW224/ *pic p042* 120 WT	A derivative of pRW224 carrying *pic042* with a consensus CRP site.	This study
pRW224/ *pic p042* 120 UKO	A derivative of pRW224 carrying *pic042* with the upstream promoter knockout and with a consensus CRP site.	This study
pRW224/ *pic p042* 120 DKO	A derivative of pRW224 carrying *pic042* with the downstream promoter knockout and with a consensus CRP site.	This study

pRW224/ *pic p042* 120 BKO	A derivative of pRW224 carrying *pic042* with both promoter knockouts and with CRP consensus.	This study
pRW225/ *pic p042* WT	A derivative of pRW225 carrying the *pic p042* fragment with 12bp of *pic* gene ORF.	This study
pRW225/ *pic p042* UKO	A derivative of pRW225 carrying the *pic p042* fragment with 12bp of *pic* gene ORF and with the upstream promoter knockout.	This study
pRW225/ *pic p042* DKO	A derivative of pRW225 carrying the *pic p042* fragment with 12bp of *pic* gene ORF and with the downstream promoter knockout.	This study
pRW225/ *pic p042* BKO	A derivative of pRW225 carrying the *pic p042* fragment with 12bp of *pic* gene ORF and with both promoter knockouts.	This study
pRW225/ *pic p073* WT	A derivative of pRW225 carrying the *pic p073* fragment with 12bp of *pic* gene ORF.	This study
pRW225/ *pic p073* UKO	A derivative of pRW225 carrying the *pic p073* fragment with 12bp of *pic* gene ORF and with the upstream promoter knockout.	This study
pRW225/ *pic p073* DKO	A derivative of pRW225 carrying the *pic p073* fragment with 12bp of *pic* gene ORF and with the downstream promoter knockout.	This study
pRW225/ *pic p073* BKO	A derivative of pRW225 carrying the *pic p073* fragment with 12bp of *pic*	This study

	gene ORF and with both promoter knockouts.	
pBAD/AggR	A derivative of pBAD30 carrying the aggR gene with its promoter.	(Sheikh *et al.*, 2002)
pRW50/cc41.5	A derivative of pRW50 carrying the CC (-41.5) promoter fragment.	(Savery *et al.*, 1995)
pRW50/*pkpsMII*	A derivative of pRW50 carrying the *pkpsMII* promoter fragment.	This study
pRW50/*p0225 VI*	A derivative of pRW50 carrying the *p0225 VI* promoter fragment.	This study
pRW224/*p0224 VI*	A derivative of pRW224 carrying the *p0225 VI* promoter fragment.	This study
pRW50/*kpsM* WT	A derivative of pRW50 carrying the *pkpsM* promoter fragment.	This study
pRW224/ *kpsM*216	A derivative of pRW224 carrying the *kpsM* promoter fragment.	This study
pRW224/ *kpsM*216R	A derivative of pRW224 carrying the *kpsM* promoter fragment reverse version of *kpsM*216.	This study
pRW50/*kpsM* 751	A derivative of pRW50 carrying a *kpsM* promoter fragment with a deletion of 100bp from *kpsM* WT promoter fragment.	This study
pRW50/*kpsM* 651	A derivative of pRW50 carrying a *kpsM* promoter fragment with a deletion of 200bp from the *kpsM* WT promoter fragment.	This study
pRW50/*kpsM* 551	A derivative of pRW50 carrying a *kpsM* promoter fragment with a deletion of 300bp from the *kpsM* WT promoter fragment.	This study

pRW50/*kpsM* 451	A derivative of pRW50 carrying a *kpsM* promoter fragment with a deletion of 400bp from the *kpsM* WT promoter fragment.	This study
pRW50/*kpsM* 351	A derivative of pRW50 carrying a *kpsM* promoter fragment with a deletion of 500bp from the *kpsM* WT promoter fragment.	This study
pRW224/*kpsM* 440	A derivative of pRW224 carrying a *kpsM* promoter fragment constructed from the *kpsM* WT promoter fragment.	This study
pRW224/*kpsM* p1KO	A derivative of pRW224 carrying a *kpsM* promoter fragment with promoter 1 knocked out, constructed from the *kpsM* WT promoter fragment.	This study
pRW224/*kpsM* p2KO	A derivative of pRW224 carrying *kpsM* promoter fragment with promoter 2 knocked out, constructed from the *kpsM* WT promoter fragment.	This study
pRW224/*kpsM* p3KO	A derivative of pRW224 carrying *kpsM* promoter fragment with both promoters knocked out, constructed from the *kpsM* WT promoter fragment.	This study
pSR/ *pkpsM* 440	A derivative of pSR carrying the *kpsM* 440 promoter fragment located upstream of a λ*oop* terminator site.	This study

2.5 Growth curves using growth kinetics in the Fluostar microplates

Overnight bacterial cultures of the test strain and their parent strain were set up, a day before the day of the assay. For the growth kinetics assay, 90µl of an appropriate media with an appropriate antibiotic was added to each well of a 96-well plate. Biological and technical replicates of tested strains were included in the assay. The overnights of the test cultures were diluted by adding 10 µl to 1 ml sterile LB broth. These suspensions were inoculated by adding 10µl to each well of the plate, including an empty well as the media control. The plate was covered with a BreathEasy membrane to prevent evaporation of liquid into the machine. The plate was placed in a Fluostar and the machine was absorbance measurement set to OD_{600} nm, number of cycles as 24, temperature at 37˚C and with shaking. The data was analysed using MARS analysis and Excel software.

2.6 Crystal violet biofilm formation using microtiter plates

Overnight bacterial cultures were grown from single colonies of test strains. Each strain was diluted 1:100 into 5 ml of high glucose medium (Dulbecco's modified Eagle medium (DMEM)) and incubated for 1 hour with shaking at 37˚C. DMEM was used here due to its high concentration of amino acids such as arginine, cysteine, glutamine, histidine, methionine, glycine, phenylalanine, and tyrosine. All these amino acids play an important role in regulating biofilm formation. Each well of a microtiter plate was inoculated with 150 µl of each culture. The microtiter plate was covered with a BreathEasy membrane to prevent evaporation and incubated at 37℃ overnight. After 16-17 hours of incubation, the media was removed by smacking the plate on absorbent papers upside down until no more media appeared on the paper towel. After that, 150 µl of 0.1% crystal violet was added to each well and the plate was incubated for 30 min at 4℃. The crystal violet was removed by smacking the plate again on absorbent paper upside down and the wells were washed with dH_2O three times until no dye found in the negative control. Then the plate was dried by smacking the plate on absorbent papers upside down and 150 µl ethanol/acetone solution was added to each well (mixture of 40 ml ethanol and 10 ml acetone) and was incubated the plate for 30 min at room temperature. The Labsystem Multiskan MS plate reader was used to measure absorbance at OD_{595} nm as a measure of biofilm formation.

2.7 Bacterial conjugation procedures

The transformant *Escherichia coli* S17`1 (with tested plasmids) was grown on a selective LB agar plate overnight at 37 °C and EAEC DFB 042 was grown on LB agar overnight. EAEC DFB 042 was mixed with *E. coli* S17`1 on the trans-conjugant agar plates. Control plates (EAEC DFB 042 and *E. coli* S17`1; separately) were prepared parallel to the conjugation process and all plates were incubated at 37 °C for 5 hours. When the incubation was complete, the conjugated EAEC DFB 042 and the controls (EAEC DFB 042 and *E. coli* S17`1) were streaked onto minimal salt agar with appropriate antibiotics. All plates were incubated overnight at 37 ˚C.

2.8 *E. coli* transformation with DNA plasmid DNA

2.8.1 Competent cells preparation using 0.1 M calcium chloride

1 ml of overnight bacterial culture was used to inculcate 50 ml LB media, which was incubated at 37˚C until the culture reached mid-exponential phase of growth (OD_{600} of 0.3-0.5). After that the culture was transferred into a sterile centrifuge tube and incubated on ice for 10 minutes. Then, the culture was harvested by centrifugation for 15 minutes at 4000 rpm at 4˚C in a pre-cooled centrifuge. The supernatant was removed, and the pellet re-suspended in 25 ml ice cold 0.1 M $CaCl_2$ and incubated on ice for 20 minutes before being harvested by centrifugation for 15 min at 4000 rpm at 4˚C. The pellet was re-suspended in 3.3 ml ice cold 0.1 M CaCl2 with 15% glycerol and incubated on ice for overnight before aliquoting into microfuge tubes and storage at -80˚C.

2.8.2 Plasmid transformation into competent cells

1 µl of plasmid is added to 100 µl of competent cells and incubated on ice for 45 minutes, including competent cells alone as a control for the experiment. Then, all transformations are heat-shocked at 42˚C for 2 minutes followed by incubation on ice for a further 5 minutes. 900 µl of LB is added to the cells and incubated at 37˚C for 1 hour with shaking. Then cells are harvested by centrifugation at 13,000 rpm for 1 min and re-suspended in 100 µl LB before being plated out using a sterile glass spreader onto appropriate agar plates with appropriate antibiotics. The plates were incubated at 37˚C overnight.

2.9 Plasmid DNA extraction method

2.9.1 Miniprep Kit for small scale of plasmid DNA preparation (Qiagen)

A bacterial overnight culture was centrifuged at 4000 rpm and 15°C for 15 minutes. Pelleted bacterial cells were suspended in 250 µl buffer P1, then transferred it to a microcentrifuge tube. 250 µl of Buffer P2 was added to the suspended cells and mixed by inverting 6 times until the solution becomes clear. Then, 350 µl of Buffer N3 was added and mixed immediately by inverting 6 times before being centrifuged it for 10 minutes at 13,000 rpm. 800 µl of the supernatant was applied to a QlAprep 2.0 spin column by pipetting, then the column was centrifuged at 13,000 rpm for 1 minutes and the flow-through discarded. The QlAprep 2.0 spin column was washed by adding 0.5 ml of Buffer PB, then centrifuged for 60 seconds, and the flow-through discarded. Then, the QlAprep 2.0 spin column was washed by adding 750 µl of Buffer PE, centrifuged for 1 minutes and the flow-through was discarded, and the column centrifuged again for 1 minute to remove residual wash buffer. After that, the QlAprep 2.0 spin column was transferred to a fresh microcentrifuge tube to collect the plasmid DNA. To do this, 50 µl of Buffer EB was added to the centre of the QlAprep 2.0 spin column, and, after standing for 1 minute, the column was centrifuged for 1 minute.

2.9.2 Maxiprep Kit for large scale of DNA plasmid (Qiagen)

A single colony from a fresh transformation or restreak plate was inoculated into 50 ml of LB medium with appropriate antibiotic. The following day, the culture was transferred into a 50 ml centrifuge tube and centrifuged at 4000 rpm for 15 minutes at 4°C to collect the cells. The pellet was resuspended in the P1 buffer provided in the kit and the plasmid extraction was continued following the manufacture' instructions.

2.10 B- galactosidase assays

Single colonies of the strains to be assayed were transferred into 5 ml LB in flasks with appropriate antibiotics. Three biological replicates were incubated overnight at 37°C, separately. The following day, subcultures were set up by adding 100 µl of overnight culture into 5 ml of fresh LB with appropriate antibiotics. Each culture was and incubated at 37°C until the OD_{650} was between 0.4-0.6. The subcultures then were then lysed by adding three drops of 1 % sodium deoxycholate and three drops of toluene (last step were done in the hood) and vortexed for 5-10 seconds. After 20 minutes incubation in the hood, 100 µl of lysed cells was added into 2 ml aliquots of Z buffer solution (100

µl of Z buffer with 270 µl of β- Mercaptoethanol) that had been equilibrated in a water bath at 30˚C. The reaction was started by adding 500 µl *o*-nitrophenyl-β -D-galactopyranoside (ONPG) in Z buffer and stopped with 1000 µl of 1M Na_2CO_3 when sufficient yellow colour had developed. The OD_{420} measurements of the final solutions were used to measure the amount of *o*-nitrophenol and recorded for further calculations. β -galactosidase activity was determined by using the following formula:

$$\beta - \text{galactosidase activity (Miller Units)} = \frac{1000 \times 2.5 \times A \times OD420}{4.5 \times T \times V \times OD650} \text{nmol/minutes/mg}$$

Where:

1000/4.5= Conversion factor of OD_{420} into nmol Ortho-nitrophenyl-b-D- galactopyranoside (ONPG), based on the assumption that 1 nmol/ml ONPG absorption of OD_{420} is equivalent to 0.0045.

2.5= Conversion factor of OD_{650} into bacterial mass(mg) based on the assumption that OD_{650} of 1 is equivalent to 0.4 mg/ml bacterial dry weight.

T=incubation time(minutes).

A=final assay volume (3.6 ml).

V= lysed culture volume (0.1 ml)

The β -galactosidase activities were measured using three independent transformants that carried the recombinant plasmids with desired promoter fragments. Therefore, β–galactosidase levels were calculated with standard deviation for each promoter fragment. Negative controls were included in all assays with empty vector and titled with 'No Insert'.

2.11 Construction of promoter fragments and cloning into plasmids

2.11.1 Polymerase chain reaction (PCR)

Q5 DNA polymerase (New England BioLabs (NEB)) was mainly used in the construction of desired products, due to its high fidelity. A typical PCR thermocycle method is outlined in Table 2.3. DNA template for the PCR experiment was prepared either from bacterial lysates or purified plasmids. Isolated DNA from bacterial cells came from boiling a single colony of EAEC strain 042 at 100˚C, followed by centrifugation to then use the supernatant (bacterial lysate) which contains the DNA. The

primers used for PCR are listed in Table 2.4. All oligonucleotides (Table 2.4) were made by Merck Life science UK limited.

The typical PCR reaction volume was 50 µl using Q5 DNA polymerase. This includes 5 µl of DNA template, 10 µl of 10 µM Forward primer, 10 µl of 10 µM Reverse primer, 1 µl of 1mM dNTPs (Bioline), 0.5 µl of Q5 DNA polymerase, 10 µl of 5x Q5 buffer, 3.5 µl of dH_2O and 10 µl of Q5 enhancer. The samples were transferred to the PCR machine and set up the cycling conditions as detaild in Table 2.3. After that, PCR products were run on polyacrylamide gels, calibrated by 100 bp DNA ladders (New England BioLabs (NEB)).

2.11.2 Site-directed mutagenesis (SDM)

SDM was used to introduce selected point mutations into desired promoters using the megaprimers listed in Table 2.4. Accuzyme DNA polymerase was used in the two rounds of PCR. In the first round, Megaprimer and primer D19897 were used, then the reaction was checked by running 1 µl of the PCR product on a polyacrylamide gel. Then, if the desired product was detected, the product was purified for the second round of PCR where the PCR product from the first round was used as a reverse primer togather with the Forward primer, D10520. The DNA templates for *pic p042* and *pic p073* mutagenesis were provided by Rita Godfrey.

For designing nested deletion fragments, a plasmid carrying the WT full length target promoter was used as the template for PCR. Then, a set of forward primers, listed in Table 2.4, were used to construct the desired fragments with the same reverse primer. Forward and reverse primers were flanked with restriction sites for EcoRI (New England BioLabs (NEB)) and HindIII (New England BioLabs (NEB)) respectively. The DNA products from each PCR reactions were then digested with these restriction enzymes and ligated into appropriate plasmid vectors.

Table 2.3 PCR thermocycle conditions

Cycle Step	Temperature	Time	Cycles
Initial step	98˚C	30s	1
Denaturation	98˚C	10s	x35
Annealing	X˚C (Typically 55˚C)	30s	
Extension	72˚C	30s	
Final extension	72˚C	10 minutes	1
	4˚C	Hold	

Table 2.4 Oligonucleotide primer sequences used in this study

Oligo name	Sequence (5′-3′)	Use
D10520	CCCTGCGGTGCCCCTCAAG	Forward primer for upstream EcoRI site in pRW50/ pRW224/ pRW225. Used in sequencing and amplification of the inserts.
D19897	GGCGATTAAGTTGGGTAACG CCAGGG	Reverse primer for HindIII site downstream in pRW224/pRW225. Used in sequencing and amplification of the inserts.
D10527	GCAGGTCGTTGAACTGAGCCTGAA ATTCAG	Reverse primer for HindIII site downstream in pRW50. Used in sequencing and amplification of the inserts.
pic p042 MG	CATTAATGCAGTAACTCTATTTTCC	Megaprimer for knockout the upstream promoter in *pic p042*
pic p073 MG	CGCCCGGGTCATCACGTCTGGTTATC	Megaprimer for knockout the upstream promoter in *pic p073*
pic p042.225 R	GGGGGAAGCTTAACTTTATTCACTAT GGAT	Reverse primer for construction of the *pic p042* protein fusion fragment including HindIII site.
pic p073.225 R	GGGGGAAGCTTAACTTTATTCACTGC GGACTCTCCATGATGTTTAAGT	Reverse primer for construction of the *pic p073* protein fusion fragment including HindIII site.
pic p042 120 F	GGGGGAATTCCTCCAGGAAACCCGGT GTGATTCAGTTCACAAAAACAC	Forward primer for construction of the *pic p042* 120 fragment including EcoRI site.
pic p042 28 F	GGGGGAAGCTTATGATGTTTAAGTAC TAATGATAACCCG	Forward primer for construction of the *pic*

		p042 28 fragment including EcoRI site.
pic p042 U F	GGGGGAATTCGAAAAGTATTTCACTA GGCGGTTCAGTTCACAAAAACACA	Forward primer for construction of the *pic p042 U* fragment including EcoRI site.
SP1	CTGGCGAAAGGGGGATGTGCTGCAA	Specific primer for 5'RACE kit (SP1)
pJET1.2 forward sequencing primer	CGACTCACTATAGGGAGAGCGGC	Forward sequencing of insert colony PCR in pJET1.2
pJET1.2 reverse sequencing primer	AAGAACATCGATTTTCCATGGCAG	Reverse sequencing of insert colony PCR in pJET1.2
EC042_0224 F	GGGGGAATTCACGAACAACCCGTCGT TAATCCC	Forward primer for construction of the *EC042_0224* fragment including EcoRI site.
EC042_0224 R	GGGGGAAGCTTCCGTGTTCTGAAATT GACTAAA	Reverse primer for construction of the *EC042_0224* fragment including HindIII site.
KpsM 201 F	GGGGGGAATTCTAATTACCTTCGGGA TTATTGATG	Forward primer for construction of the *KpsM* 201 fragment including EcoRI site.
KpsM 201 R	GGGGGAAGCTTCTGTTTCAAACCAGA ACGCACGCG	Reverse primer for construction of the *KpsM* 201 fragment including HindIII site.
EC042_0414 F	GGGGGGAATTCATGTTGCAATCTTCT GCTGACAAAGC	Forward primer for construction of the *EC042_0414* fragment including EcoRI site.
EC042_0414 R	GGGGGAAGCTTTTAACTTATAATTAA GAGAAAAAAC	Reverse primer for construction of the *EC042_0414* fragment including HindIII site.
EC042_3143 F	GGGGGGAATTCATGTCCGTTTGCGGA CAAGCAATAG	Forward primer for construction of the *EC042_3143* fragment including EcoRI site.
EC042_3143 R	GGGGGAAGCTTGGGAAACCGGTGTTT TGAAAACAGT	Reverse primer for construction of the *EC042_3143* fragment including HindIII site.
EC042_3644 F	GGGGGGAATTCTTTAACATTAATGCC AAAAACCGGG	Forward primer for construction of the

		EC042_3644 fragment including EcoRI site.
EC042_3644 R	GGGGGAAGCTTTTGATAAACGTTCGACGCATAGTAA	Reverse primer for construction of the *EC042_3644* fragment including HindIII site.
EC042_0224 F	GGGGGGAATTCATGAGCAAAATGAACAACAATGGCG	Forward primer for construction of the *EC042_0224* fragment including EcoRI site.
EC042_0224 R	GGGGGAAGCTTCGGTTGATAGCAATGGCAGTTCGCT	Reverse primer for construction of the *EC042_0224* fragment including HindIII site.
EC042_4012 F	GGGGGGAATTCGGTGGAGATCTCTGTCACCAGCCAG	Forward primer for construction of the *EC042_4012* fragment including EcoRI site.
EC042_4012 R	GGGGGAAGCTTCGCAGCGTGTTCGCCGTCGAACCGT	Reverse primer for construction of the *EC042_4012* fragment including HindIII site.
EC042_3975 F	GGGGGGAATTCAGATTCGGTTTTTCAGACCCCCATC	Forward primer for construction of the *EC042_3975* fragment including EcoRI site.
EC042_3975 R	GGGGGAAGCTTATAAGAGTTATTCAAAATATTTTGT	Reverse primer for construction of the *EC042_3975* fragment including HindIII site.
EC042_3970 F	GGGGGGAATTCCAATAGCTAATTTTTTTGGCGCACC	Forward primer for construction of the *EC042_3970* fragment including EcoRI site.
EC042_3970 R	GGGGGAAGCTTATCGCTCACCAGAGTCGTACAGG	Reverse primer for construction of the *EC042_3970* fragment including HindIII site.
EC042_4604 F	GGGGGGAATTCAACTATTGAGGCCAGCCTGATTTTG	Forward primer for construction of the *EC042_4604* fragment including EcoRI site.
EC042_4604 R	GGGGGAAGCTTATCGCTCCCCTCTTTAACCATTTGA	Reverse primer for construction of the *EC042_4604* fragment including HindIII site.
virK F	GGGGGGAATTCTCATGTTTTCCGGCAATTGAGATAC	Forward primer for construction of the *virK*

		fragment including EcoRI site.
virK R	GGGGGAAGCTTTTTCGGTACTCAGAG CGTTTTTTAC	Reverse primer for construction of the *virK* fragment including HindIII site.
EC042_0225 F	GGGGGGAATTCAACTCGAATAAAGA AAAGGGTGTG	Forward primer for construction of the *EC042_0225* fragment including EcoRI site.
EC042_0225 R	GGGGGAAGCTTATGGGGTTGGCATTT ATG	Reverse primer for construction of the *EC042_0225* fragment including HindIII site.
KpsM527 F	GGGGGGAATTCATAACACCATTTAAA TGTGATAT	Forward primer for construction of the *KpsM 527* fragment including EcoRI site.
KpsM WT R	GGGGGAAGCTTACTTCTTGCCATTTG ATGATGTG	Reverse primer for construction of the *KpsM* WT fragment including HindIII site. Primer was used as the reverse primer for all nested deletion fragments.
KpsM WT F	GGGGGGAATTCTTATTAATAGTTGCA ATAAATCA	Forward primer for construction of the *KpsM* WT and *KpsM* 440 fragments including EcoRI site
KpsM751 F	GGGGGGAATTCAAAATTCCTGGAGAT AATCAGAAA	Forward primer for construction of the *KpsM* 751 fragment including EcoRI site
KpsM651 F	GGGGGGAATTCACTGAGGGATGGTGT TGGTTGTAA	Forward primer for constructing *KpsM* 651 fragment including EcoRI site
KpsM551 F	GGGGGGAATTCGATGCGACTTAAATA ACACCATTT	Forward primer for construction of the *KpsM* 551 fragment including EcoRI site
KpsM451 F	GGGGGGAATTCAAATAGGGAAATAG TTTCTCGGTG	Forward primer for construction of the *KpsM* 451 fragment including EcoRI site

KpsM351 F	GGGGGGAATTCGGGGTATTATAATCAAGTAGTTA	Forward primer for construction of the *KpsM* 351 fragment including EcoRI site
KpsM216 F	GGGGGAATTCTAACACTGGTTAAAATAAATA	Forward primer for construction of the *KpsM* 216 fragment including EcoRI site
KpsM216 R	GGGGGAAGCTTTGTTCACCGAGAAACTATTT	Reverse primer for construction of the *KpsM* 216 and *KpsM* 440 fragment including HindIII site.
KpsM216R F	GGGGGAATTCTGTTCACCGAGAAACTATTT	Forward primer for construction of the *KpsM* 216R fragment including EcoRI site
KpsM216R R	GGGGGAAGCTTTAACACTGGTTAAAATAAATA	Reverse primer for construction of the *KpsM* 216R fragment including HindIII site.
KpsM p1KO F	GGGGGGAATTCTTATTAATAGTTGCAATAAATCATTGAGTAACAATTGATAGGCCAAAACATATAGGATAATTCTTGTGTGATCTGT**G**TTTTGTGTAGC	Forward primer for construction of the *KpsM* p1KO and *KpsM* p3KO fragments including EcoRI site
KpsM p2KO R	GGGGGAAGCTTTGTTCACCGAGAAACTATTTCCCTATTTAAAATTCACTCGTGTACTTCTTATTTATATCTACAGCCCCCTCTTTACAGTCATATTTGTGATTTA**C**ATCACATTTA	Reverse primer for construction of the *KpsM* p2KO and *KpsM* p3KO fragments including HindIII site.

**Grey shaded bases represent the site for restriction enzymes EcoRI or HindIII.

2.12 Polyacrylamide gel electrophoresis

Polyacrylamide gels were used to check PCR products and purify the desired products. Each gel was prepared by adding 15 µl Tetramethylethylenediamine (TEMED) and 100 µl 10% Ammonium persulphate (APS) into 10 ml 7.5% acrylamide solution, which was then poured into vertical glass casting chambers. The chamber was placed into a vertical electrophoresis tank, and then the tank was filled with 1x TBE buffer. The PCR product and 100 bp DNA ladder were loaded using 6x loading dye, and run for around 35 min, at a constant current of 40 mA. Gels were stained with ethidium bromide before visualisation on a UV light box. The desired DNA fragments bands were cut out according to their size and the gel slices were placed into clean micro centrifuge tubes, for purification by DNA electroelution, phenol chloroform extraction, and ethanol precipitation.

2.13 Electroelution of DNA fragments from polyacrylamide gel

Slices of gel containing DNA fragments were placed in sterile dialysis tubing (Medicell International Ltd), which had been washed out with H_2O, and then clipped at one of the ends. Then 200 µl 0.1x TBE buffer was added into the tubing, and the other end was clipped. 0.1X TBE buffer was put in an electroelution tank, and then, the dialysis tubing was placed in the tank and electrophoresed for 20 min at 30 mA. The DNA solution was removed from the tubing into a clean micro centrifuge tube then the dialysis tube was washed with 200 µl with sterile distilled (SDW) water, which was added to the DNA solution already in the microfuge tube.

2.14 Phenol/chloroform extraction and ethanol precipitation of DNA fragments

After electroelution, samples containing DNA were made up to 400 µl with SDW, then 400 µl phenol/chloroform (1:1) was added and mixed by vortexing until the solution turns white. After centrifuging for 2 minutes at 13,500 rpm, the upper layers of the samples were taken into a clean micro centrifuge tubes and made up to 400 µl with SDW. For ethanol precipitation, 40 µl 3 M sodium acetate (pH 7.0), 4 µl 1 M $MgCl_2$ and 888 µl 100% ice cold ethanol (stored at -20℃) was added and incubated overnight at -20℃. The following day, the samples were centrifuged at 13,000 rpm for 15 minutes at 4℃. After removal of the supernatant, the pellet was resuspended in 1 ml 70% iced cold ethanol (stored at -20℃) and centrifuged at 13,000 rpm for 10 minutes at 4℃. Most of the supernatant was then decanted and the samples were dried using the DNA speed vac for around 10 minutes at medium heat. DNA pellets were resuspended in 20 µl SDW.

2.15 Restriction digestion of PCR fragments and plasmids

Plasmids, pRW50, pRW224 and pSR were cut using restriction enzymes to facilitate cloning of promoter fragments. before. Hence, 300 µl of plasmid minipreps DNA (pRW50 (9 µg), pRW224 (9.9 µg), and pSR (15 µg)) was digested by adding 10 µl EcoRI (200 $U\mu l^{-1}$), 10 µl HindIII (200 $U\mu l^{-1}$), 40 µl 10x CutSmart buffer, with the volume then made up to 400 µl with SDW. The mixture was incubated for 3 hours at 37℃ and then 4 µl calf intestinal alkaline phosphatase (CIP) was added to the mixture and incubated for a futher 1 hour at 37℃. The digested plasmid was purified using phenol/chloroform extraction and ethanol precipitation and then run on a 0.8 % agarose gel to estimate the amount of plasmid needed for DNA cloning. For digesting PCR products, 4 µl EcoRI, 4 µl HindIII, and 6 µl CutSmart buffer were added to 20 µl of PCR product, and the mixture was made up to 60 µl with SDW. The mixture was incubated for 3 hours at 37℃ then purified by phenol/chloroform extraction and ethanol precipitation. The amount of PCR product was estimated before cloning by running 1 µl of the PCR product on a small polyacrylamide gel.

2.16 Ligation and sequencing of segments of plasmid clones

The desired promoter was cloned into an appropriate vector by adding 1-2 µl cut plasmid, ,1 µl T4 DNA ligase (200 $U\mu l^{-1}$), and 2 µl 10x T4 DNA ligase buffer, with 8-15 µl restricted insert DNA, and the final volume made up to 20 µl with sterile dH_2O. The mixture was incubated overnight at 16℃. The following day, the mixture was transformed into 100 µl $CaCl_2$ competent cells. Cells were plated out onto MacConkey agar plates with selective antibiotics and incubated for overnight at 37℃. For sequencing analysis, single colonies from each ligation plate were inoculated into 5 ml 2x LB and incubated overnight with shaking at 37℃. The following day, plasmid DNA was prepared from the cultures and then 3 µl of 1 µM D10520 primer was added to 7 µl plasmid DNA and sequenced in the Functional Genomics and Proteomics Laboratory, University of Birmingham, UK.

2.17 Rapid amplification of cDNA ends (5` RACE)

Overnight cultures of BW25113 cells, containing either pRW224/ *pic p042* or pRW224/ *pic p073*, were grown in LB medium at 37 °C with shaking until OD_{600} = 1. Total RNA was isolated using an Isolate II RNA Mini Kit (Meridian Bioscience), as specified by the manufacturer. Specific mRNA was then converted to cDNA using primer SP1 and a 5`Race kit 2$^{nd}$ Generation (Roche), according to the manufacturer's instructions. Strand-specific cDNA was A-tailed on the 3` end and then amplified by PCR using primers dT-anchor and lacZRev. PCR products were cloned into the pJET1.2

vector using the Clone JET PCR Cloning Kit (Thermo Scientific), as specified by the manufacturers, and sequenced using Sanger Sequencing.

2.18 *In vitro* assays

2.18.1 Radiolabelled DNA fragments

The desired fragments were constructed with flanking EcoRI and HindIII restriction sites and amplified using PCR (primers listed in Table). This was done to generate promoter fragments that could be cloned into pSR plasmid for sequencing. After verifying the fragment sequence, DNA was produced using Qiagen plasmid miniprep kits following the manufacturer's instructions, DNA was prepared for radiolabelling by cutting the fragments with EcoRI and HindIII followed by treatment with Calf Intestinal Alkaline Phosphatase (CIP) and was purified by gel extraction. Then fragments were radiolabelled at both ends using T4 polynucleotide kinase (following manufacturer's guidelines) and [γ-^{32}P]-ATP by incubation of the reaction for 30 minutes at 37℃. Next, the fragment was purified to remove the excess [γ-^{32}P]-ATP using a 200 µl of G-50 Sephadex columns. Labelled fragments were then used in EMSA assays.

2.18.2 Electrophoretic mobility shift assay (EMSA)

Purification of CRP is described in Kumari *et al.* (2000) and CRP purified in this way was used in EMSA assays. The labelled fragments were incubated with various concentrations of purified CRP, depending on the purpose of the experiment. One µl of the labelled DNA fragment was added into the reaction. Reaction buffer was provided by Dr. Doug Browning (contained 10x HEPES, 2mM cAMP, 50% (*v/v*) glycerol, 25 µg ml^{-1} herring-sperm DNA, 0.5 mg ml^{-1} BSA in dH_2O). The final volume of the reaction was 10 µl. The reactions were then incubated at 37 °C for 30 minutes, and then run on 0.25 x TBE on a 6% polyacrylamide gel supplemented with 250 nM cAMP, with 0.25 x TBE as running buffer, and the gel was run for 3 hours at 166V. After electrophoresis, the gel was fixed with (10% *v/v*) methanol and (10% *v/v*) acetic acid for one hour. Next, the gel was dried using a vacuum drier. After that, the gel was exposed overnight to a phosphor-screen for analysed using a Bio-Rad Molecular Imager FX and Quantity One software (Bio-Rad).

2.18.3 Multi-round transcription assays

Desired fragments, cloned into pSR, were purified from bacterial hosts grown overnight using Qiagen Maxiprep kits, following the manufacturer's instructions. The recombinant plasmids were used for *in vitro* transcription reactions as described by Kolb *et al.* (1995). Reaction mixtures contained 40 mM Tris pH 7.9, 10 mM $MgCl_2$, 50 µM DTT, 0.1 M KCl, 2 µg ml^{-1} BSA, 5 mM ATP/GTP/CTP, and UTP and 5 µCi α-^{32}P- UTP, using 572 µg ml^{-1} pSR template. The reaction mixture was incubated at 37°C for 10 minutes. Then 2 µl of 400 nM holo-RNAP were added into 11 µl reaction mixture and incubated further for 10 minutes. The reaction mixture was then stopped using stop solution. Four µl of each reaction was analysed by denaturing PAGE (at 60 W for 2 hr). Then the gel was dried and exposed to a Biorad phosphorscreen. The gel image was analysed using Bio-Rad Molecular Imager FX and Quantity One software (Bio-Rad).

2.19 Chromatin immunoprecipitation followed by sequencing assay (ChIP-seq)

The method used for ChIP-seq has been modified from that described by Singh *et al.* (2014). Data analysis was carried out and the quality of samples was assessed using the Galaxy platform (Afgan *et al.*, 2018). Experiments were carried out in duplicate. EAEC 042 cells were sub-cultured from overnights and grown to exponential phase in 1/100 LB medium (OD_{600} between 0.3 and 0.6). The culture was fixed using formaldehyde at a final concentration of 1% (*v/v*) and incubated for 20 minutes then quenched with 10 ml of 2.5 M glycine. Cells were collected by centrifugation for 5 minutes at 4000 rpm at 4°C then re-suspended in 1x of the culture volume in 1 ml of 1x TBS. This step was repeated twice. After washing, the cells were resuspended in 1 ml of FA lysis buffer (containing 2 mg/ml lysozyme) and incubated for 30 minutes at 37°C. Then the cell lysates were chilled on ice for around 5 minutes before the sonication step using a Bioruptor sonicator. This was set for three rounds of 10-minutes cycles at 4°C (30s on, 30s off), and cells were kept on a water/ice slurry during the sonication. Then cell lysates were centrifuged at 13,000 rpm in a cooled benchtop centrifuge, and the supernatant was kept. The DNA concentration in clarified lysates was estimated using a nandrop, and 150 µg of chromatin was used per immunoprecipitation. Immunoprecipitations were made up to 700 µl using FA-1 buffer, and two immunoprecipitations were set up for each culture, one containing antibody as a 'treatment' sample, and the other as a mock. The remainder of the supernatant was stored at -80°C. For the immunoprecipitation step, 2 µl of anti-CRP antibody provided by Dr. Douglas Browning (1mg/ml) was added into treatment samples. For the the mock

samples, I added 2 µl of FA-1 buffer. All samples (the treated and mock samples) were incubated on a rotating wheel at 4°C overnight.

Magnetic beads (Dynabeads protein A and protein G (ThermoFisher Scientific)) with a magnetic rack were used to form and wash the immunocomplexes. Fifteen µl of each protein A and protein G magnetic beads were prepared for each sample (scaled to the number of samples being processed). Beads were washed with 1 ml FA-1 buffer, before resuspending in the equivalent of 100 µl of FA-1 buffer per samples. 100 µl of beads were added into each sample and incubated on a rotating wheel for 1 hour at 4°C. Then a series of washes were done, using FA buffers as follows; the beads were washed three times using 1 ml FA-1 buffer each time. For the last wash, 1 ml of FA-1 buffer was added, and beads were rotated for 5-10 minutes at room temperature. Then the beads were resuspended with 1 ml of FA-1 and transferred into a new tube and washed twice with 1 ml of FA-1 (total of six washes were done). One ml of FA-2 buffer was used to wash the beads then 1 ml of FA-3 buffer was used to wash the beads. Afterward, 1 ml of TE buffer was used to wash the beads. Then the beads were resuspended in 195 µl of ChIP Elution buffer and incubated for 10 minutes at 65°C. Next, 5 µl of Proteinase K (50mg/ml, Sigma) were added into samples. Afterwards, the reversal of crosslinks was done by placing the tubes at 42°C for one hour and then at 65°C for 4 hours. Next the DNA was purified using a PCR purification kit (Qiagen), following the manufacturer's instructions. Then the DNA was eluted with 50 µl of dH_2O. Next, DNA libraries were generated using NEBNext® Ultra II DNA library prep kit (New England BioLabs (NEB)), following the manufacturer's instructions. After the final purification of amplified libraries using Agencourt AMPure XP magnetic SPRI (Solid Phase Reversible Immobilisation) beads, the library quality was checked using the Tapestation instrument and quantified by qPCR using a NEBNext library quantification kit, following the manufacturer's instructions. Then samples were diluted to 20nM stocks, and pooled in an equimolar ratio, before being sent for sequencing. Samples were sent for sequencing on an Illumina HiSeq platform at the Genewiz Next Generation Sequencing facility.

2.20 Bioinformatic analysis of ChIP-seq data

Raw sequences were aligned to the EAEC 042 genome (chromosome genome reference FN554766.1; and pAA plasmid genome reference FN554767) using Bowtie selection in Galaxy software. Next, the quality of the reads was generated by studying the correlation among samples and the strength of signals of ChIP-seq reads (Figure 2.7). Here I used plot correlation to generate a heat map for studying

the correlation between the input and treated samples and the signal from the ChIP-seq enrichment. The parameters for further analysis were done following the instructions on the Galaxy website, under analysis of the ChIP-seq data tutorial. Peaks were called using MACS2 in Galaxy software, each replicate was assessed with the appropriate controls separately then peaks were called only if the peaks were found in both replicates. The thresholds indicate the total number of reads, and these were compared with control samples to discard any artefactual enrichment. I extracted DNA sequence from 50 bp upstream to 50 bp downstream of the peak centre and applied the sequence to MEME SUITE software (Bailey *et al*., 2009) to locate CRP binding sites and generate a CRP motif. The nearest annotated genes to peaks were compared to *E. coli* K-12 strain MG1655 using BLASTp to identify the common CRP targets or new targets. CRP targets in the EAEC 042 genome was viewed in the Artemis genome browser (Carver *et al.*, 2011). The EAEC 042 DNA plotter of chromosome and pAA plasmid were created using the Artemis genome browser. Statistical analyses, calculation of peak locations, and determination of CRP binding site positioning relative to genomic features were done manually in an Excel spreadsheet.

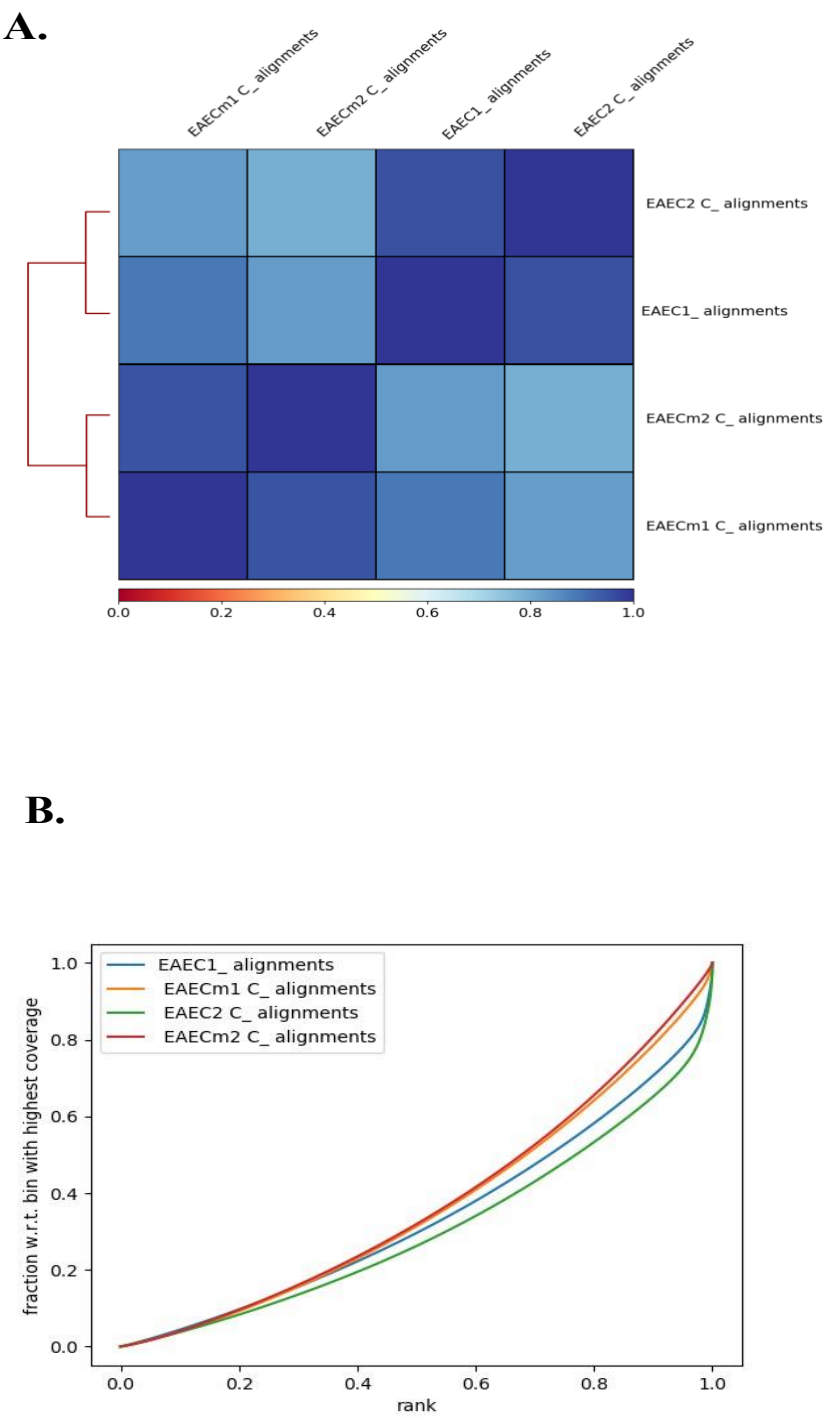


Figure 2.1: Assessing the quality of ChIP-seq samples

A. Heatmap shows the correlation between ChIP-seq replicates.

B. The strength of signals of ChIP-seq reads using Signal Extraction Scaling (SES).

Chapter 3

Investigation of the *pic* gene promoter region in Enteroaggregative and Uropathogenic *E. coli*

3.1 Introduction

Enteroaggregative *E. coli* and uropathogenic *E. coli* both express a "protein involved in colonisation" (Pic) in the early stage of infection of human hosts. Pic is a serine protease autotransporter of the *Enterobacteriaceae* (SPATE) and is secreted via a type V protein secretion system. This virulence determinant is an enzyme that targets mucus in the intestines and the bladder, hence the name mucinase. Degradation of the mucin component of mucus, by Pic, permits pathogens to attach to the host cells, hence, initiating the formation of biofilms. The *pic* gene is usually found on the pathogenic chromosome: in EAEC strain 042, *pic* is located at position 4,924,840 (accession no. FN554766.1), and in UPEC strain CFT073 it is at position 326,209 (accession no. NC_004431.1) (Navarro-Garcia *et al.*, 2010) (Figure 3.1).

Pic was originally identified as a virulence determinant in EAEC by Henderson *et al.* (1999). Comparison of the *pic* promoter region in EAEC strain 042 and UPEC strain CFT073 shows some differences in promoter DNA sequence. Although transcription of the genes encoding most virulence determinants in EAEC is controlled by the virulence transcription factor, AggR, the *pic* gene appears not to be part of the AggR regulon (Morin *et al.*, 2013). Pic is needed at the very first stage of infection to break down the host's first line of defence (the mucus layer), hence the regulation of *pic* needs a transcription factor that can manage to regulate in these circumstances.

Both Henderson *et al.* (1999) and Harrington *et al.* (2009) studied the *pic* gene in EAEC strain 042 (accession no. AF097644). These papers focused on the Pic protein itself, its secretion, and its transportation, and how Pic contributes to EAEC 042 virulence. Behrens *et al.* (2002) studied the transcriptional regulation of the *pic/set* locus in EAEC strain 042 and *Shigella flexneri* 2a strain 245T. They concluded that multiple promoters were located upstream of *pic* gene, and that expression of *pic* reached high levels in the exponential phase of bacterial growth, but there was no consideration of which regulator drives *pic* expression. In this chapter, I focus on the regulation of *pic* expression by the cyclic AMP receptor protein (CRP), and the differences in the *pic* regulatory region between EAEC strain 042 and UPEC strain CFT073.

A. EAEC 042

B. UPEC CFT073

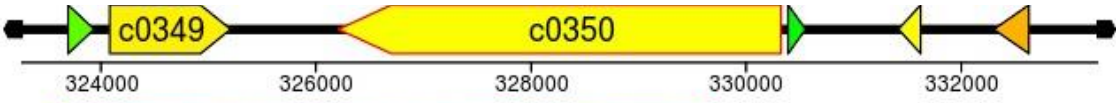


Figure 3.1: Position of *pic* gene in the EAEC and UPEC genomes

The panel shows the diagrammatic representation of the position of *pic* gene in (**A**) EAEC strain 042 chromosome (*pic)* (**B**) UPEC strain CFT073 chromosome (*c0350*). The target gene is outlined with red (Chaudhuri *et al.,* 2007).

In *E. coli* K-12, CRP is a global transcription factor as it regulates hundreds of promoters. It is usually considered that CRP is involved in regulating the use of alternative carbon sources, in the absence of glucose or other nutritional stresses. CRP can act as an activator or a repressor of genes expressions, depending on its binding position with respect to RNAP at target promoters (Grainger and Busby, 2008). CRP binds to specific DNA sequences after it becomes activated via cAMP (Busby and Ebright, 1999; Zheng *et al.*, 2004). The consensus DNA site for CRP on DNA contains two 5 base pairs elements separated by 6 base pairs: 5`-**TGTGA**$N_6$**TCACA**-3`, each bold sequence is recognised via the recognition helix on one subunit of the CRP homodimer. After binding, CRP often recruits RNAP to the target promoter and this positions the RNAP to initiate the transcription. This happens by interaction of domain 2 in sigma (σD2 in holo-RNAP) with promoter -10 hexamer element to open the transcription bubble to facilitate insertion of the template strand into the RNAP active site. Such promoters are classified as Class I or Class II for activation by CRP. In Class I activation, the CRP binding site is located 41 bp or more upstream from the promoter -10 hexamer element and the CRP-RNAP interaction requires activating region 1 (AR1) of CRP. In Class II activation, the CRP binding site is located 21 bp upstream of the promoter -10 hexamer element, and so bound CRP overlaps with the promoter -35 element. Class II activation requires both CRP activating region 1 (AR1) and activating region 2 (AR2) to interact with RNAP (Rhodius *et al*., 1997) (Figure 3.2).

To study the regulation of *pic* promoters, site-directed mutagenesis was used to find the functional regulatory elements at the *pic* promoter in both EAEC strain 042 and UPEC strain CFT073 (*pic p042* and *pic p073*). Surprisingly, I found that both promoters were dependent on CRP. In next two sections, I outlined the strategy I have used, involving two *lacZ*-based reporter vectors. The promoter-less *lacZ* expression vector, pRW224 was used to clone the *pic* promoters using EcoRI and HindIII sites, and then measure transcription (Figure 3.3). The other reporter vector was pRW225, which carries *lacZ* without an SD sequence and start codon. This was used to clone *pic* promoter fragments that included the *pic* gene translation start and this facilitated study of the efficiency of *pic* gene translation initiation (Figure 3.3). Comparisons between *pic p042* and *pic p073* were conducted in the following sections. Additionally, I studied other TFs, such as Fis (another global regulator) and AggR (master regulator for virulence determinants in EAEC strain 042) to assess if they played any role in *pic* regulation.

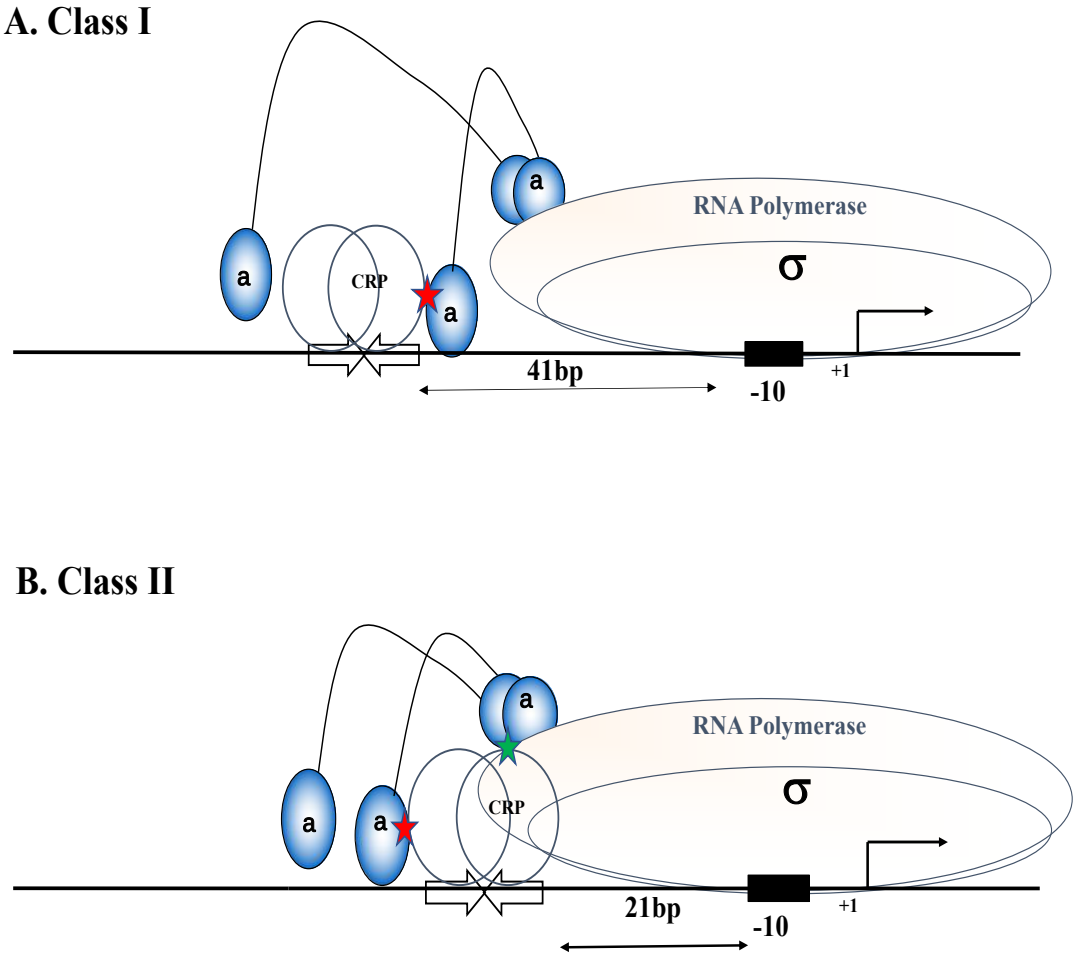


Figure 3.2: CRP activation Classes

The figure schematically shows the two CRP activation classes:

A. The figure shows a diagrammatic representation of Class I CRP activation. CRP, shown as light maroon subunits, interacts with alpha-CTD of RNAP (light blue domain) using active region 1 (red star). The -10-hexamer element of the target promoter is showed as a black box separated from the CRP binding site by 41 bp. +1 represent transcription start site.

B. The figure shows a diagrammatic representation of Class II CRP activation. CRP, shown as light maroon subunits interacts with alpha-CTD of RNAP (light blue domain) using CRP active region 1(red star), and interacts with alpha-NTD of RNAP (light blue domain) with CRP active region 2 (green star). The -10-hexamer element of the target promoter is shown as a black box separated from CRP binding site by 21 bp. +1 represent transcript start site.

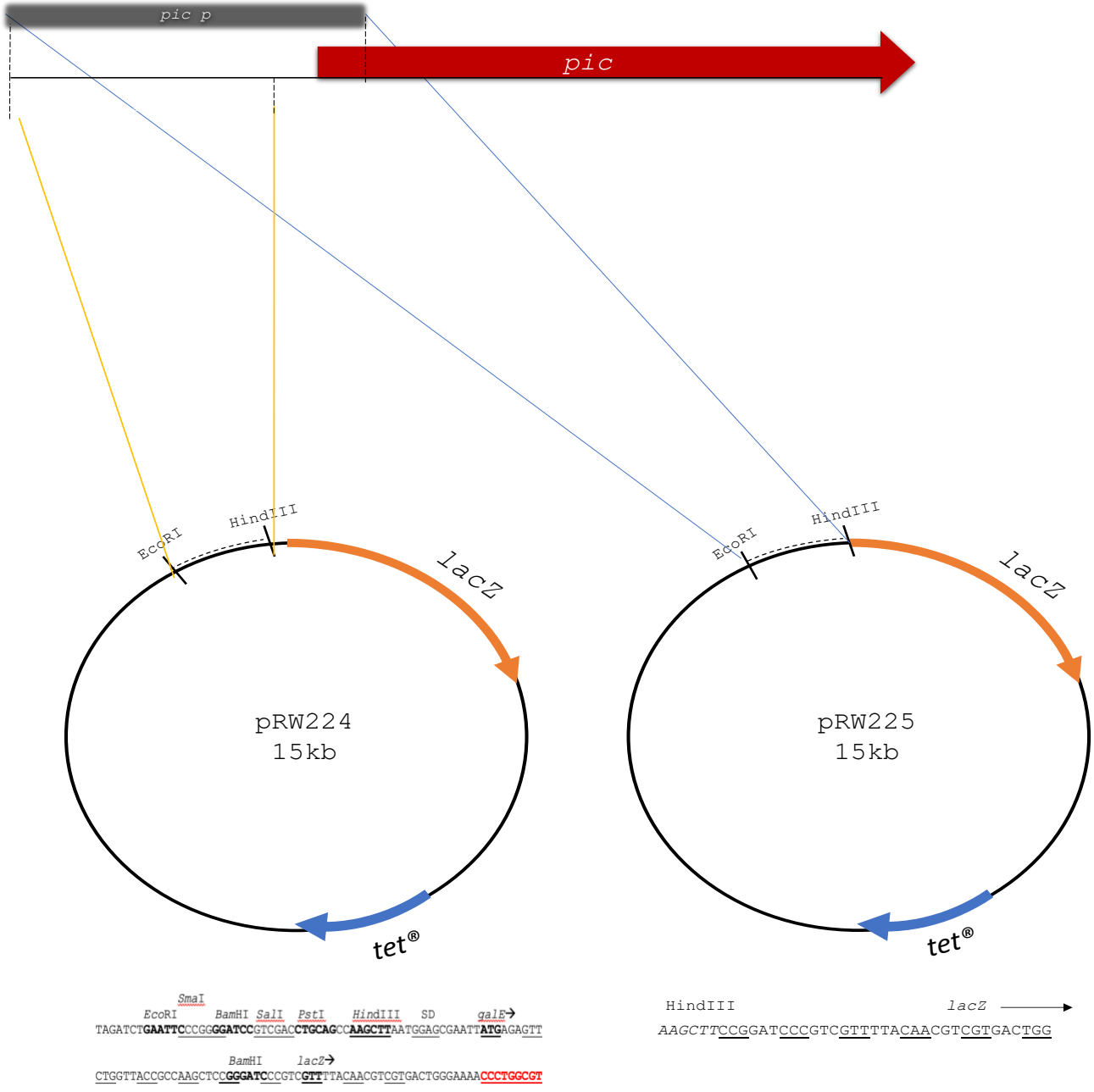


Figure 3.3: Cloning vectors for *pic* promoter fragments

pic promoter fragments were cloned into two plasmids to study their regulation. *lacZ* reporter, pRW224 (left), was used to construct operon fusions for studying and detecting *pic* promoter activity and regulation. *lacZ* reporter, pRW225 (right), was used to construct protein fusions for studying *pic* gene translation efficiency, as well as *pic* promoter- directed transcription. Lower part of figure shows the sequence of cloning site for each plasmids. The bases after HinDIII site until BamI site in pRW224 are missing bases in pRW225 cloning site (SD bases missed) .

3.2 The EAEC 042 *pic* regulatory region

3.2.1 Investigation of the *pic p042* promoter region and transcription regulation

To investigate the activity and regulation of *pic p042*, the 167 bp DNA sequence upstream of the *pic* gene was amplified by PCR, using DNA from EAEC strain 042, and primers listed in Table 2.4. The PCR product was cloned, using EcoRI and HindIII sites into the promoter-less *lacZ* expression vector, pRW224. This gave an opportunity to measure the promoter activity by measuring ß-galactosidase enzyme levels due to expression of *lacZ.* The recombinant plasmid, pRW224/ *pic p042* WT was used to investigate regulation by CRP, Fis and AggR to define functional -10 element, and for mapping transcript start sites. Figure 3.4 shows the base sequence of inserted promoter fragment.

To study the activity of promoters cloned in *lacZ* expression vector, *E. coli* K-12 Δ*lac* host strains were used. Thus, the recombinant plasmid, pRW224/ *pic p042* WT, was transformed into *E. coli* K-12 strains BW25113 Δ*lac* and its Δ*crp* derivative. The transformed cells were then grown in LB medium at 37˚C, with shaking to mid-exponential phase, with the OD_{650} between 0.4-0.6. Cells were lysed and measurements of ß-galactosidase levels were used as a proxy for promoter activity. The negative control, pRW224 with no insert, was included in the experiments.

Data in Figure 3.5 show an increase in ß-galactosidase level by 350 Miller units, in cells that carrying pRW224/ *pic p042* WT comparing with negative control with no insert, indicating the existence of a promoter in the cloned fragment. The level of ß-galactosidase expression was reduced by almost 4-fold in Δ*crp* cells, indicating a CRP-dependent promoter. A potential DNA site for CRP binding site is shown in Figure 3.4 starting at positioned 118 5`-**G**C**G**GTTCAGT**TCACA**-3`. A calculation of AT% of the bases downstream of CRP binding site (103 bp) showed a hight percentage of AT bases, around 68.93 AT%.

> *pic p042* WT

```
                          150
GAATTCGATCTGGCAGCCTGAGTTCACAGATAAAACAATCTCCAGGAAA

                    100
CCCGGGGCGGTTCAGTTCACAAAAACACATTAATGCAGTAACTATATTT

                50
TCCTTTCTGGTGATAACGTCGGGTTATCATTAGCTTCTTCAGCTATTTT

                1
ACTTTTATATCCCTTGTAAACATCATAAGCTT
```

Figure 3.4: Fragment carrying the EAEC 042 *pic* promoter DNA sequence

The figure shows the DNA sequence of *pic p042* promoter, 167 base pairs upstream from the of *pic* gene Shine Dalgarno sequence (see Figure 3.5B below). The DNA site for CRP is highlighted in light green. The sequences highlighted with dark green are the flanking EcoRI and HindIII restriction site. The numbering of bases in fragment starts from the HindIII site.

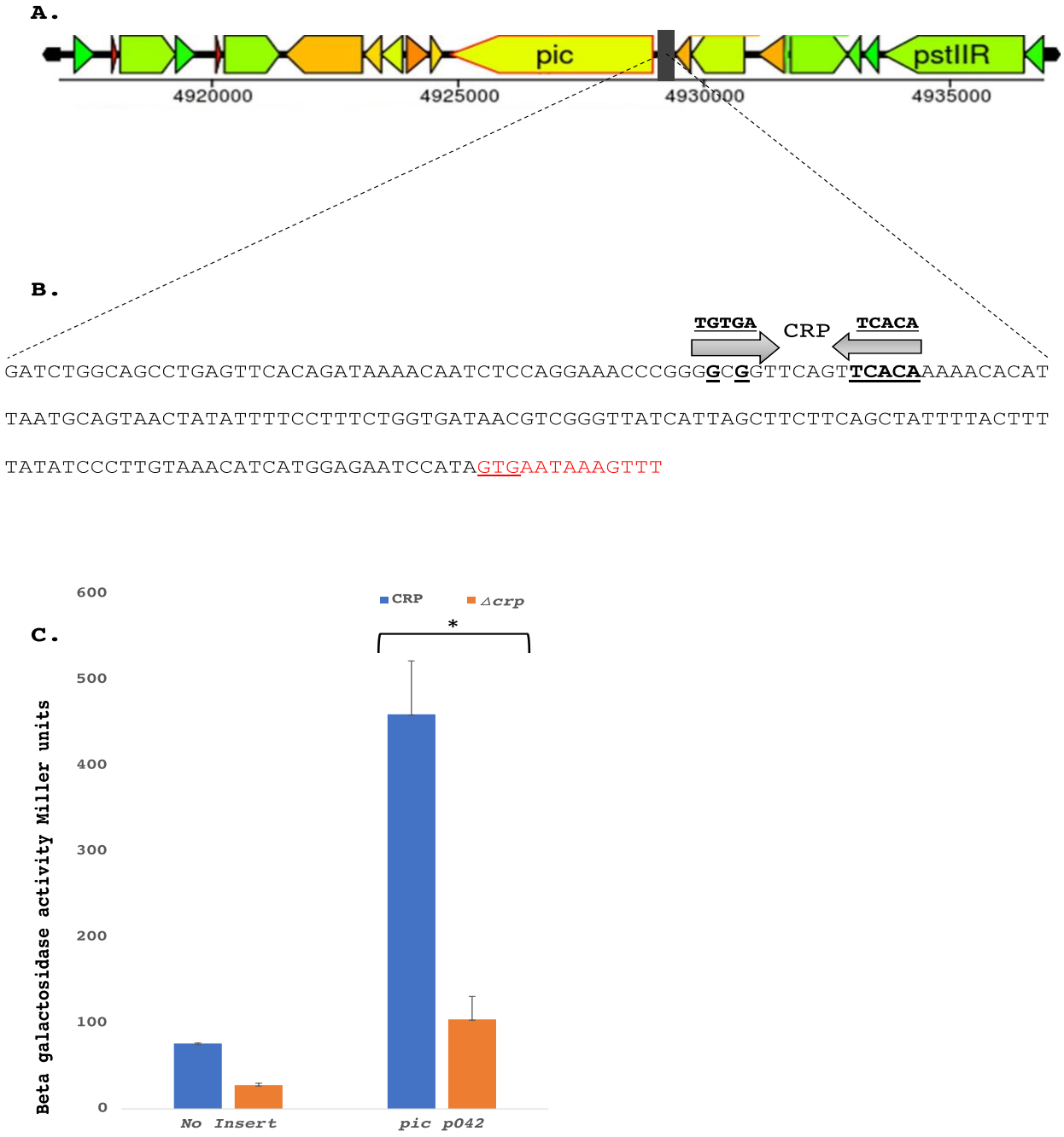


Figure 3.5: The *pic p042* promoter

A. The panel shows a diagrammatic representation of the position of the *pic* gene in the EAEC chromosome (strain 042) (Chaudhuri *et al.,* 2007).

B. The panel shows the DNA sequence of the *pic p042* promoter region. The proposed DNA site for CRP is identified by horizontal arrows with bold type denoting bases that correspond to the consensus. The *pic* gene coding region is shown in red with start codon underlined.

C. The bar chart illustrates measured ß- galactosidase levels in *E. coli* K-12 strain BW25113 WT *Δ lac* (blue bars) and *E. coli* K-12 strain BW25113 *Δ lac Δcrp* (orange bars) carrying pRW224 with the *pic p042* promoter. Each measurement is the average of three independent β-galactosidase assays and the error bars represent the standard deviation (STDEV) and * indicates $P < 0.0008$ using a student's *t*-test.

3.2.2 Identification of the transcript start site at the *pic p042* promoter

The location of the *pic p042* transcript start site (TSS) can help identify promoter -10 hexamer element. Hence, the 5` RACE (rapid amplification of cDNA ends) method was used to determine the first base of the transcript and plasmid pRW224/ *pic p042* WT was used here. Cells carrying this plasmid were grown in LB medium at 37˚C, with shaking to late-logarithmic phase (OD_{600}= 1.0), and total RNA was isolated from the lysate using a kit. Using thr 5`RACE kit, this RNA was copied into cDNA and then tailed with poly(A). After that, the tailed cDNA was amplified using anchor primer 5`-GACCACGCGTATCGATGTCGACTTTTTTTTTTTTTTTV-3` as forward primer and SP2 as reverse primer (listed in Table 2.4). The PCR product from this last step was then used in cloning using the CloneJET PCR cloning kit. Here, the product from the PCR step was ligated into pJET1.2 vector, then the ligation mixture was used directly for transformation into *E. coli* K-12 strain, M182. Transformants were then selected to isolate the recombinant plasmid pJET1.2/*pic p042* TSS and sequenced by reverse or forward primers, provided by the CloneJET PCR cloning Kit (the primers are listed in Table 2.4).

The result, illustrated in Figure 3.6, pinpoints the *pic p042* transcript start to position 50 in the sequence shown in Figure 3.4. Inspection of the base sequence upstream from the experimentally determined start point (Figure 3.6) suggests 5`-TAACGT-3` as the likely *pic p042* -10 hexamer element and 5`-AACTAT-3` as the likely -35 hexamer element. The *pic p042* promoter TSS is 62 bases upstream from *pic* gene started codon 5`-GTG-3`. Note that the suggested -10 hexamer element carries an upstream 5`-TG-3` motif and is located 41bp downstream of the proposed DNA site for CRP (see Figure 3.6).

>pic p042

```
                            -100                      -80                       -60
GATCTGGCAGCCTGAGTTCACAGATAAAACAATCTCCAGGAAACCCGGGGCGGTTCAGTTCA
                         -40       -35            -20          -10           +1
CAAAAACACATTAATGCAGTAACTATATTTTCCTTTCTGGTGATAACGTCGGGTTATCATTA
                    +20                       +40                      +60
GCTTCTTCAGCTATTTTACTTTTATATCCCTTGTAAACATCATGGAGAATCCATAGTGAATA

AAGTTT
```

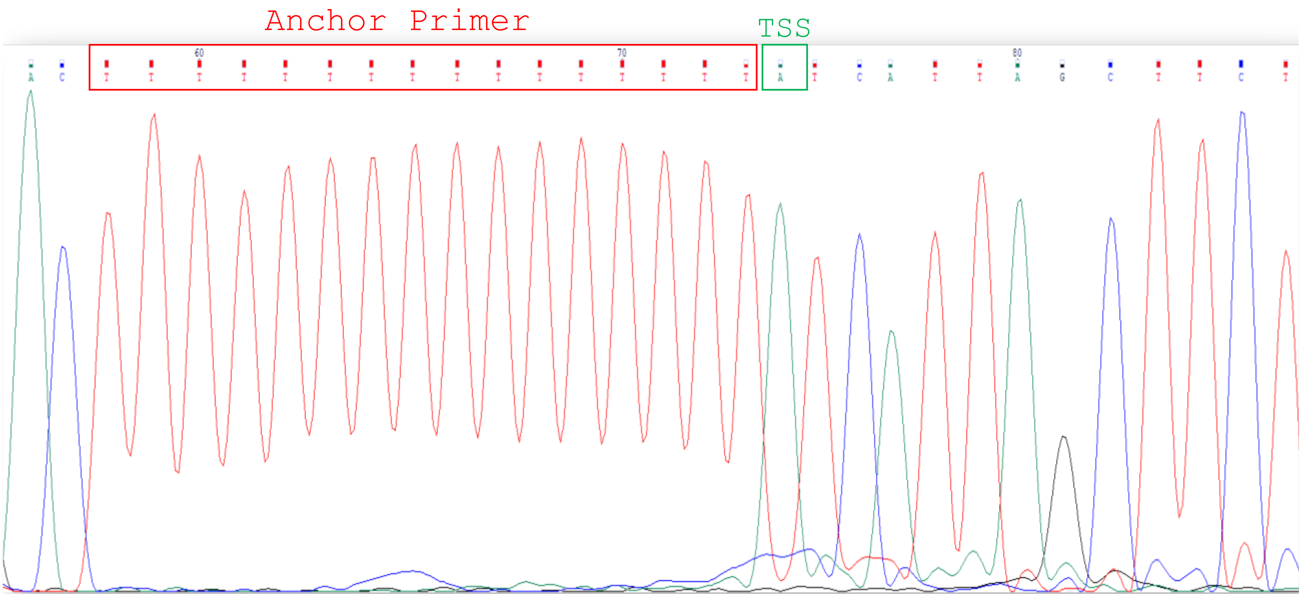


Figure 3.6: Mapping the *pic p042* promoter TSS

The figure shows the identification of the TSS at the *pic p042* promoter. Upper part shows the base sequence of *pic p042* promoter non-template strand. The arrow and +1 indicate the TSS with the corresponding base shaded bold green. The DNA site for CRP is highlighted in yellow. The likely corresponding promoter -10 and -35 hexamer elements are shown with rectangles. The *pic* gene coding sequence is shown in red typeface with the start codon in bold. The lower graph shows the 5`RACE sequencer readout for the *pic p042* TSS.

3.2.3 Identification of functional -10 hexamer element at the *pic p042* promoter

The purpose of this experiment was to locate functional -10 hexamer element and assess their location with respect to the proposed DNA site for CRP, using the 167 bp DNA fragment, *pic p042* WT, shown in Figure 3.4. As well as carrying the proposed extended -10 element 5`-TGATATAACGT-3`, I noted a second potential -10 hexamer 20 bp upstream 5`-TATATT-3`. Hence, site-directed mutagenesis was used to: (1) introduce a point of mutation in the upstream -10 element at base 2**A** 5`- T**A**TATT-3` (Upstream knockout, UKO), and (2) introduce double mutations at -10 element base **A** and at extended -10 element **G** 5`-T**G**AT**A**ACGT-3` (Downstream knockout, DKO), and (3) combine the mutations 5`-CT**A**TATTTTCCTTTCTGGT**G**AT**A**ACGT-3` (Both knockout, BKO). The primers are listed in Table 2.4, and alignments of the mutated constructs are shown in Figure 3.7. with schematic diagrams of the four constructs: WT, UKO, DKO, BKO, shown in Figure 3.8A.

Each of the three mutant constructs, UKO, DKO and BKO was cloned into the pRW224 *lac* expression vector and the resulting recombinant plasmids were transformed into *E. coli* strains M182 Δ*lac* and M182 Δ*lac* Δ*crp.* Transformed cells carrying pRW224/*pic p042* WT (positive control), pRW224/*pic p042* UKO, pRW224/*pic p042* DKO, pRW224/*pic p042* BKO, and empty vector pRW224 with no insert as a negative control, were grown and assayed to measure ß-galactosidase expression.

Figure 3.8B shows a strong reduction in promoter activity with the knockout of the downstream -10 element (DKO). This promoter activity was decreased more than 2-fold in the absence of the transcriptional factor CRP. In the case of *pic p042* BKO, the promoter activity here was recorded as the lowest activity with all previous promoter activities as the ß-galactosidase level shows only 158 Miller units and in Δ*crp* cells shows half of this expression. On the other hand, *pic p042* UKO promoter activity was less than the *pic p042* WT. These results indicate that *pic p042* is a CRP-activated promoter and both upstream and downstream play some role. However, this promoter prefers to use the downstream -10 element and thus is predominantly a Class I CRP-activated promoter, as predicated from the experimentally determined transcript start site.

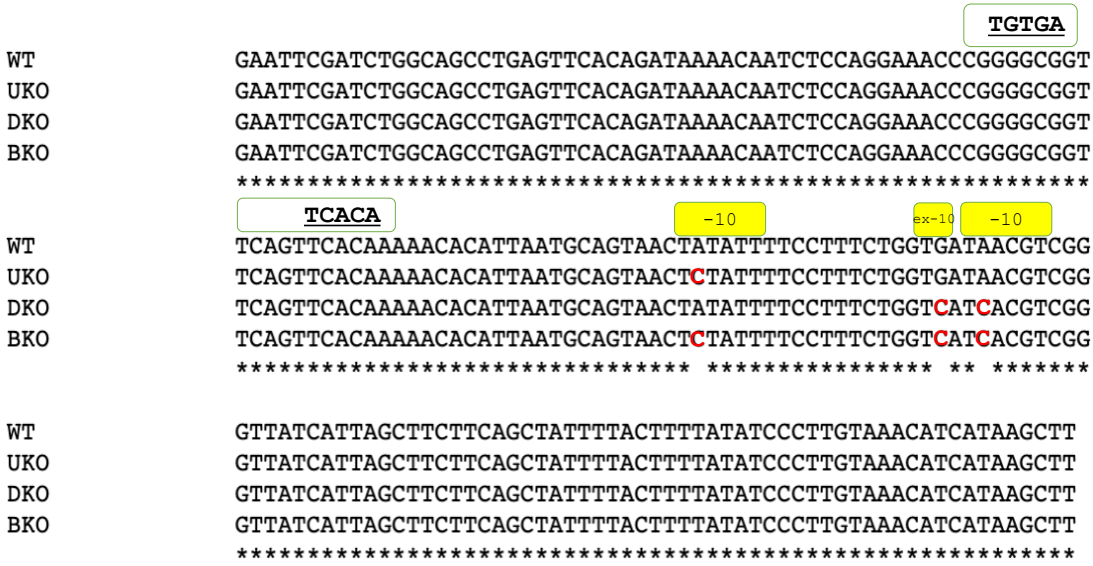


Figure 3.7: Alignment of *pic p042* mutated derivatives with different at -10 hexamer elements

The figure presents base sequence alignment of the wild type *pic p042* promoter (WT), the upstream -10 element knock out (UKO), both -10 element knock outs (BKO), and the downstream -10 element knock out (DKO) respectively. The red-letters C show the points of mutations at -10 elements in the region of *pic p042* promoter. Yellow rectangles are located above the two potential -10 elements and the extended -10 elements in the *pic* promoter region. The green rectangle denotes the DNA site for CRP. (Thompson *et al.*, 2003)

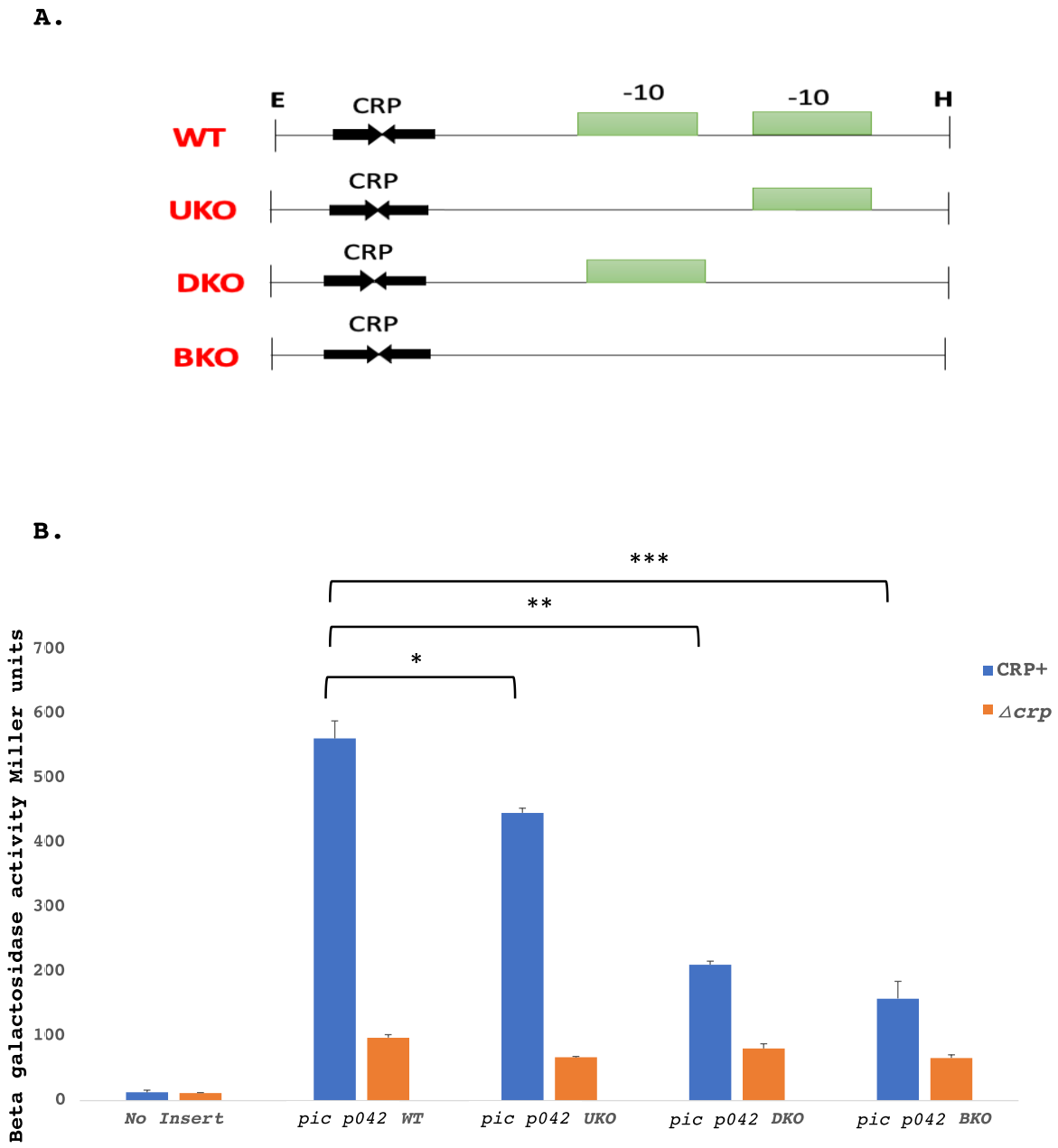


Figure 3.8: Mutational analysis of *pic p042* promoter

A. The figure illustrates *pic p042* promoter derivatives, WT, UKO, DKO and BKO. The green bars represent the -10-hexamer element, and the black arrows denoted the CRP binding site. Restriction sites are shown as E (EcoRI) and H (HindIII).

B. The figure shows measured ß- galactosidase levels expressed in *E. coli* strains M182 Δ*lac* and M182 Δ*lac* Δ*crp* carrying the *lacZ* expression vector (pRW224) or pRW224/*pic p042* WT or pRW224/*pic p042* UKO or pRW224/*pic p042* DKO or pRW224/*pic p042* BKO**.** Each column of promoter activity indicates the average of three independent ß- galactosidase assays and the error bars

represent the standard deviation (STDEV) and * indicates $P < 0.002$, ** indicates $P < 0.000002$; and *** indicates $P < 0.000005$ using a student's *t*-test.

3.2.4 CRP activation at the *pic p042* promoter

To investigate how CRP, activates transcription at the *pic p042* promoter, the experiment in Figure 3.8 was repeated but with CRP mutants lacking specific Activating Regions. Hence, genes encoding wild type CRP, CRP with inactive region 1 (CRP AR1$^-$), CRP with inactive region 2 (CRP AR2$^-$) and CRP with inactive region 1 and inactive region 2 (CRP AR1$^-$ AR2$^-$), were each cloned into pDU9 and transferred into *E. coli* K-12 strain M182 Δ*lac* Δ*crp* that also contained *pic p042*::*lac* fusions in pRW224. In these experiments, I used the *pic p042* WT::*lac* fusion, the *pic p042* UKO::*lac* fusion, the *pic p042* DKO::*lac* fusion or the *pic p042* BKO::*lac* fusion, individually. This resulted in host cells carrying two plasmids (one with the *pic* promoter in pRW224 and the other with CRP in pDU9), thereby permitting the measurement of CRP-dependent activation using ß-galactosidase assays.

The results presented in Figure 3.9 show that the *pic p042* WT::*lac* fusion is highly expressed in the presence of pDCRP or pDCRP AR2$^-$. Less than 150 Miller Units of *pic p042* promoter activity were recorded in the negative control and with pDCRP AR1$^-$ or pDCRP AR1$^-$ AR2$^-$. With the *pic p042* UKO::*lac* fusion, ß-galactosidase levels with pDCRP or pDCRP AR2$^-$ were almost the same level as with the *pic p042* WT::*lac* fusion, suggesting that RNAP does not use the upstream -10 element of *pic p042*, and it prefers the downstream one. In sharp contrast, knocking out the downstream -10 element in the *pic p042* DKO::*lac* fusion greatly decreased *lacZ* expression. Knocking out both -10 elements resulted in even lower ß-galactosidase levels, so it is possible that RNAP might use the upstream -10 element to drive some transcription at *pic p042*. Taken together, these results indicate that CRP mainly uses activating region 1 (AR1) to interact with RNAP and recruit it to the downstream promoter for transcription. Hence, *pic p042* can be considered as Class I CRP-dependent.

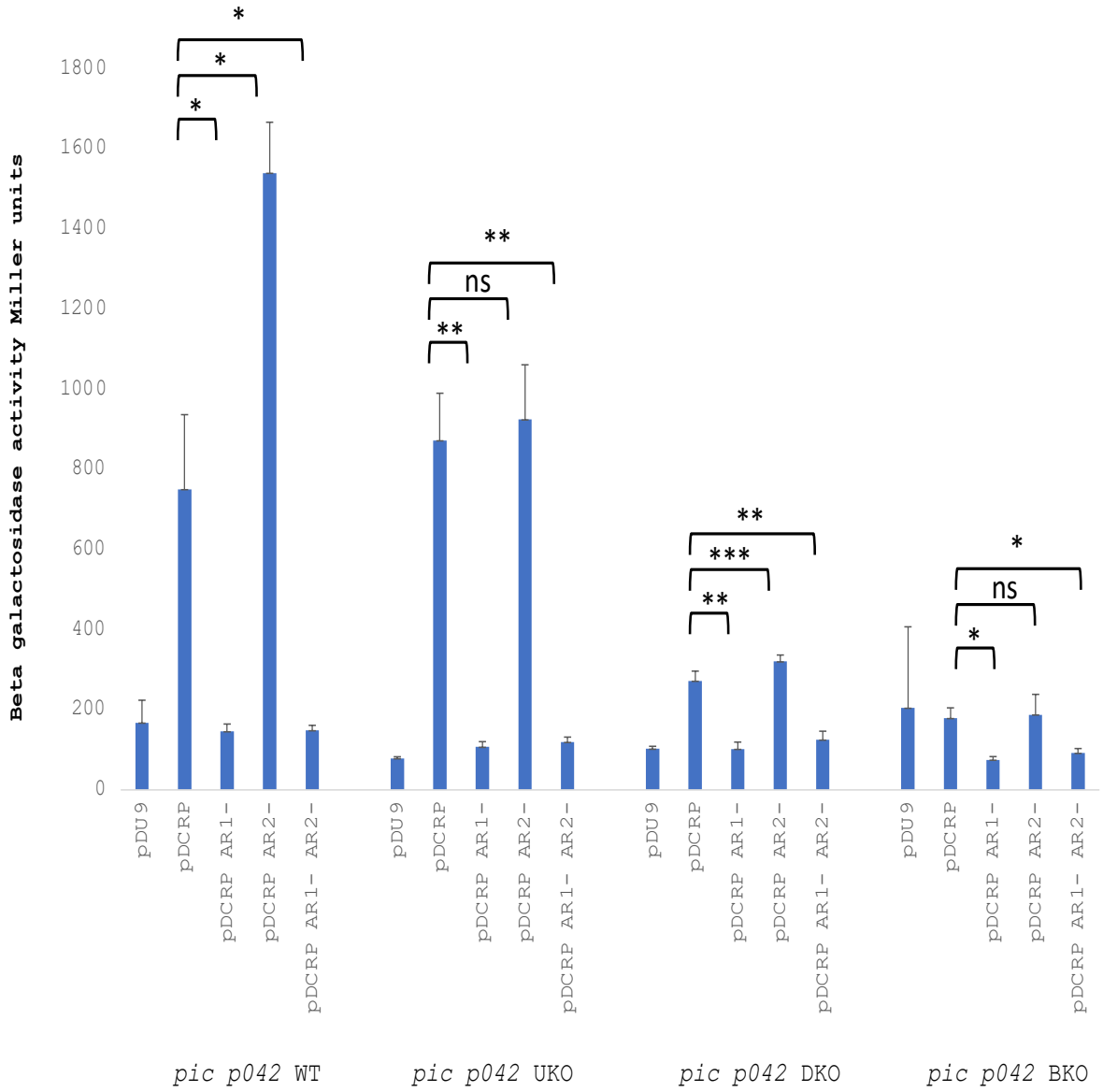


Figure 3.9: Measured *pic p*042 promoter activities with different CRP derivatives

The Figure shows measured ß-galactosidase activities in *E. coli* strain M182 Δ*lac* Δ*crp* carrying two plasmids. One encodes a *pic p042*::*lac* fusion vector *(*pRW224/*pic p042* WT or pRW224/*pic p042* UKO or pRW224/*pic p042* DKO or pRW224/*pic p042* BKO). The other encodes wild type of mutant CRP in pDU9 vector (pDU9 or pDCRP or pDCRP AR1$^-$ or pDCRP AR2$^-$ or pDCRP AR1$^-$AR2$^-$). The bar chart shows measured levels of ß-galactosidase in each strain with each measurement driving from the average of three independent assays and the error bars represent the standard deviation (STDEV). * Indicates $P < 0.005$, ** indicates $P < 0.0005$, *** indicates $P < 0.05$; and **ns** not significant using a student's *t*-test.

3.3 The UPEC CFT073 *pic* regulatory region

3.3.1 Investigation of the *pic p073* promoter region and transcription regulation

The purpose here was to study the *pic p073* promoter and compare it with the *pic p042* promoter. A 107 bp DNA segment from upstream of the UPEC strain CFT073 *pic* gene was amplified by PCR, using chromosomal DNA and primers listed in Table 2.4. The PCR product was then cloned into the promoter-less *lac* expression vector, pRW224, using flanking EcoRI and HindIII sites. This plasmid was used in assays to identify active -10 hexamer element for mapping the TSS, and for investigating any regulation by Fis and AggR. The base sequence of the promoter fragment and other details are shown in Figure 3.10 and Figure 3.11.

The recombinant plasmid, pRW224/ *pic p073* WT, was transformed into *E. coli* K-12 BW25113 Δ*lac* and *E. coli* K-12 BW25113 Δ*lac* Δ*crp.* Then the transformed cells were grown in LB medium at 37˚C, with shaking to mid-logarithmic phase (OD_{650}=0.4-0.6). The culture then was lysed to release the cell contents for measuring the ß-galactosidase levels. Plasmid pRW224, with no insert was used as a negative control.

Sequence analysis shows a potential DNA site for CRP in *pic p073* WT: starting at base 90 5`-**TGT**AA**C**AGACA**TCACA-**3`, (Figures 3.10 and 3.11B). An AT% calculation of the bases downstream of CRP DNA site shows a percentage of 60 %. Data in Figure 3.11C show that high levels of ß-galactosidase are recorded with *pic p073,* and these are reduced 4-fold in Δ*crp* cells. This result indicates the presence of a promoter in the insert fragment, *pic p073* WT, and this promoter is a CRP-activated promoter.

>*pic p073* WT

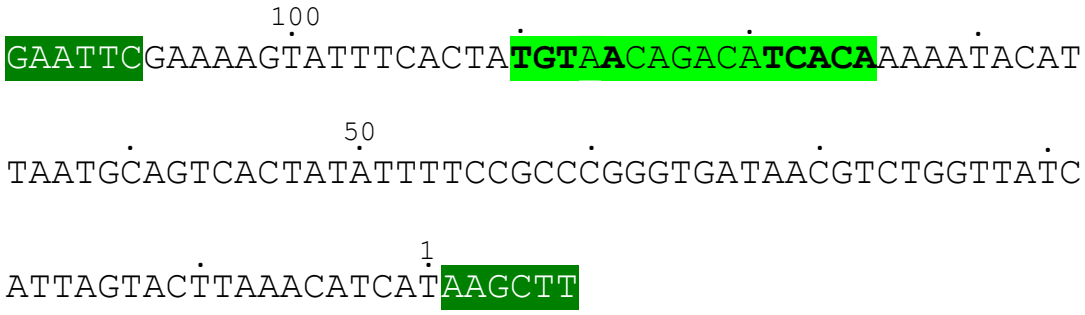


Figure 3.10: Fragment carrying the UPEC CFT073 *pic* promoter DNA sequence

The figure shows the 107 bp base sequence of the starting *pic p073* promoter fragment (WT), from upstream of the *pic* gene. The CRP binding site is highlighted in light green. The sequence highlighted with dark green correspond to the EcoRI and HindIII restriction sites. The numbering of the fragment starts from HindIII site.

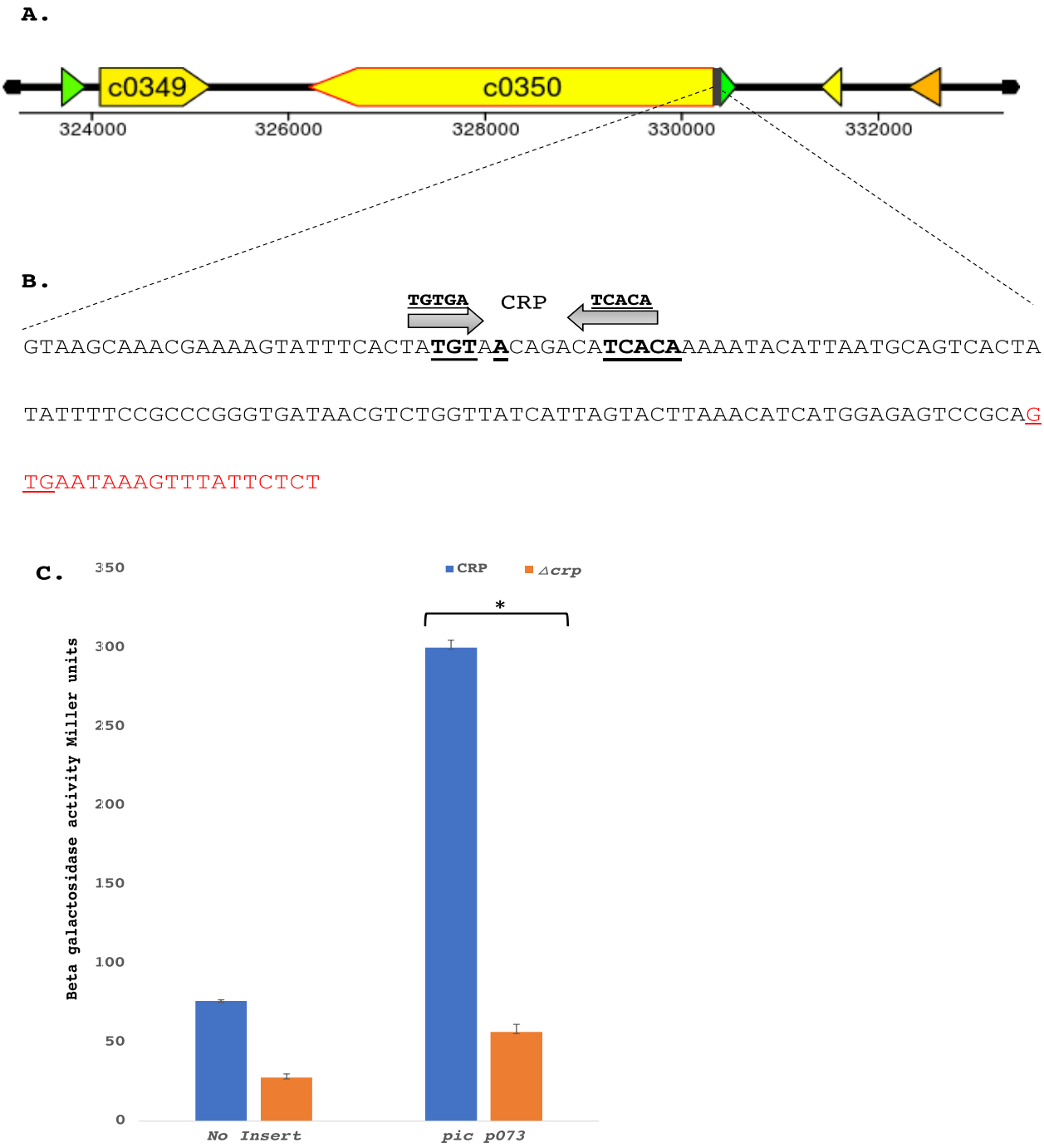


Figure 3.11: Activity of the *pic p073* promoter

A. The panel shows a diagram of the position of *pic p073* gene in the UPEC chromosome strain CFT073 (Chaudhuri *et al.,* 2007).

B. The panel shows the DNA sequence of the *pic p073* promoter region. Bases in black bold typeface correspond to the consensus DNA site for CRP. *pic* gene encoding sequence is shown by red underlined bases.

C. The bar chart illustrates measured ß- galactosidase activities in *E. coli* K-12 strain BW25113 WT *Δ lac* (blue bars) and *E. coli* K-12 strain BW25113 *Δcrp Δ lac* (orange bars). The strains carried empty pRW224 or pRW224/ *pic p073* WT as indicated. Each data bar indicates the average of three independent ß- galactosidase assays and the error bars represent standard deviation (STDEV). * Indicates $P < 0.00000004$ using a student's *t*-test.

3.3.2 Identification of the transcript start site at the *pic p073* promoter

The purpose of this section was to locate the actual TSS in the *pic p073* promoter region, using the 5` RACE method and pRW224/ *pic p073* WT. Hence, *E. coli* K-12 strain M182 Δ*lac* with this plasmid was grown in LB medium at 37˚C, with shaking to late-logarithmic phase (OD_{600}=1.0) and total RNA was isolated using a kit.

The TSS for *pic p073* was identified using the 5`RACE kit (following same procedure in the *pic p042* TSS Section). RNA, extracted from growing cells that carry pRW224/ *pic p073* WT, was converted into cDNA then tailed with poly(A). The mixture from this last step was amplified by PCR using the anchor primer (forward primer) 5`-GACCACGCGTATCGATGTCGACTTTTTTTTTTTTTTTV-3` and SP2 (reverse primer) (listed in Table 2.4). The product from this PCR was then cloned into pJET1.2 vector, using the CloneJET PCR cloning kit, and transformed into *E. coli* K-12 strains M182 Δ*lac.* Recombinant plasmid pJET1.2/ *pic p073* TSS was isolated from cells and sequenced by primers using the CloneJET PCR kit (primers listed in Table 2.4).

Figure 3.12 shows the results, the starting base of the *pic p073* promoter transcript is 5`**C**`3 located downstream from the likely -10 hexamer element 5`TATATT`3, and a possible -35 elements 5`-ACAAAA-3`, which overlaps with the likely DNA site for CRP binding. The start codon 5`-GTG-3` of *pic p073* is separated from TSS by 55 bp. This result suggests that RNAP selects the upstream promoter in *pic* promoter fragment to drive *lacZ* expression. The result also suggests that RNAP uses a different -10 elements in *pic p073* from *pic p042* which indicates the possibility of multiple promoter elements in the *pic* regulatory region.

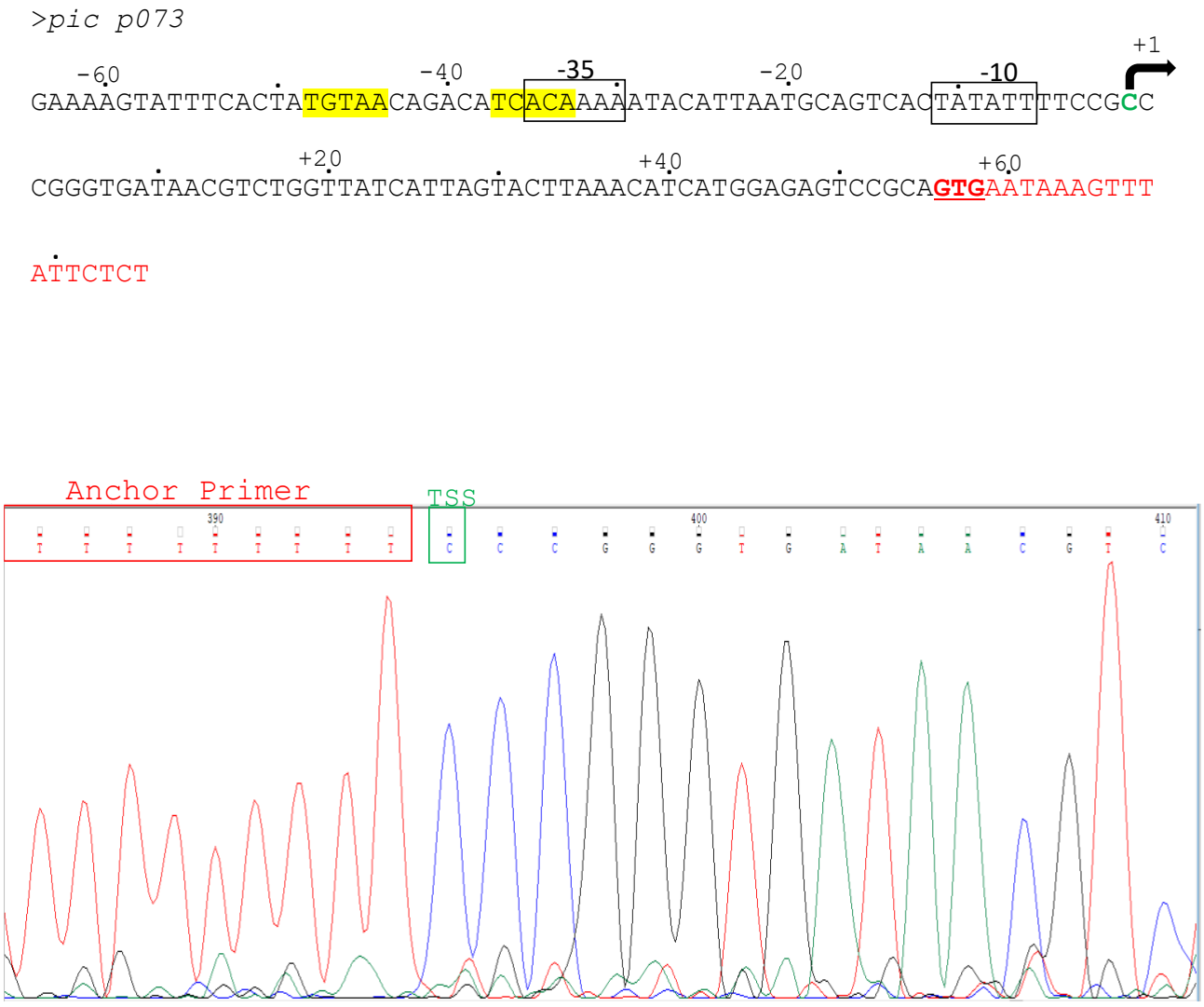


Figure 3.12: Mapping the *pic p073* promoter TSS

The figure shows the identification of the TSS at the *pic p073* promoter. The upper part shows the *pic p073* promoter region base sequence. The arrow and +1 represent the TSS (green bold base). The DNA site for CRP is highlighted in yellow. The suggested promoter -10 and -35 hexamer elements are boxed. The lower part shows the 5`RACE sequencing. The start codon of the *pic* gene is shown with bold red type and underlining bases. The numbering of the fragment starts from the TSS.

3.3.3 Identification of functional -10 element at the *pic p073* promoter

The aim was to locate the functional -10 hexamer element in order to confirm the mechanism of activation by CRP. As for *pic p042*, hexamer elements that concurred with the -10 consensus could be found at different locations downstream of the DNA site for CRP at the *pic p073* promoter from UPEC strain CFT073. Hence, the 107 bp *pic p073* WT fragment, shown in Figure 3.10, was used to introduce a point of mutation at position 2A of 5`-T**A**TATTTT-3` (Upstream knockout, UKO), and, also, double point mutations in downstream extended -10 elements 5`-T**G**AT**A**ACGT-3` (Downstream knockout, DKO). In addition, the UKO and DKO mutations were combined 5`-CT**A**TATTTTCCGCCCGGGT**G**AT**A**ACGT-3` (Both knockout, BKO). Alignment of the base sequences in the different constructs is shown in Figure 3.13, together with a schematic diagram of the four constructs: WT, UKO, DKO and BKO, in Figure 3.14A.

The different constructs were cloned into pRW224, and recombinant plasmids were transformed into *E. coli* strain M182 Δ*lac* and M182 Δ*lac* Δ*crp.* Transformed cells carrying pRW224/ *pic p073* WT (positive control), pRW224/ *pic p073* UKO, pRW224/ *pic p073* DKO, pRW224/ *pic p073* BKO, and empty vector pRW224, used as a negative control, were grown, and assayed to measure ß-galactosidase expression.

Data in Figure 3.14B show that knockout of the upstream -10 element (UKO) results in a significant decrease in *pic p073* promoter activity compared with the positive control (*pic p073* WT). Note that the residual activity is reduced in Δ*crp* cells by 2-fold. In contrast, the DKO mutations resulted in less of a decrease in promoter activity, whilst, as expected, the BKO construct showed the greatest decreased production in promoter activity. Note that for each promoter, there was a reduction in promoter activity in the Δ*crp* strain. Taken together, these results indicate that RNAP prefers the upstream promoter in the *pic p073* promoter region and that activation by CRP predominantly works by a Class II mechanism, with the DNA site for CRP separated from functional -10 hexamer element by 21bp.

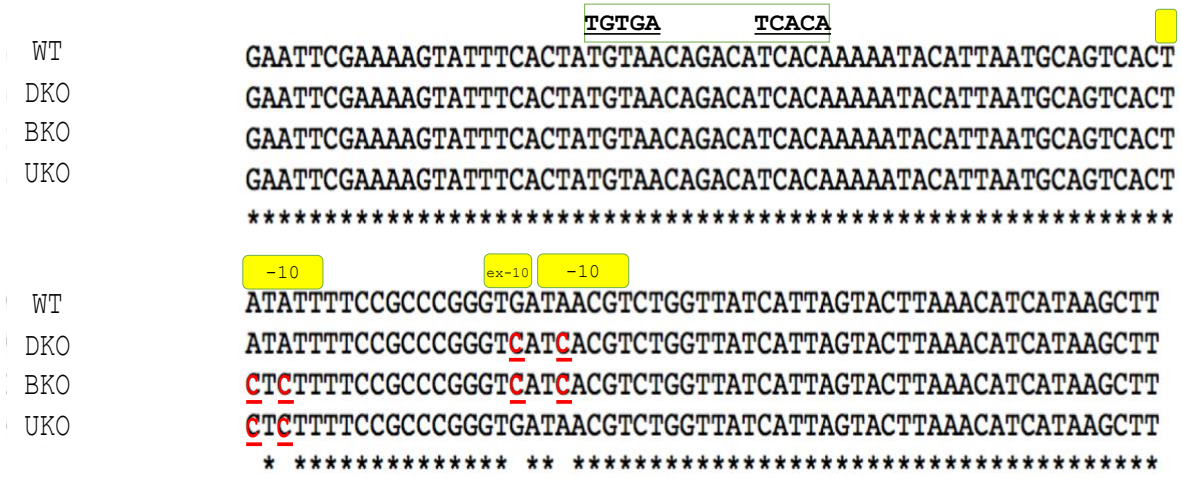


Figure 3.13: Alignment of *pic p073* derivatives with point mutations in alternative -10 hexamer elements

The figure presents the bae sequence alignment of wild type *pic p073* promoter (WT), downstream -10 element knock out (DKO), both -10 element knock out (BKO); and upstream -10 element knock out (UKO), as shown on the left side. The red-letters C denote the inserted mutations. The yellow rectangles indicate potential -10 elements and extended -10 elements in the promoter. The green rectangle represents the DNA site for CRP. (Thompson *et al.*, 2003)

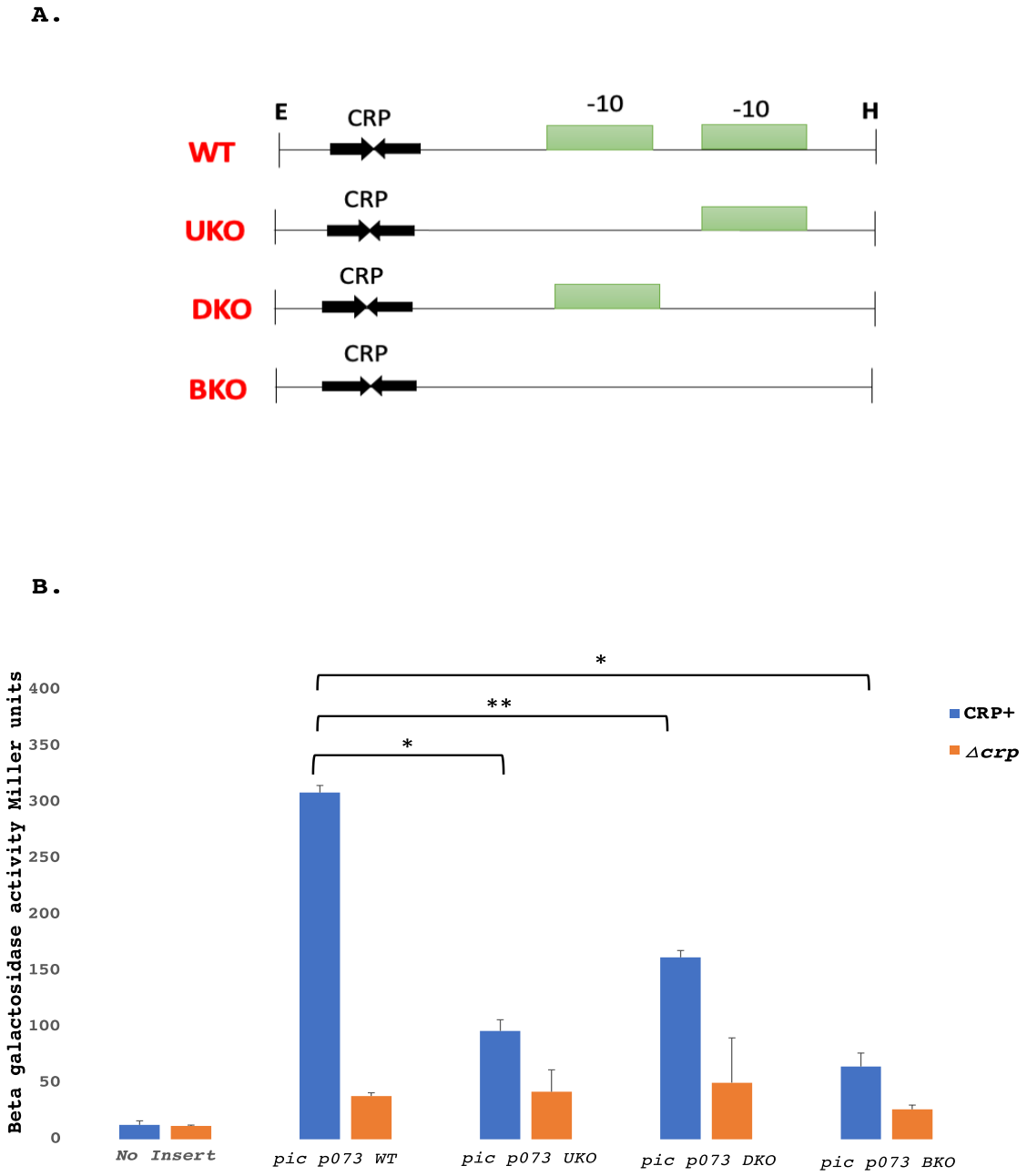


Figure 3.14: Mutational analysis of the *pic p073* promoter

A. The figure shows schemes of the WT, UKO, DKO and BKO *pic p073* promoter fragments. The green bars represent -10 element. The black arrow represents the DNA site for CRP binding site. Restriction sites are shown as E (EcoRI) and H (HindIII).

B. The figure shows measured ß-galactosidase activities in *E. coli* K-12 strains M182 Δ*lac* (blue bars) and *E. coli* K-12 strain M182 Δ*lac* Δ*crp* (orange bars) carrying the empty pRW224, or pRW224/*pic p073* WT or pRW224/*pic p073* UKO or pRW224/*pic p073* DKO or pRW224/*pic p073* BKO. Each measurement indicates the average of three independent ß- galactosidase assays and the error bars

represent the standard deviation (STDEV). * Indicates $P < 0.0000006$, ** indicates $P < 0.0000008$ using a student's *t*-test.

3.3.4 CRP- dependent activation at the *pic p073* promoter

To investigate how CRP activates transcription at the *pic p073* promoter, the experiment in Figure 3.14 was repeated, but with CRP mutants lacking specific Activating Regions. Hence, genes encoding wild type CRP, CRP with inactive region 1 (CRP AR1$^-$), CRP with inactive region 2 (CRP AR2$^-$) and CRP with inactive regions 1 and 2 (CRP AR1$^-$ AR2$^-$) were each cloned into pDU9 and transformed into *E. coli* K-12 strain M182 Δ*lac* Δ*crp* that also contained *pic p073*::*lac* fusions in pRW224.Then negative control, pDU9, was also included. In these experiments, I also used the *pic p073* WT::*lac* fusion, the *pic p073* UKO::*lac* fusion, the *pic p073* DKO::*lac* fusion or the *pic p073* BKO::*lac* fusion, individually.

Results in Figure 3.15 show that *pic p073* activity is highest in the presence of WT CRP and this is the benchmark in this experiment. With CRP AR2$^-$, expression is halved, showing the need for CRP AR2 to recruit RNAP on the *pic p073* promoter. Note that CRP AR1$^-$ gives low ß-galactosidase expression, similar to the negative control (pDU9). All other three promoter constructs show lower in ß-galactosidase expression. The experiments with the *pic p073* UKO::*lac* fusion underscores the importance of the upstream promoter *pic p073* expression. In the situation where both promoters are knocked out in *pic p073*, the ß-galactosidase level was recorded as the lowest level with all the CRP derivatives. Taken together, the data indicate that *pic p073* is an ambiguous promoter, but it is predominantly a Class II CRP-dependent promoter and so both CRP AR1 and AR2 are used for RNAP recruitment.

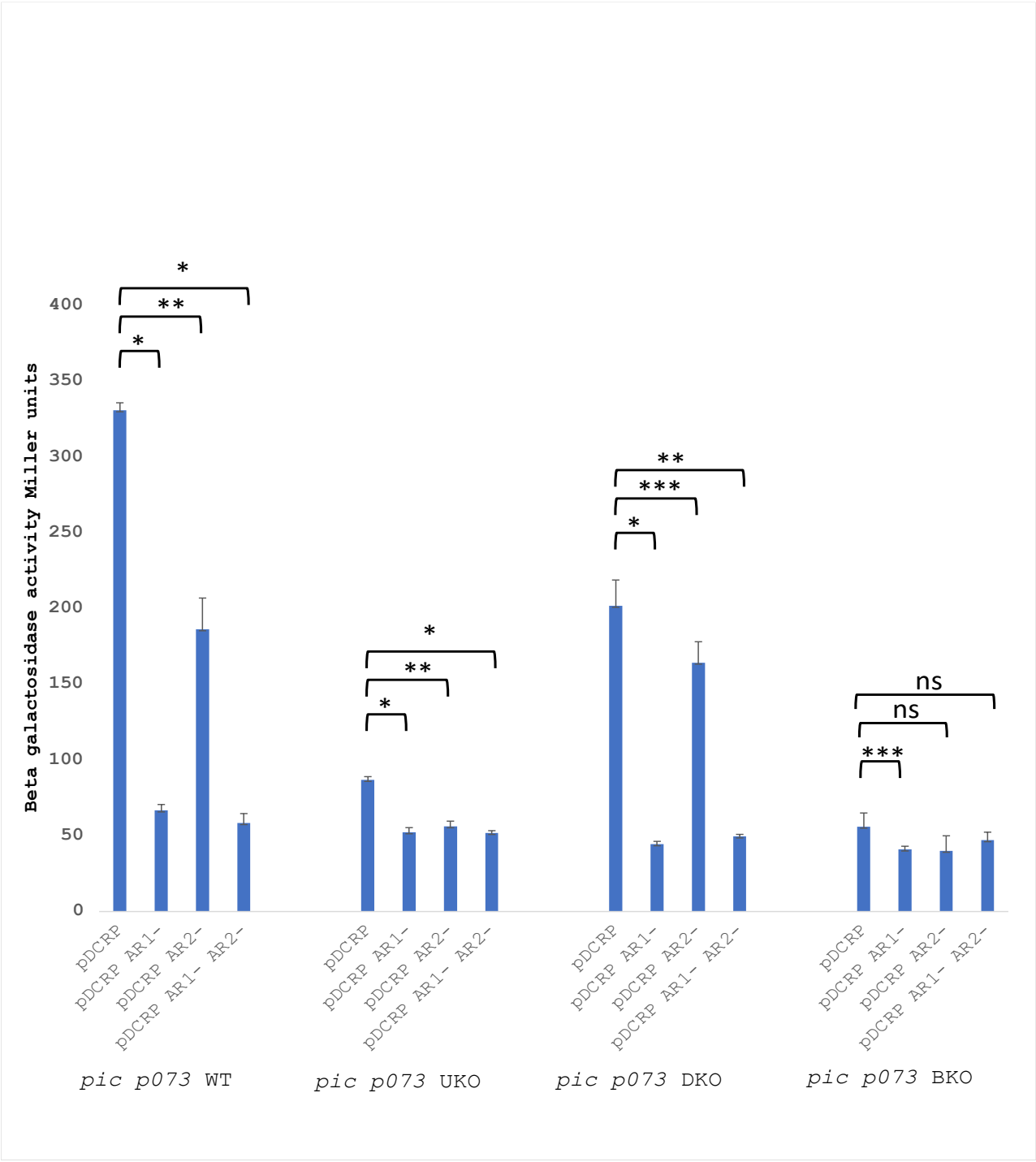


Figure 3.15: Measured *pic p073* activity in CRP derivatives

The Figure shows measurements of expression of the *pic p073*::*lac* fusion with different CRP derivatives. *E. coli* K-12 strain M182 Δ*lac* Δ*crp* carries two plasmids (1) one is the *lacZ* expression vector pRW224/ *pic p073* WT or pRW224/ *pic p073* UKO or pRW224/ *pic p073* DKO or pRW224/ *pic p073* BKO. And (2) the CRP mutant construct (pDCRP WT, pDCRP AR1^{-}, pDCRP AR2^{-}, pDCRP

AR1^{-}AR2^{-}) separately**.** Here, pDU9 was used as negative control, and pDCRP as positive control. Each activity measurement is the average of three independent ß-galactosidase assays and the error bars represent the standard deviation (STDEV). * Indicates $P < 0.000005$, ** indicates $P < 0.0005$, *** indicates $P < 0.05$; and **ns** not significant using a student's *t*-test.

3.4 Comparison of the EAEC 042 and UPEC 073 *pic* regulatory regions

3.4.1 Alignment of *pic p042* and *pic p073* base sequences

The base sequence alignment of the *pic* promoter region from EAEC strain 042 and in UPEC strain CFT073 is shown in Figure 3.16, which highlights some differences. For example, *pic p073* as a 28 bp deletion near the ORF. A DNA site for CRP appears in both promoters. The site in *pic p073* shows a 9/10 match to the CRP consensus binding site, whilst the site in *pic p042* shows only a 7/10 match. Both promoter regions carry plausible -10 hexamer elements located 21bp downstream and 41 bp downstream of the DNA site for CRP binding site. These hexamer elements are respectively, 5`-TATATTT-3` and 5`-TAACGT-3` and they are conserved in both promoter regions. The existence of two -10 elements suggests two promoters that I named 'upstream promoter' and 'downstream promoter': the names were dependent on the promoter location from the *pic* gene. A difference to point out is a cluster of GC bases, the "GC marker", found at a different position in each promoter. For the *pic p073* promoter, the GC marker is located downstream of the upstream promoter, whereas in the *pic p042* promoter, it appears adjacent to and upstream of the CRP binding site.

The aim of this section was to study the differences in base sequence between the two promoter regions, that show 68% base sequence identity and 32% divergence (Figure 3.1) I wanted to understand how RNAP switches between the two possible promoters, changing from one activation class to another, for the same virulence factor gene, *pic*. To do this, I used "transplant" methodology which is the transfer of short sequence segments from one promoter to another. The first difference that I assessed was the location of CG marker, the second difference was the 28 bp deletion in *pic p073*, and the third difference was the deviation from consensus of the DNA site for CRP in *pic p042*. I also examined the trans dominance of CRP mutants at different *pic* promoters and the consequences of transcript start site selection on translation efficiency.

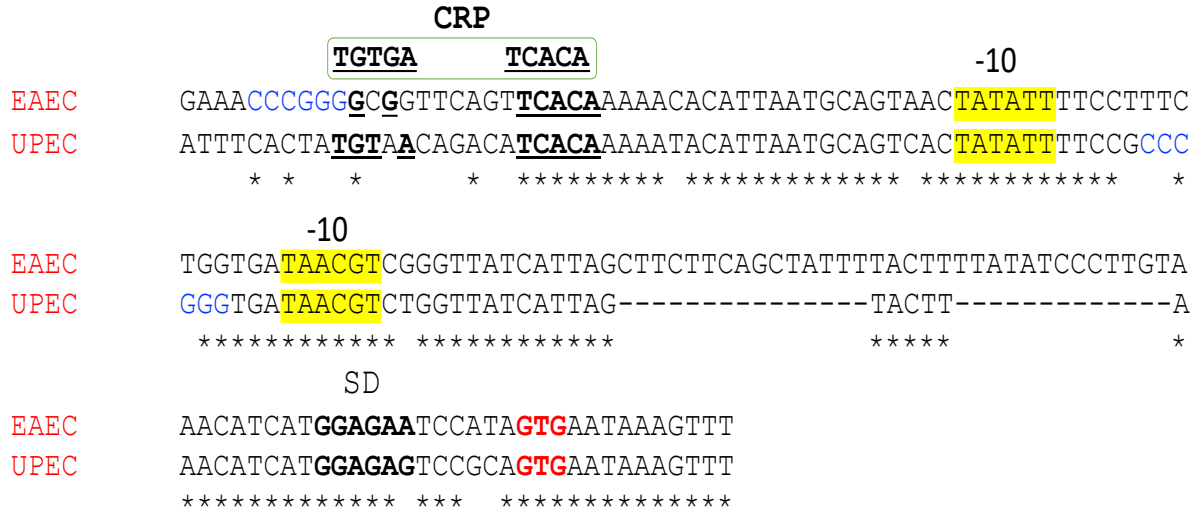


Figure 3.16: Alignment of the EAEC strain 042 and UPEC strain CFT073 *pic* promoter region base sequences

The figure illustrates the alignment of the EAEC and UPEC *pic* promoter region sequences. The DNA site for CRP is shown in bold with bases that match consensus (highlighted in green) underlined bases. The two -10 hexamer elements are highlighted in yellow. The SD sequence is in bold, and the coding region highlighted with red bases. The GC marker is marked with blue bases. Starts indicating identical bases and dashes indicate gaps introduced to support the alignment. (Thompson *et al.*, 2003).

3.4.2 The influence of the GC marker on *pic p042* activity

To investigate effects of the GC marker on *pic p042* promoter activity, a new construct was made with the marker removed from the *pic p042* promoter. This was done by transplanting the 15 bp upstream of the CRP binding site in *pic p073* into *pic p042,* as shown in Figure 3.17A. This was done using PCR and primers listed in Table 2.4. The new construct named *pic p042 U*, size 147 bp, was cloned into pRW224, again using EcoRI and HindIII sites. The same procedure was applied to the other *pic p042* derivatives: *pic p042* UKO, *pic p042* DKO and *pic p042* BKO. The recombinant plasmids pRW224/ *pic p042 U* WT, pRW224/ *pic p042 U* UKO, pRW224/ *pic p042 U* DKO and pRW224/ *pic p042 U* BKO were then transformed into *E. coli* K-12 strains M182 Δ*lac* and *E. coli* K-12 strain M182 Δ*lac* Δ*crp* individually. Transformed cells was then grown in LB medium at 37˚C, with shaking to mid-logarithmic phase (OD_{650}= 0.4-0.6). As before, ß-galactosidase assays were used as a proxy for promoter activity.

The result, illustrated in Figure 3.17C, shows that the replacement of the GC marker here causes an increase in promoter activity, but the CRP-dependence is unchanged. Additionally, the same pattern of expression is found with the different *pic p042* derivatives, with the downstream promoter preferred, as CRP functions predominantly as a Class I activator. A possible explanation for the higher *pic p042 U* activity is that the increase of AT% in the upstream sequence allows RNAP alpha-CTD to bind properly, but this cannot be enough to switch the selection from Class I activation by CRP to Class II.

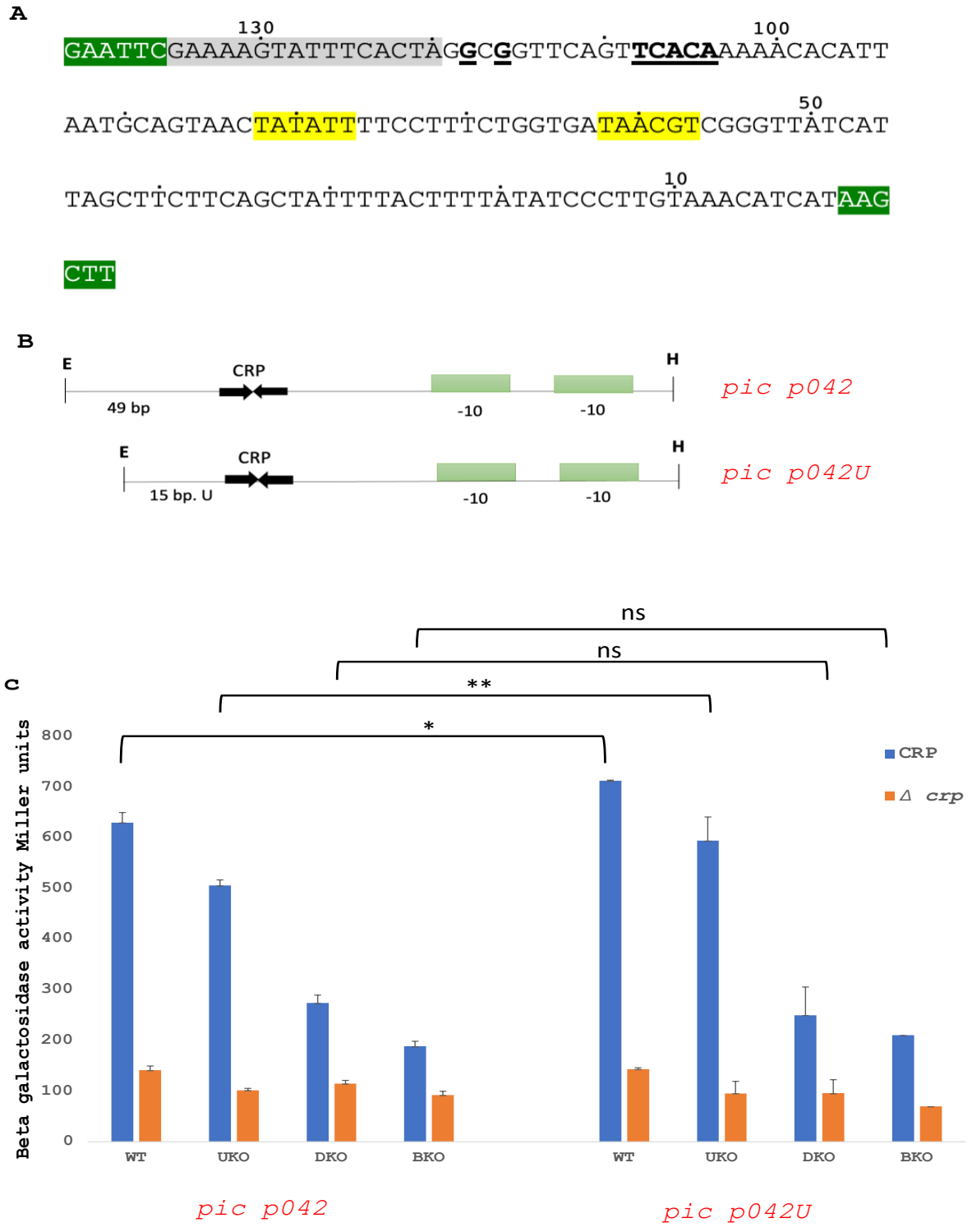


Figure 3.17: Influence of the GC marker on *pic p042 U* promoter activity

A. The figure shows the 147 bp DNA fragment carrying *pic p042 U*. The DNA site for CRP is shown with bold underlined bases. Yellow highlighter denotes the two -10 hexamer elements. Base

sequence highlighted with light grey denotes the sequence taken from *pic p073* that was transplanted into *pic p042* DNA sequence. The sequences highlighted dark green indicate the EcoRI and HindIII restriction sites. The numbering starts from the HindIII site.

B. Schematic diagrams of the *pic p042* promoter WT and the derivative *pic p042 U*. The green bars represent the two -10 elements. The black arrow represents the CRP binding site. E represents EcoRI and H represents HindIII.

C. The figure shows measured ß- galactosidase levels in *E. coli* K-12 strain M182 Δ*lac* (blue bars) and *E. coli* K-12 strain M182 Δ*lac* Δ*crp* (orange bars) carrying the *lacZ* expression vector pRW224/ *pic p042* WT or pRW224/ *pic p042* UKO or pRW224/ *pic p042* DKO or pRW224/ *pic p042* BKO with or without the GC marker. The left part shows activities of the starting *pic p042* derivatives, and the right part shows corresponding *pic p042 U* promoter activities. Each promoter activity is the average of three independent ß- galactosidase assays and the error bars represent the standard deviation (STDEV). * Indicates $P < 0.01$, ** indicates $P < 0.0001$;and **ns** not significant using a student's *t*-test.

3.4.3 Leader sequence deletions in *pic p042*

The *pic p073* base sequence has a 28 bp deletion downstream of the promoter, compared with the *pic p042* sequence (Figure 3.16), which will result in a shorter untranslated 5` end to the *pic* message. Here, guided by *pic p073,* I deleted 28 bases in *pic p042*, to study if this affects -10 element selection by RNAP, and hence, CRP activation class, or if there is any effect of this deletions on promoter activity. Hence, a new construct was made, named *pic p042* 28 with fragment size of 151 bp, using the reverse primer of *pic p073* (listed in Table 2.4). Figure 3.18A shows the difference in base sequence between *pic p042* and *pic p042* 28. The deletion was also introduced into the other *pic p042* derivatives: *pic p042* 28 UKO, *pic p042* 28 DKO and *pic p042* 28 BKO. Then, each construct was ligated into pRW224, using the EcoRI and HindIII sites and the recombinant plasmids were transformed into the *E. coli* K-12 strain M182 Δ*lac* for ß-galactosidase assays. A negative control, pRW224, with no insert, was included in the experiment.

Results illustrated in Figure 3.18B shows that with *pic p042* 28, in general, ß-galactosidase levels were lower by 2-fold in comparison with the starting *pic p042* derivatives. Interestingly, these lower levels were similar to those observed with *pic p073* (Figure 3.11C). However, the data in Figure 3.18B indicates that the 28 bp deletions does not impact promoter selection by RNAP, and CRP continues to activate in Class I mode.

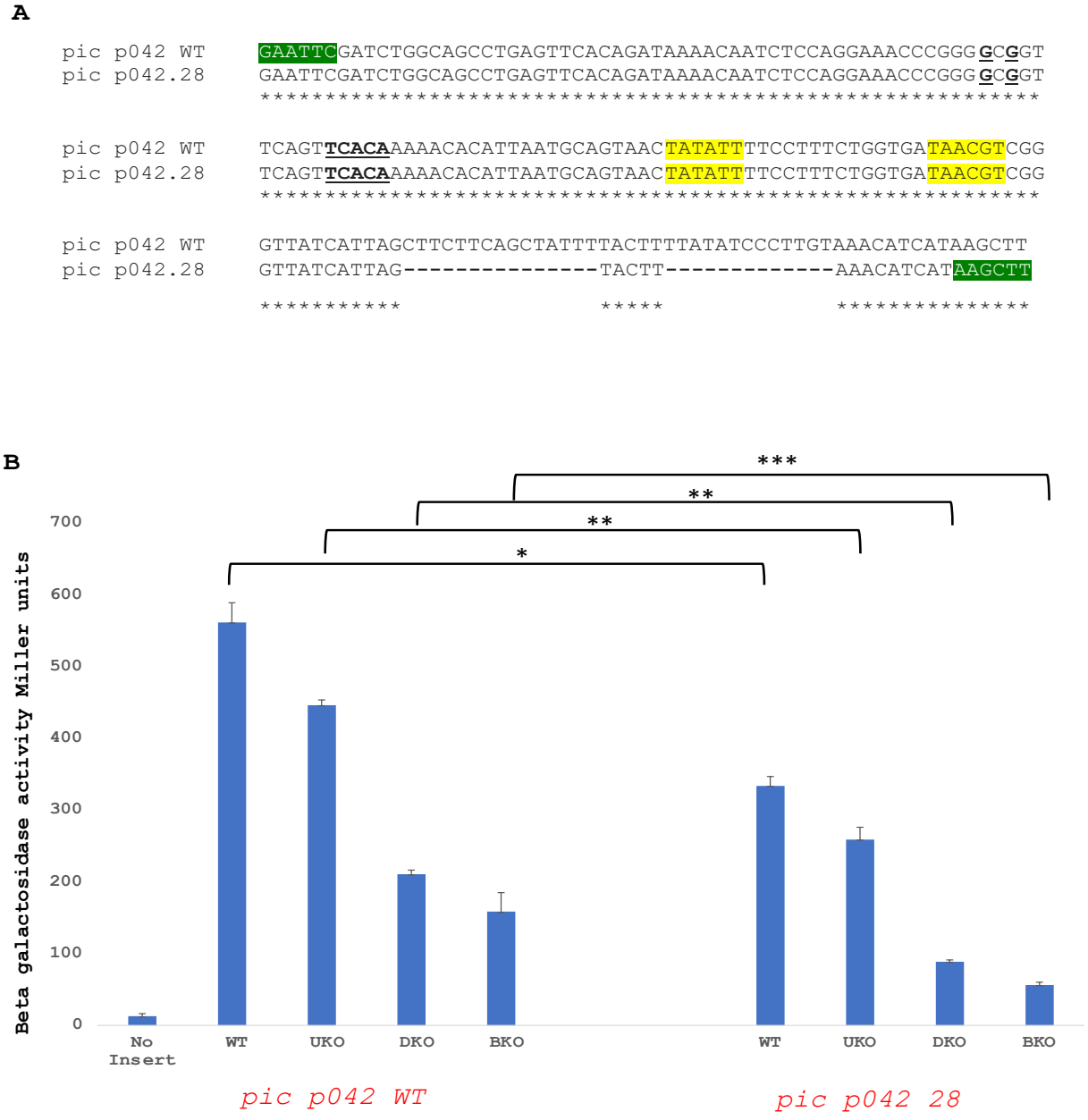


Figure 3.18: Effects of shortening *pic p042* downstream sequences

A. The figure illustrates the alignment of the base sequence of *pic p042* with *pic p042* 28. (Thompson *et al.*, 2003). The DNA site for CRP is shown with bold underlined bases. Yellow highlighter denotes the two -10 hexamer elements. The sequences highlighted dark green indicate the EcoRI and HindIII restriction sites.

B. The left part of the figure shows measured ß- galactosidase levels in *E. coli* K-12 strain M182 Δ*lac* carrying pRW224/ *pic p042* WT or pRW224/ *pic p042* UKO or pRW224/ *pic p042* DKO or pRW224/ *pic p042* BKO. The right part of figure shows activity with *pic p042* 28 derivatives

(pRW224/ *pic p042* 28 or pRW224/ *pic p042* 28 UKO or pRW224/ *pic p042* 28 DKO or pRW224/ *pic p042* 28 BKO). Each promoter activity indicates the average of three independent ß-galactosidase assays and the error bars represent the standard deviation (STDEV). * Indicates $P < 0.0001$, ** indicates $P < 0.000005$; and *** indicates $P < 0.002$ using a student's *t*-test.

3.4.4 Improving the *pic p042* promoter DNA site for CRP

The *pic p042* DNA site for CRP 5`-G**G**C**G**GTTCAGT**TCACA**-3` shows three mismatches from the consensus that, likely, weaken CRP binding. Here, a consensus CRP binding site was introduced into *pic p042*, to study its effect on promoter activity. In this experiment, a new 179 bp construct was made, named *pic p042* 120, with a consensus DNA site for CRP 5`-**TGTGA**TTCAGT**TCACA**-3` (Figure 3.19A). Corresponding, CRP consensus sites were also introduced to *pic p042* UKO 120, *pic p042* DKO 120 and *pic p042* BKO 120. All these fragments were cloned into pRW224 and transformed into *E. coli* K-12 strains M182 Δ*lac* for testing via ß-galactosidase assays.

Results of the experiment are shown in Figure 3.19B. All ß-galactosidase levels were much higher with the *pic p042* 120 derivatives compared to the corresponding *pic p042* derivatives. This indicates the importance of the CRP consensus sequence in setting expression levels. *pic p042* 120 DKO showed lower promoter activity than *pic p042* 120 UKO indicating that the nature of the DNA site for CRP does not affect -10 hexamer choice and mechanism of activation at *pic p042*.

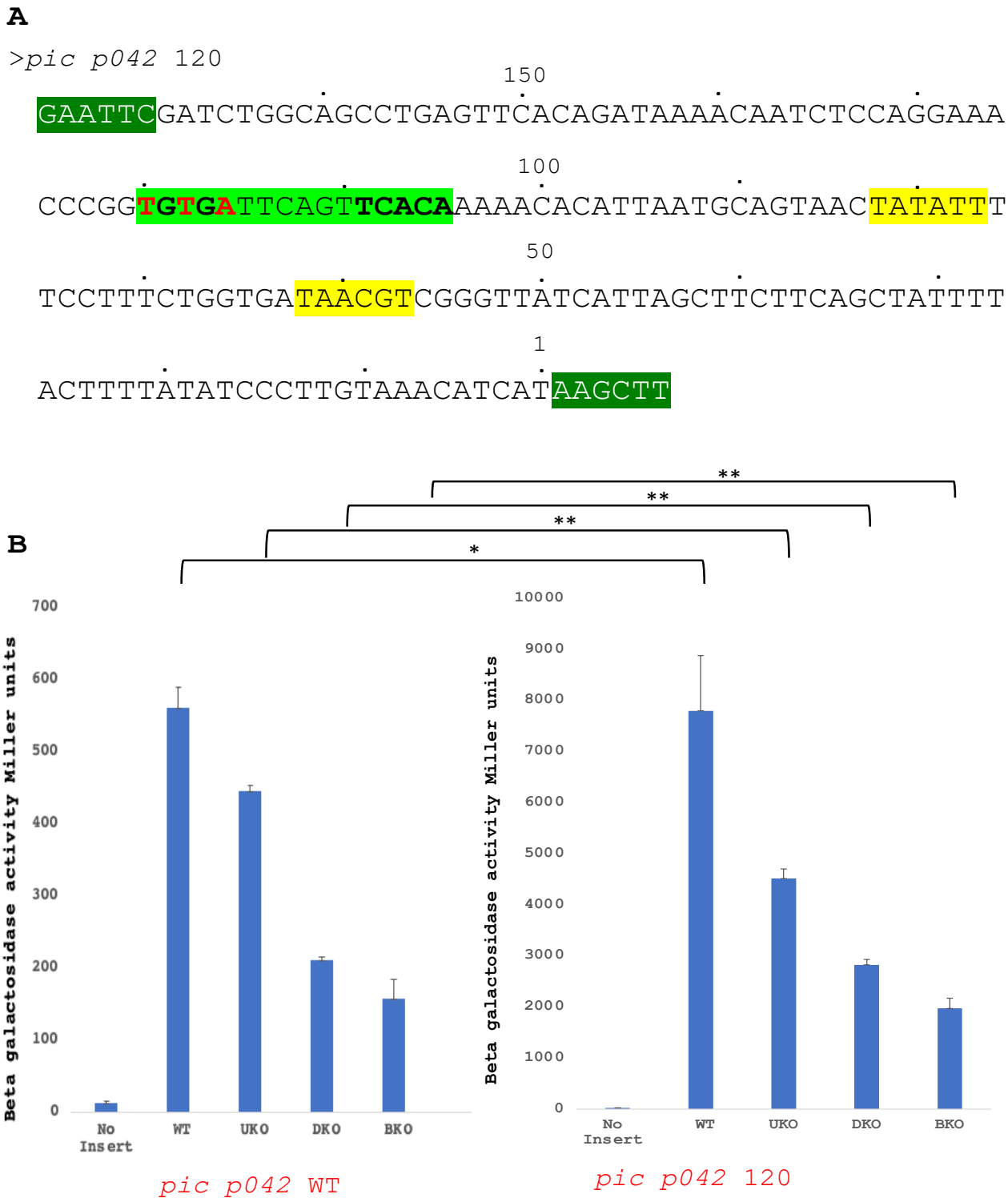


Figure 3.19: *pic p042* activity is affected by the quality of its DNA site for CRP

A. The figure shows the DNA sequence of *pic p042* 120 promoter fragment. The CRP binding site is highlighted in light green, and bases changed to consensus are in red bold bases. The

two -10 elements are highlighted in yellow. Restriction enzyme EcoRI and HindIII sites are highlighted in dark green. The numbering of the fragment starts from the HindIII site.

B. The left part of the figure shows measured ß-galactosidase levels in *E. coli* K-12 strain M182 Δ*lac* carrying the *lacZ* expression vector pRW224/ *pic p042* WT or pRW224/ *pic p042* UKO or pRW224/ *pic p042* DKO or pRW224/ *pic p042* BKO. The right part of the figure shows measured ß-galactosidase levels with the corresponding *pic p042* 120 derivatives (pRW224/ *pic p042* WT 120 or pRW224/ *pic p042* UKO 120 or pRW224/ *pic p042* DKO 120 or pRW224/ *pic p042* BKO 120). Each promoter activity indicates the average of three independent ß- galactosidase assays and the error bars represent the standard deviation (STDEV). * Indicates $P < 0.0003$,and ** indicates $P < 0.000005$ using a student's *t*-test.

3.4.5 Trans dominance of CRP mutants at the *pic* promoters

Trans dominance experiments were used to investigate the behaviour of CRP at the EAEC strain 042 and UPEC strain CFT073 *pic* promoter regions. Here, the test strain is *E. coli* K-12 M182 Δ*lac* carrying pRW224 with either *pic p042* WT or *pic p073* WT. These cells were then used for another transformation with pDU9 (negative control), or with pDCRP (CRP wild type), or with pDCRP $AR1^-$ $AR2^-$ (CRP with inactive region 1 and inactive region 2) leading to strains carrying two plasmids, pRW224 with *pic* promoter and pDU9 encoding wild type or mutant CRP. These transformed cells were grown in LB medium at 37˚C, with shaking to mid-logarithmic phase (OD_{650}= 0.4-0.6), lysed and assayed for ß-galactosidase activity. The negative control, pRW224, with no insert, was included in these experiments.

The results of these assays are shown in Figure 3.20. For *pic p042,* ß- galactosidase level shows a significant increase when cells expressed more wild type CRP, indicating the boost of transcription of *pic p042* by RNAP with the help of CRP, leading to higher expression of ß- galactosidase. However, when the extra CRP carries inactive of region 1 and region 2, the impact on promoter activity is less. With *pic p073*, ß- galactosidase expression is the same with or without added WT CRP, but promoter activity is decreased in cells carrying CRP with inactive region 1 and region 2. This trans-dominant effect must be due to defective CRP binding at *pic p073* and blocking access for WT CRP. Note that, at *pic p073*, activating regions are required in both subunits of the CRP dimer because the Class II activation mechanism is operating (Zhou *et al.,* 1994). In contrast, at *pic p042*, activating regions are only functional in one subunit, since the Class I activation mechanism is operating, and this likely explains why the trans-dominance is so weak.

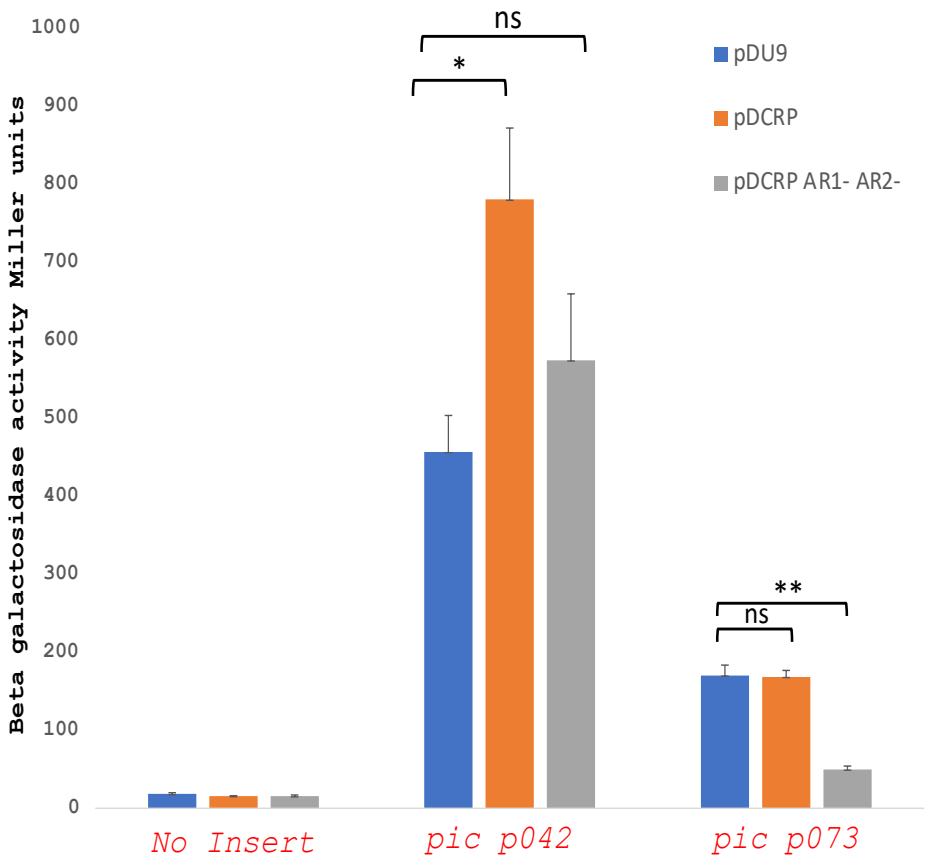


Figure 3.20: Trans-dominance of defective CRP at *pic* promoters

The figure illustrates *pic* promoter activity measured in *E. coli* K-12 strain M182 Δ*lac* cells carrying two plasmids: pRW224/*pic p042* WT (middle) or pRW224/*pic p073* WT (right) and pDU9 (blue bars), pDCRP (orange bars), pDCRP AR1$^-$AR2$^-$ (grey bars)**.** Empty pRW224 was included as a control (left). Each activity measure is the average of three independent ß- galactosidase assays and the error bars represent the standard deviation (STDEV). * Indicates $P < 0.005$, ** indicates $P < 0.0001$; and **ns** not significant using a student's *t*-test.

3.4.6 Translation efficiency at the *pic p042* and *pic p073* regulatory regions

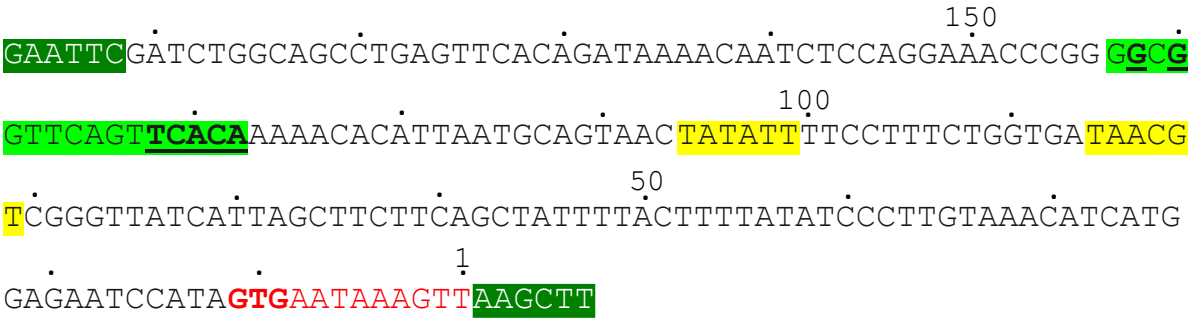
One of the consequences of the differences between *pic p073* and *pic p042* is that the untranslated leader sequence will differ from promoter to promoter (and mutant to mutant, UKO and DKO). Hence, I wanted to measure the efficiency of *pic* gene translation. In this section, the *lacZ* expression vector, pRW225, was used to study the efficiency of *pic* translation. As well as being promoter-less, the *lacZ* gene in pRW225 lacks a SD sequence and start codon. Here, in my constructions, when *pic p073* and *pic p042* fragments were inserted into pRW225, the SD sequence and start codon were provided by the cloned EcoRI-HindIII promoter fragment. Hence this allowed me to study the full 5' UnTranslated Region (5`UTR) of the *pic* promoters. The fusion of the *pic* promoter and the translation start of the *pic* gene with the pRW225 *lacZ* gene results in a protein fusion, whereas, with pRW224, the *pic*::*lac* fusion is an operon fusion.

To serve my purpose, new *pic p073* and *pic p042* fragments were constructed carrying the *pic* SD sequence and start codon 5`-GTG-3` of *pic*. The *pic p042* new fragment size was 203 bp and *pic p073* new fragment size was 143 bp (Figure 3.21). Derivatives of both promoters were also included in fragment constructions: *pic p042* UKO, *pic p042* DKO, *pic p042* BKO; and *pic p073* UKO, *pic p073* DKO, *pic p073* BKO. All the new constructs were ligated into pRW225 using the EcoRI and HindIII sites, and the recombinant plasmids were then transformed into *E. coli* K-12 strain, M182 Δ*lac*. Transformed cells, carrying the recombinant plasmids, were grown in LB medium at 37°C, with shaking to mid-logarithmic phase (OD_{650}= 0.4-0.6). Cultures were then lysed and assayed for ß-galactosidase activity.

The results of assays with the *pic p042* protein fusion are shown in the left part of Figure 3.22. The starting *pic p042* construct shows high ß-galactosidase levels, almost 7000 Miller Units. Knocking out the upstream promoter in *pic p042* reduced promoter activity by 1000 Miller Units, whereas the knocking out the downstream promoter in *pic p042* DKO caused a reduction of 3800 Miller Units. As expected, knocking out both promoters, in *pic p042* BKO, gave the lowest ß-galactosidase levels with the protein fusion situation in pRW225.

A

> *pic p042* (203bp)



B

> *pic p073* (143bp)

Figure 3.21: Base sequences of *pic p042* and *pic p073* fragments for protein fusions

The figure shows the DNA sequence of the (**A**) EcoRI-HindIII *pic p042* wild type promoter fragment, (**B**) EcoRI-HindIII *pic p073* wild type promoter fragment, including the start of the *pic* gene. The CRP binding site is highlighted in light green. The yellow highlighter denotes the two -10 hexamer elements. The start codon is shown in bold red bases. Sequences highlighted with dark green denote restriction sites EcoRI and HindIII.

With the different *pic p042* protein fusions, the hierarchy of the activities was similar to that found with operon fusions (Figure 3.8). Surprisingly, the results with the *pic p073* protein fusion (Figure 3.22 right) showed different results from the operon fusions (Figure 3.14). The data illustrated in the right part of Figure 3.22, show that *pic p073* appears to rely more on the downstream promoter. Hence, with *pic p073* DKO, apparent promoter activity is reduced by 5200 Miller Units, whilst knocking out the upstream promoter, with *pic p073* UKO, showed only a reduction 2500 Miller Units. As expected, knocking out both promoters, in *pic p073* BKO, gave the lowest ß-galactosidase levels with protein fusion situation in pRW225.

Figure 3.23 shows a combination of data from both *pic p042* and *pic p073* operon fusions and protein fusion. With *pic p042*, I found the same pattern of activation in both the operon fusion and protein fusion experiments. This argues that the efficiency of translation here is independent of the transcript start. In contrast, with *pic p073* derivatives, there is an apparent switch in transcription activation class between the operon fusion and the protein fusion. However, the more likely explanation is that the longer transcript, driven by the upstream -10 element, is less efficiently translated. This is likely due to an inhibitory 5'UTR structure (see later, Section 3.6).

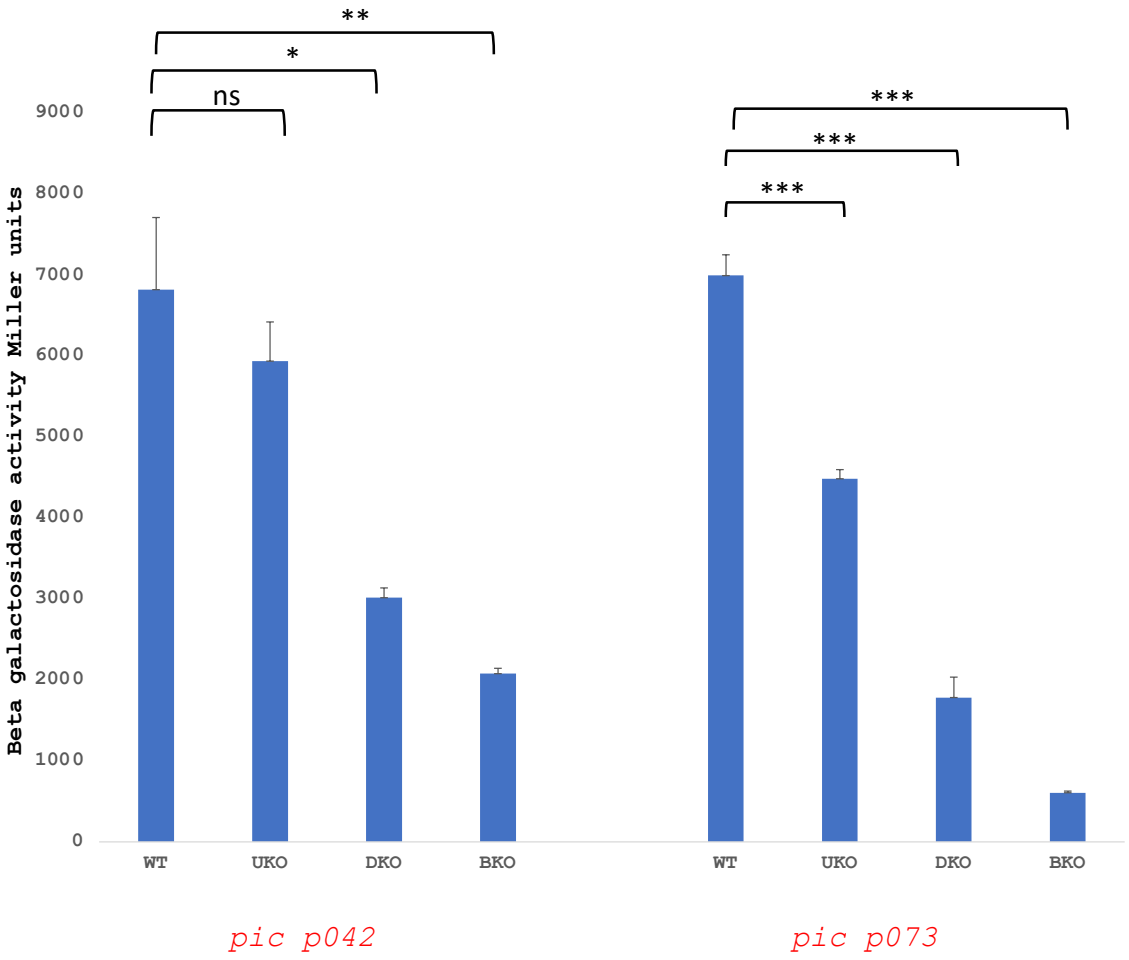


Figure 3.22: Expression of *pic p042* and *pic p073* protein fusions

The Figure shows measurements of ß-galactosidase activities in *E. coli* K-12 strain M182 Δ*lac* carrying pRW225 with *pic p042* promoter derivatives (left part of graph) or *pic p073* promoter derivatives (right part of graph). Each promoter activity indicates the average of three independent ß-galactosidase assays and the error bars represent the standard deviation (STDEV). * Indicates $P < 0.001$, ** indicates $P < 0.0007$, *** indicates $P < 0.000005$;and **ns** not significant using a student's *t*-test.

A

pic p042	*Protein fusions*	*Operon fusions*
WT	100%	100%
UKO	77.22%	79.48%
DKO	41.22%	37.55%
BKO	27.62%	28.20%

pic p073	*Protein fusions*	*Operon fusions*
WT	100%	100%
UKO	50.49%	31.24%
DKO	22.49%	52.47%
BKO	7.30%	20.98%

B

Figure 3.23: Comparison of measured *pic p042* and *pic p073* activities in operon fusions and protein fusions

A. Figure shows the measured expression of each derivative as a percentage of the value with the starting *pic p042* and *pic p073* fragments.

B. Graph shows a combination chart detailing relative expression from *pic p042* and *pic p073* derivatives, comparing operon fusion data (pRW224) shown as the blue line, with protein fusion data (pRW225), showed with green bars.

3.5 Involvement of other TFs in *pic* promoter regulation

The regulation of the expression of some virulence genes can be complex, with more than one transcription factor involved. Here, I tested two transcription factors, Fis and AggR, to study their impact on *pic* promoter activity. To do this, I used fusions of *pic* promoters with the *lacZ* reporter in pRW224 vector, and different bacterial host strains. The reason for selecting these factors was that Fis is a co-activator, with CRP, of transcript initiation at the EAEC 042 *pet* promoter, and AggR is known to be the master virulence regulator of EAEC virulence (Rossiter *et al.*, 2011; Yasir *et al.*, 2019).

3.5.1 Investigation of the role of Fis at the *pic* promoter

Analysis of the *pic* promoter region in UPEC and EAEC, Figure 3.24A, shows a binding site for Fis that resemble the consensus 5`-GNTcAaatTTtGaNC-3`, located after the downstream -10 hexamer element in both promoters (*pic p042* and *pic p073*). With this location, Fis could affect *pic* expression negatively. To study this, *E. coli* K-12 strain BW25113 Δ*lac* Δ*fis* was used. Recombinant plasmids pRW224/ *pic p042* WT (Figure 3.4) and pRW224/ *pic p073* WT (Figure 3.10) were transformed into *E. coli* K-12 strains BW25113 Δ*lac* and BW25113 Δ*lac* Δ*fis*. Then, transformed cells were grown in LB medium at 37˚C, with shaking to mid-exponential phase (OD_{650}= 0.4-0.6), and the level of ß-galactosidase in cell lysates was measured. A negative control, pRW224, with no insert, was included in the assay.

The influence of Fis on *pic* promoter expression is shown in Figure 3.24B. For *pic p042*, a 2-fold higher level of ß-galactosidase is found in cells that had a deletion of *fis* gene, indicating a repression role of Fis on *pic p042* expression. However, with *pic p073*, ß-galactosidase level in Δ*fis* was 90 Miller Units more than wild type indicating a smaller repressive effect of Fis on *pic p073*. Due to the location of the Fis binding site at both *pic* promoters, it is likely that Fis hinders RNAP binding or elongating, resulting in repression.

A

```
                        CRP BS
EAEC        GAAACCCGGGGCGGTTCAGTTCACAAAAACACATTAATGCAGTAACTATATTTTCCTTTC
UPEC        ATTTCACTATGTAACAGACATCACAAAAATACATTAATGCAGTCACTATATTTTCCGCCC
                 * *    *       *   ********* ************ ************    *
                                Fis BS
EAEC        TGGTGATAACGTCGGGTTATCATTAGCTTCTTCAGCTATTTTACTTTTATATCCCTTGTA
UPEC        GGGTGATAACGTCTGGTTATCATTAG----------------TACTT-------------A
             ************ ************                  *****              *

EAEC        AACATCATGGAGAATCCATAGTGAATAAAGTTT
UPEC        AACATCATGGAGAGTCCGCAGTGAATAAAGTTT
            ************* ***   **************
```

B

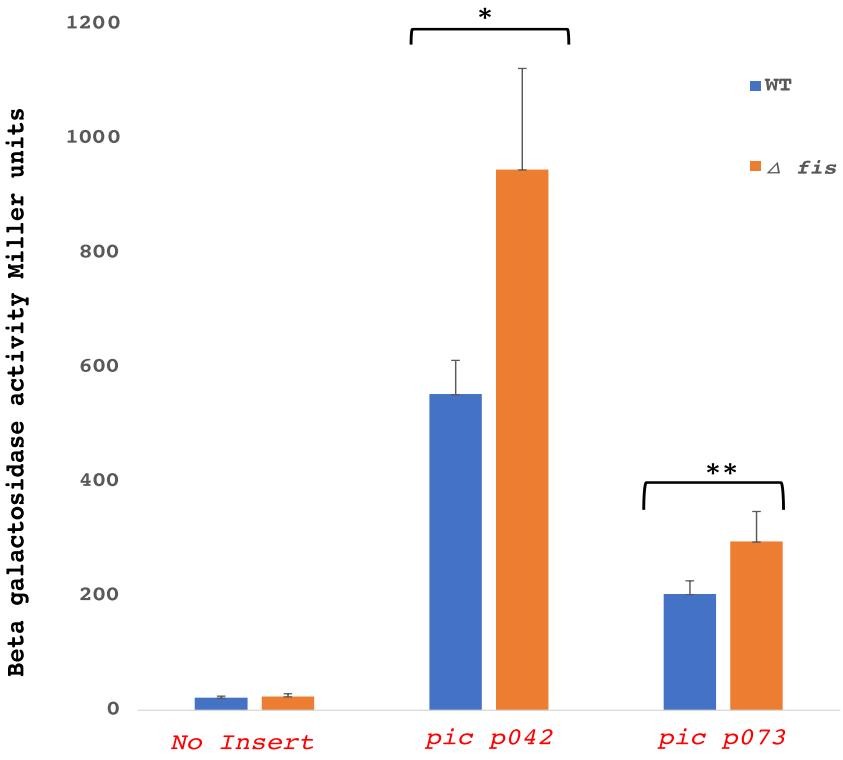


Figure 3.24: Regulation of the *pic* promoters by Fis

A. The figure shows an alignment of *pic* promoter DNA sequences in EAEC and UPEC. The two -10 elements are highlighted in yellow. The location of the CRP binding site is indicated by the green rectangle with bases corresponding to the consensus in bold. The Fis binding site

is indicated with bold underlined bases with an orange rectangle above it. The *pic* gene start codon is shown in red bold type. (Thompson *et al.*, 2003).

B. The figure shows measurement of ß-galactosidase levels in *E. coli* K-12 strain BW25113 Δ*lac* (blue bars) or BW25113 Δ*lac* Δ*fis* (orange bars) carrying pRW224 with the *pic p042* WT or *pic p073* WT promoters. A negative control, pRW224, with no insert, was included. Each promoter activity is the average of three independent ß- galactosidase assays and the error bars represent the standard deviation (STDEV). * Indicates $P < 0.02$, and ** indicates $P < 0.05$ using a student's *t*-test.

3.5.2 The influence of AggR on *pic* promoter activity

The master regulator of most virulence genes in EAEC, AggR, has been described (Morin *et al.*, 2013; Yasir *et al.*, 2019). Here, I wanted to study if and how AggR interacts with RNAP at the *pic* promoters. The base sequence of these promoters shows an interesting three putative AggR binding sites 5`-AnnnnnnTATC-3` along the *pic p042* and *pic p073* DNA sequence (highlighted in red in Figure 3.25). The first potential position AggR binding site is next to the upstream -10 hexamer element in both *pic p042* (5`-AATGCAGTAaC-3`) and *pic p073* (5`-AATGCAGTcaC-`3). The second potential site is located downstream and overlapping the downstream -10 hexamer element at both promoters, *pic p042 (*5`-gTCTGGTTATC-3`) and for *pic p073* (5`-gTCTGGTTATC-3`). The third potential site only occurs in *pic p042* (5`-ACTTTTATATC-3`) as it located in the 28 bases that are deleted in *pic p073.*

To test the impact of AggR on *pic* regulation, the recombinant *lacZ* expression vectors carrying the *pic* promoters, pRW224/ *pic p042* WT or pRW224/ *pic p073* WT, were transformed into *E. coli* K-12 strain M182 Δ*lac* containing pBAD/*aggR*. The resulting cells, with two plasmids, pRW224/ *pic p042* WT and pBAD/*aggR* or pRW224/ *pic p073* WT and pBAD/*aggR* were grown in LB medium either without or with 0.2% (*w/v*) arabinose to induce AggR expression. All cultures were incubated at 37˚C, with shaking to mid-exponential phase (OD_{650}= 0.4-0.6). Cultures were lysed and ß-galactosidase expression was measured.

Results are shown in Figure 3.26B. Measured ß-galactosidase levels are high for both promoters, *pic p073* and *pic p042,* in the absence of AggR induction. When AggR expression was induced, the level of ß-galactosidase levels falls by nearly 60% for *pic p042*, and nearly 45% for *pic p073*. These results indicate that AggR can repress each *pic* promoter, likely by blocking the promoter from RNAP. The key binding site of AggR could well be the second potential binding site: 5`-gTCTGGTTATC-3` in *pic p042* and (5`-gTCTGGTTATC-3`) *pic p073*, that was overlaps with the downstream -10 hexamer element promoter.

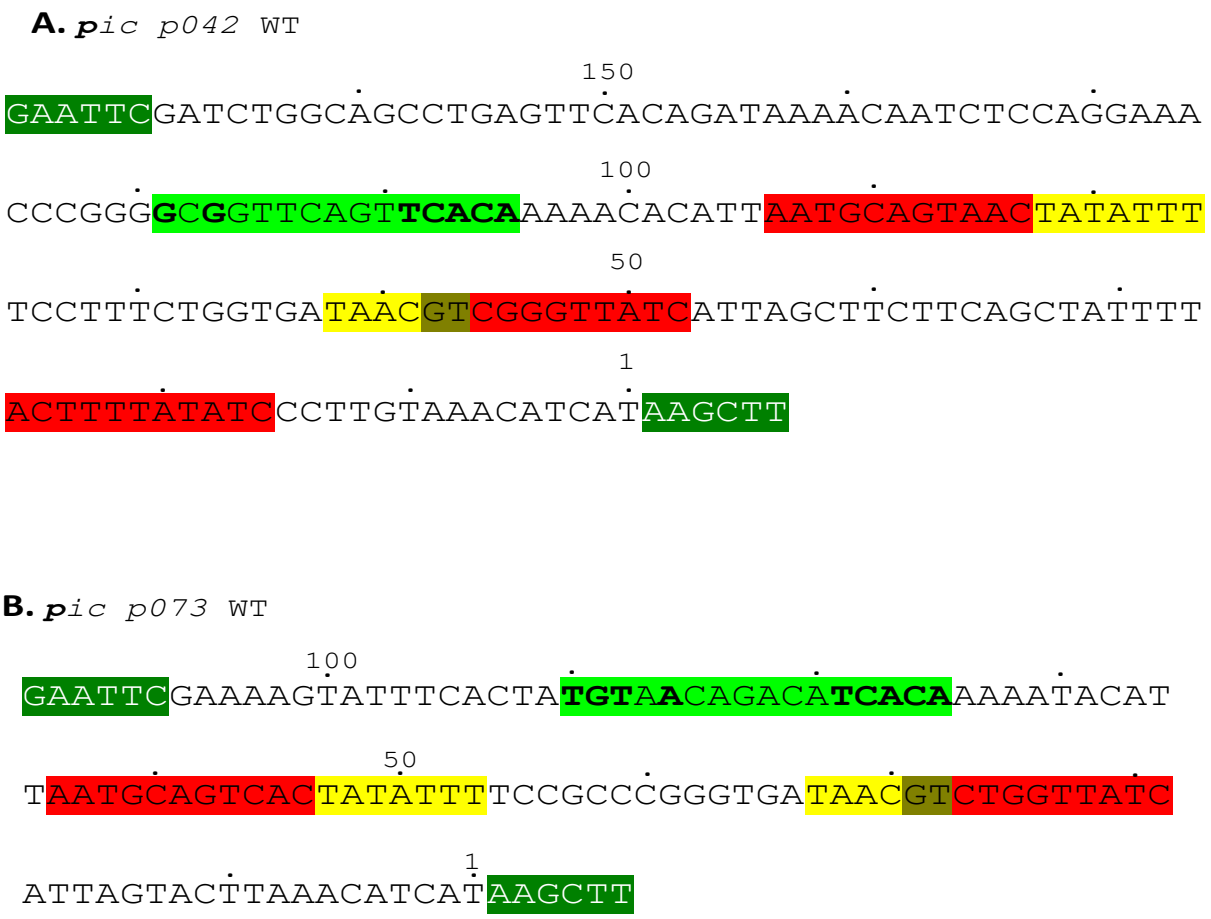


Figure 3.25: AggR site in the EAEC 042 and UPEC CFT073 *pic* promoter DNA sequences

A. The figure shows the base sequence of the *pic p042* wild type (WT) fragment. The CRP binding site is highlighted in light green. The yellow highlighter denotes the two -10 hexamer elements in *pic p042* promoter. Potential AggR binding sites are highlighted in red and olive-green highlighter represents the overlapped with the downstream -10 element. The sequences highlighted with dark green denotes the EcoRI and HindIII restriction sites. The numbering of the fragment starts from HindIII site.

B. The figure shows the DNA sequence of *pic p073* wild type (WT) fragment. The CRP binding site is highlighted in light green. The yellow highlighter denotes the two -10 hexamer elements in *pic p073* promoter. Potential AggR binding sites are highlighted in red, and the olive-green highlighter represents the overlap with the downstream -10 element. The sequences highlighted with dark green denote EcoRI and HindIII restriction sites. The numbering of the fragment starts from the HindIII site.

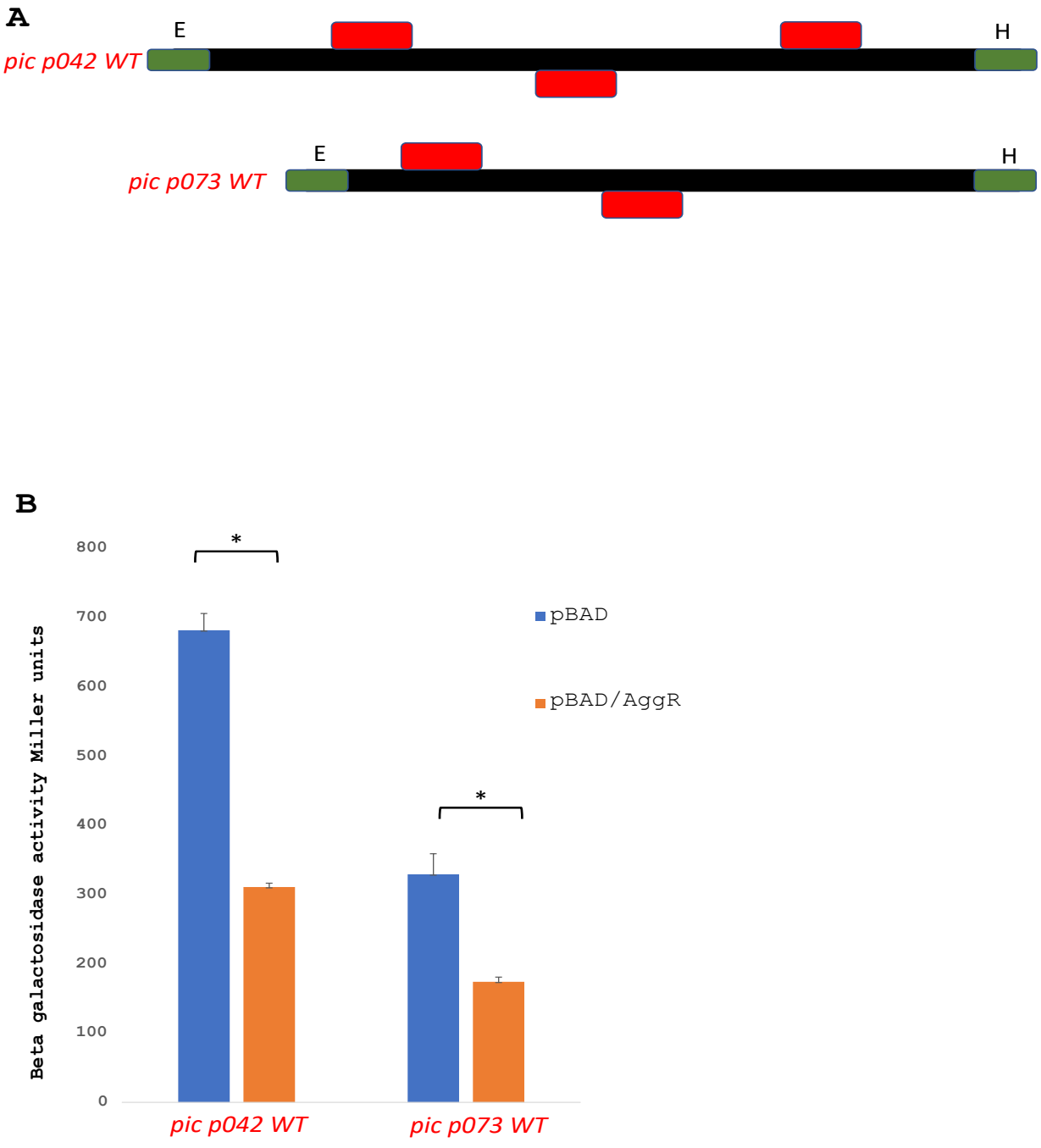


Figure 3.26: Impact of AggR on *pic* promoter activity

A. The panel shows diagrammatic representation of the position of the potential AggR binding sites at the *pic p042* and *pic p073* promoters. DNA sites for AggR are shown here as red rectangles. The green rectangles represent EcoRI (E) and HindIII (H) restriction sites.

B. The figure illustrates *pic* promoter activity, measured by ß-galactosidase activities, in wild type *E. coli* K-12 strain M182 Δ*lac* carrying two plasmids: the *lacZ* expression vector *(*pRW224/*pic p042* WT) and pBAD/AggR, or the *lacZ* expression vector *(*pRW224/*pic p073* WT) and pBAD/AggR. Blue bars indicate experiments without added (-) arabinose, whereas orange bars denote the inclusion of 0.2% (*w/v*) arabinose. Each promoter activity is the average of three independent ß- galactosidase assays and the error bars represent the standard deviation (STDEV). * Indicates $P < 0.005$ using a student's *t*-test.

3.6 Discussion

To understand the full picture of infectious pathogenesis and how pathogens attack their host, we need to focus on the early stages of infection. Human organs exposed directly or indirectly to the external environment are usually covered with slippery fluid mucus (Kim and Khan, 2013). The main role of this is to protect the organs from invasion by harmful microbes (Kim and Khan, 2013). EAEC and UPEC cause infection in two different tracts in the human body (the digestive tract and the urinary tract). The level of Pic expression differs between these two pathogens due to the difference in mucus thickness between the intestines (approximately 50 μm) and bladder (approximately 20 μm) (Johansson *et al.*, 2011; Zhang *et al.*, 2023). These pathogens produce Pic enzyme in the early stage of their pathogenesis, which helps to breakdown mucin and expose host cells (Henderson *et al.*, 1999). For this reason, I aimed to understand the regulation of *pic* gene expression, focussing on its promoter and transcription factors.

Genetic analysis of the EAEC and UPEC *pic* promoter DNA base sequences show the presence of a DNA site for CRP binding, located 21 bp and 41 bp upstream of promoter -10 hexamer element. My experiments showed that the *pic* gene in both EAEC and UPEC carries a CRP-activated promoter. The first prediction of multiple promoter elements in the *pic* promoter region was reported by Behrens *et al.* (2002). In their study, they suggested that the presence of multiple promoters in the *pic* regulatory region might be due to the need for mucinase in the early stages of infection.

A study done by MSc project student, Joseph Ingram, to locate the DNA site for CRP at the *pic p042* promoter region had shown that knocking out the underlined **T** base in the CRP binding element 5`-G**G**C**G**GTTCAGT**T**C**A**C**A**-3` in *pic p042* caused a significant reduction in promoter activity, indicating the position of CRP binding site at position -119 upstream from *pic* gene (Alhammadi *et al*., 2022). The importance of this element is supported, here, by the experiment shown in Figure 3.19, in which enhancing the CRP consensus resulted a boost in promoter activity. The mode of CRP-dependent activation at the *pic p042* promoter was tested by assessing the contribution of the upstream 5`-T**A**TATT-3` and downstream 5`-T**A**AC**G**T-3` -10 hexamer elements. The results show that RNAP prefers the downstream promoter in *pic p042,* which is subject to Class I activation by CRP. Thus, the DNA site for CRP is located at position -60 upstream from the TSS (+1), Figure 3.6. The promoter activity was lowest when both -10 elements were knocked out, but there is some expression that could be due to pervasive transcription (Lybecker *et al.*, 2014).

In the case of *pic p073*, a study done by MIBTP student, Gurdamanjit Singh, showed that mutating the **T** base in the CRP binding site 5`-**TGT**AACAGACA**T**C**ACA**-3` resulted in a complete knockout of promoter activity, thereby locating the CRP binding site and showing the need for CRP to recruit RNAP to *pic p073*. CRP activation classes at *pic p073* were tested using ß-galactosidase assays. Investigations showed that *pic p073* is predominantly Class II, and RNAP prefers the upstream promoter. Here, the bound CRP overlaps with the promoter -35 element, as shown in Figure 3.12. Tests with CRP mutants showed that RNAP needs both AR1 and AR2 for activation at *pic p073*. Hence, here, CRP interacts with RNAP by two contacts, AR2 via αNTD and AR1 via αCTD, driving transcript initiation (Browning and Busby, 2004).

The significant differences between *pic p042* and *pic p073* prompted me to examine the differences in promoter elements. My results show that different in promoter elements lead to promoter activity differences. Note that, TSS in bacterial transcription usually strats at the seventh or eighth base downstream from -10 element, however, many factors influence this selection, such as nucleotide levels, and the DNA sequence in region of TSS, for instance, in case of the *E. coli fis* promoter position 8 with nucleotide preference ATP = GTP ≫ CTP ≥ UTP is preferred (Walker and Osuna, 2002).

The presence of the GC marker in different positions in *pic p042* and *pic p073* has shown that the increase of GC bases in the UP-element sequence influenced the binding of RNAP subunit alpha-CTD. Thus, the result in Figure 3.17 showed an increase in promoter activity when the GC bases were replaced with AT bases suggesting the preference for UP-element sequences to have AT bases for alpha-CTD binding site (Estrem *et al.*, 1998; Aiyar *et al*., 1998). Another interesting point to study was the deletion block in *pic p073* and how it influences the *pic p042* activity. Figure 3.18 shows that transplanting the 28 base pair deletions from *pic p073* into *pic p042* result in loweing of expression but no change in promoter preferences. This is likely to be due to a shorter 5`-UTR. This situation has been seen in case study of *Aggregatibacter actinomycetemcomitans* where 528 base pair deletions in the 5'-non-coding region of the leukotoxin operon produces lower expression of leukotoxin (Feuerbacher *et al*., 2011).

The biggest difference between the EAEC and UPEC *pic* promoter is the location of the transcript start and this has consequences for *pic* gene translation. My data suggest that the *pic p073* transcript is sub-optimally translated, presumably due to the presence of an inhibitory element. Here, the *pic*

p073 5'UTR seems to have a challenging structure. Figure 3.27 shows the RNA secondary structure prediction of 5'UTR in *pic p073.* Here, the structure showed a stem-loop if the upstream is promoter chosen and this appears to have an impact on translation of the gene.

The other factors were tested on *pic* promoter regulation were the influence of AggR and Fis on *pic* expression. All results with these factors show a repression role on *pic* promoter regulation. The reason behind this is that the position of these regulators comes after the promoter, as it shown in Figure 3.24A (for Fis) and Figure 3.26 (for AggR). The interesting point here is that *pic p042* is more susceptible to repression by Fis and AggR than *pic p073*, and this is likely due to the position of the -10 element with respect to the AggR and Fis binding sites.

It is crucial point to understand how pathogenic strains begin infection. Understanding this might help us to target the first phase of pathogenesis and prevent its continuation. In the case of pathogenic strains of *E. coli* such as EAEC and UPEC, both express Pic in the early stage of pathogenesis which targets the mucus that covered inner tracts in the host. My result suggests that the *pic* promoter from EAEC and UPEC is an ambiguous CRP-dependent promoter.

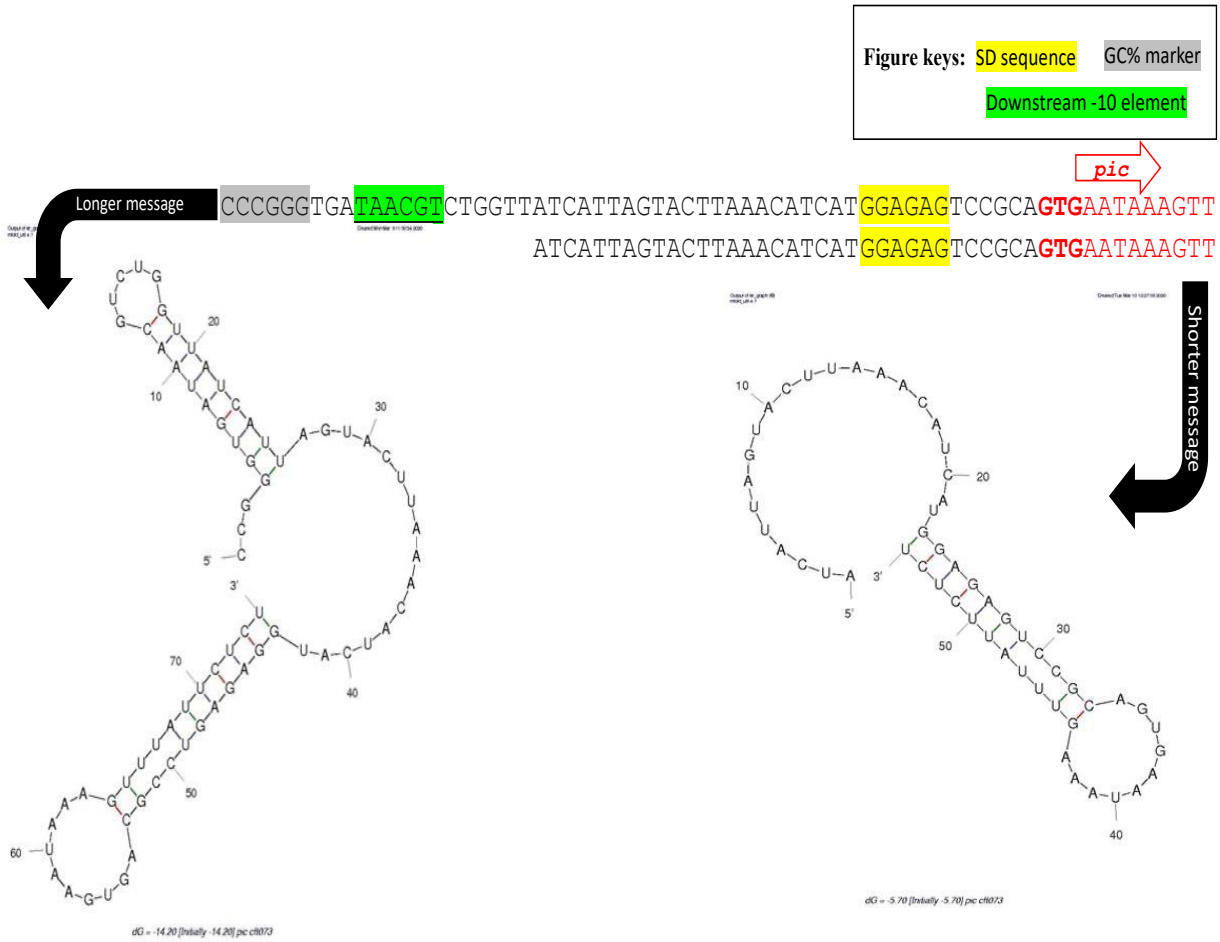


Figure 3.27: RNA secondary structure prediction of *pic* p073 expression

Figure shows a prediction of RNA secondary structure in case of using the upstream promoter in *pic p073* (Left-hand) or downstream in *pic p073* (right-hand). SD sequence highlighted in yellow in bases and structure.

Chapter 4

Identification of CRP binding sites across the EAEC 042 genome

4.1 Introduction

Some *E. coli* strains have evolved the ability to colonise the human gut, and some of these strains cause disease. Hence, EAEC strain 042 is a diarrhoeagenic human pathogen that infects human intestines. Infection by this pathogen starts with the colonisation of intestinal mucosa, followed by the expression of various virulence determinants. Genomic analysis of commensal *E. coli* K-12 and EAEC strain 042 showed a high similarity, but with some differences. Some genes have been deleted from the EAEC 042 genome, and others have been acquired that might be involved in virulence. Figure 4.1 shows a comparison between the genomes of *E. coli* K-12 strain MG1655 and EAEC strain 042.

Since the evolution of EAEC 042 required the acquisition of virulence determinants genes, it is important to study their expression and knowledge from such a study could help to develop treatments for infection that are independent of antibiotics. Yasir *et al.* (2019) had studied the virulence genes of the AggR regulon using RNA-sequencing and this led to understanding of the regulation of three virulence factors: attachment adherence fimbria (AAF), anti-aggregative protein dispersin (Aap) and a dedicated type I system (T1SS). The RNA-seq approach revealed AggR-regulated genes both on the EAEC 042 chromosome and the virulence plasmid (pAA2). However, some virulence genes are not involved in the AggR regulon, for example, *pic* which encodes the mucinase, and *pet* which encodes a cytotoxic protease. The regulation of *pic* and *pet* requires the global regulator cAMP receptor protein (CRP), since these virulence determinants are needed in conditions of nutritional stress (Rossiter *et al.*, 2011; Alhammadi *et al.,* 2022).

Due to the need for a full understanding of EAEC strain 042 pathogenesis, the regulation of the most critical virulence factors needs to be studied. Here, I contribute with a genome-wide study of the influence of the global regulator, CRP, in EAEC strain 042. My aim was to find other virulence determinants that are controlled by CRP, in addition to *pic* and *pet*. A study was undertaken to determine the binding profile of CRP across the EAEC genome, focusing on the location of the CRP-binding sites within the chromosome and plasmid pAA. To achieve this goal, I used chromatin immunoprecipitation (ChIP) combined with next generation sequence to study how CRP copes with the evolution of EAEC; and if there are other virulence genes besides *pic* and *pet* involved in the CRP regulon.

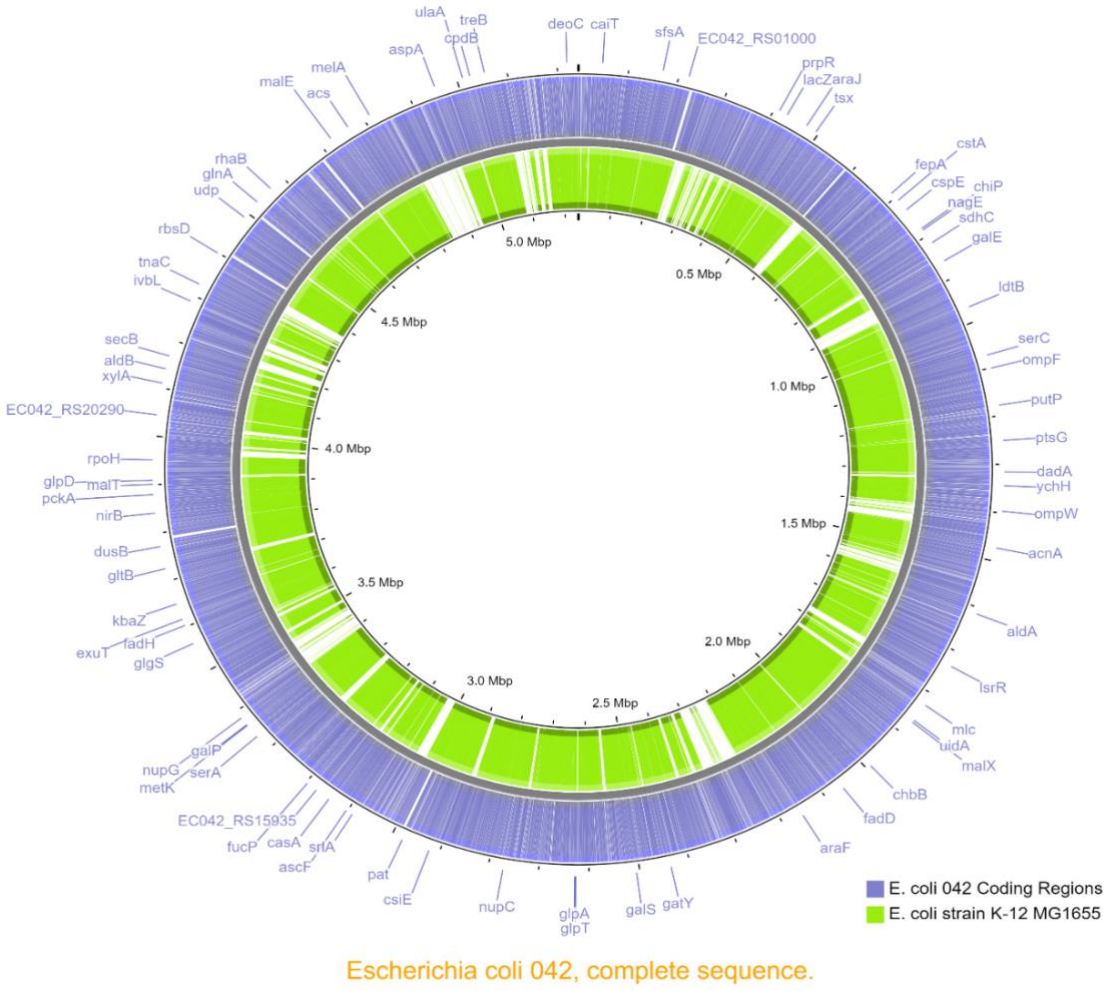


Figure 4.1: Alignment of the *E. coli* K-12 genome with the EAEC 042 genome

The figure illustrates a comparison of the *E. coli* K-12 strain MG1655 genome (green inner circle) and the EAEC 042 genome (blue outer circle) using the BLAST tool of the CGview server (Grant and Stothard, 2008). Well-known genes of the CRP regulon are named outside of circle.

4.2 The influence of CRP on EAEC strain 042 biofilm formation

My first aim was to assess the broad role of CRP on the biofilm formation of EAEC strain 042. Here I used two modified strains: EAEC 042DFB and a derivative with a *crp* gene knock out (EAEC 042DFB Δ*crp*) (Table 2.1). Firstly, the recombinant plasmids pDCRP, pDCRP AR1$^-$ AR2$^-$and pDU9 (empty vector) were transformed into *E. coli* S17`1. Then the transformed *E. coli* S17`1 was grown overnight on selective agar and the EAEC 042DFB and EAEC 042DFB Δ*crp* were grown on LB agar in preparation for a conjugation experiment. The following day, the transformed *E. coli* S17`1 was mixed with EAEC 042DFB and EAEC 042DFB Δ*crp* separately on agar plates, and these trans-conjugant agar plates and all other control plates, were incubated at 37 °C for 5 hours. Then the conjugated EAEC DFB 042; conjugated EAEC DFB 042 Δ*crp;* and *E. coli* S17'1 was streaked onto minimal salt agar with appropriate antibiotics. All plates were incubated at 37 °C overnight.

The conjugated strains (EAEC DFB 042 and EAEC DFB 042 Δ*crp*) that carry pDCRP, pDCRP AR1$^-$ AR2$^-$ or pDU9 were grown overnight at 37˚C, and the following day, cultures were diluted 1:100 in DMEM medium supplemented with tetracycline and ampicillin, and incubated for 1 hour at 37˚C. Then each strain was added to 8 wells of a 96-well microtitre plate and incubated for 17 hours to allow cells to form biofilm. Then biofilm formation was assessed using the crystal violet biofilm assay by measuring the blue dye absorbance with a Labsystem Multiskan MS plate reader at OD595 nm.

Figure 4.2 illustrates the measured absorbance for EAEC 042DFB and EAEC 042DFB Δ*crp* with plasmids pDCRP, pDCRP AR1$^-$AR2$^-$ or pDU9. The wells containing EAEC 042DFB carrying pDCRP show higher absorbance than the wells that contains EAEC 042DFB with inactivated CRP (pDCRP AR1$^-$AR2$^-$). The reason for the low absorbance in the case of EAEC 042DFB with pDCRP AR1$^-$AR2$^-$ must be the trans-dominance of inactive CRP blocking the activity of the functional CRP. These data argue strongly that functional CRP is essential for biofilm formation in EAEC 042. Hence, as expected, with EAEC 042DFB Δ*crp*, the absorbance was high only when the functional CRP is present.

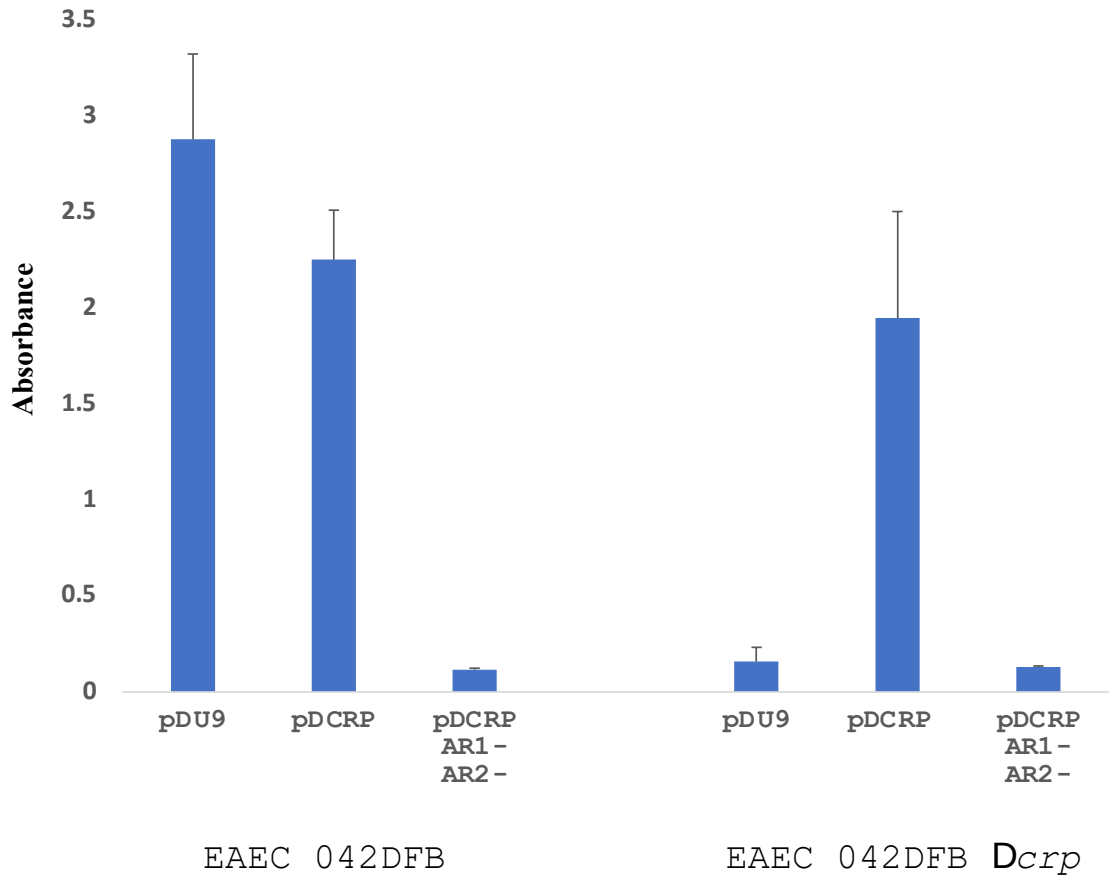


Figure 4.2: Effect of CRP on biofilm formation

Biofilm formation of EAEC 042 DFB (left) and EAEC 042DFB Δ*crp* (right) that carry pDCRP, pDCRP AR1^{-}AR2^{-} , or pDU9 (empty vector) as negative control, grown in wells. The bar chart illustrates blue dye absorbance of each well using the plate reader at OD595 nm. The error bars denote the standard deviation (STDEV) of the technical repeat of each column.

4.3 ChIP-seq analysis of CRP binding sites across the EAEC strain 042 genome

Chromatin immunoprecipitation (ChIP) coupled with high-throughput sequencing was used to map CRP binding sites across the EAEC strain 042 genome. This method was done in collaboration with Dr. James Haycocks. For the immunoprecipitation, mouse mono-clonal anti-CRP, provided by Dr. Douglas Browning, was used here to study the binding of CRP under the native condition. Here, I used the wild type strain of EAEC 042, genome reference number FN554766.1 (Chaudhuri *et al.*, 2010), which was grown to mid-exponential phase in LB medium at 37˚C, with shaking. When the culture reached OD_{600}=0.313, cells were fixed with formaldehyde and harvested and lysed with a Bioruptor sonicator. Clear cell lysates were produced and subjected to immunoprecipitation with the anti-CRP antibody (treated samples) alongside control no-treatment samples. Afterwards, the DNA products were washed using AMPure magnetic beads. DNA libraries were generated using an NEB Next Ultra II DNA Library Prep Kit, and then the library quality was checked by Tapesation, quantified by qPCR using NEB Next library quantification kit, and sequenced on an Illumina HiSeq sequencing platform.

The data were analysed using several software packages: the number of reads mapped to each position within the EAEC 042 genome was estimated by MACS2 in Galaxy (Afgan *et al.*, 2018); the traces were visualized, and plotted on the DNA sequence using the Artemis genome browser (Carver *et al.*, 2011); and CRP binding motifs within peaks were obtained from the MEME SUITE (Bailey *et al.*, 2009). I compared the EAEC 042 data with *E. coli* K-12, using the protein sequence provided in the EAEC 042 genome profile, NCBI Blast; and the CRP regulon viewed in the RegulonDB databases (Santos-Zavaleta *et al.*, 2019; Altschul *et al*., 1990).

4.3.1 A catalogue of CRP binding sites in the EAEC strain 042 chromosome

From the ChIP-seq data, I identified 322 CRP binding peaks on the EAEC 042 chromosome, and 10 CRP binding peaks on the pAA plasmid. These peaks, plotted onto the EAEC 042 chromosome and the pAA plasmid, are shown in Figure 4.3 (A. EAEC 042 chromosome DNA plotter and B. EAEC 042 plasmid DNA plotter).

The sequence around each peak centre, provided by MACS2, was extracted from 50 bp upstream to 50 bp downstream of the peak centre point, resulting in a 101 bp search window for CRP binding peak sequences. These sequences were applied to the MEME SUITE to obtain likely CRP binding motifs. The results revealed 322 loci where CRP might bind to a target that resembles the consensus DNA binding site for CRP (Appendix B). Five of these loci had two recognisable CRP binding motifs: *mlc, sdhC, EC042_3992, gntR, aslA*. A total of 10 CRP binding motifs were located between divergent gene: *EC042_2583/fadL*, *putA/putP*, *EC042_2480/EC042_2478*, *nagE/nagB*, *aspA/fxsA*, *malX/malI*, *plsB/dgkA*, *EC042_4888/sltY*, *glk/ EC042_2608* and *dacD/sbcB*.

Next, the EAEC 042 genes listed in Appendix B were compared with *E. coli* K-12 strain MG1655, using NCBI Blast protein and RegulonDB databases. This showed that 291 of the EAEC targets from the CRP ChIP-seq assay are also present in *E. coli* K-12 strain MG1655, whereas 31 genes were specific to EAEC 042. These are: *kpsMII, EC042_0414, EC042_3143, EC042_0536, EC042_0225, EC042_0224, EC042_4604, EC042_3975, virK, EC042_4012, EC042_4187, EC042_3970, EC042_4606, EC042_2431, aadA1, EC042_3998, EC042_4746, EC042_0305, EC042_0769, EC042_4744, EC042_3956, EC042_3465, EC042_3075, tetR, EC042_4011, EC042_0363, EC042_4524, EC042_2670, EC042_3190, set1A* and *EC042_2130* (selected CRP targets are shown in Figure 4.4). This points to new genes that may have been recruited into the CRP regulon in EAEC.

The classification of the genes associated with CRP binding was obtained from gene ontology (GO) based on the protein classes (Mi *et al.,* 2019). Figure 4.5 shows these categories. A total of 10 genes are involved in nucleic acid metabolism (DNA and RNA). The metabolite interconversion enzyme category recorded 30 genes. This category includes genes that encoded enzyme products for catalysing the hydrolysis of a variety of bonds, for example peptides (*serC*) or glycosides (*glK* and *maa* and *mtlD*). A total of 20 genes encoding transcription factors were found including GalS, AaeR, LysR, RpoH, LrhA, MalI, TdcA, CytR, DhaR, Mlc, PdhR and CecR. The categories with the lowest number of genes, with one gene in each category, were transfer/carrier proteins (*ptsH*) and storage proteins (*ftnB)*. Categories with two genes are chaperones (*ppiA* and *ppiA*), scaffold/adaptor proteins (*clpS* and *mobB*), and protein-binding activity modulators (*engB* and *glnB*). In addition, 26 genes were involved in the transporter category, such as *ydeA*, a sugar efflux transporter*, cutC*, which is important in copper homeostasis, and other proteins involved in transporter such as *malE, malK* and

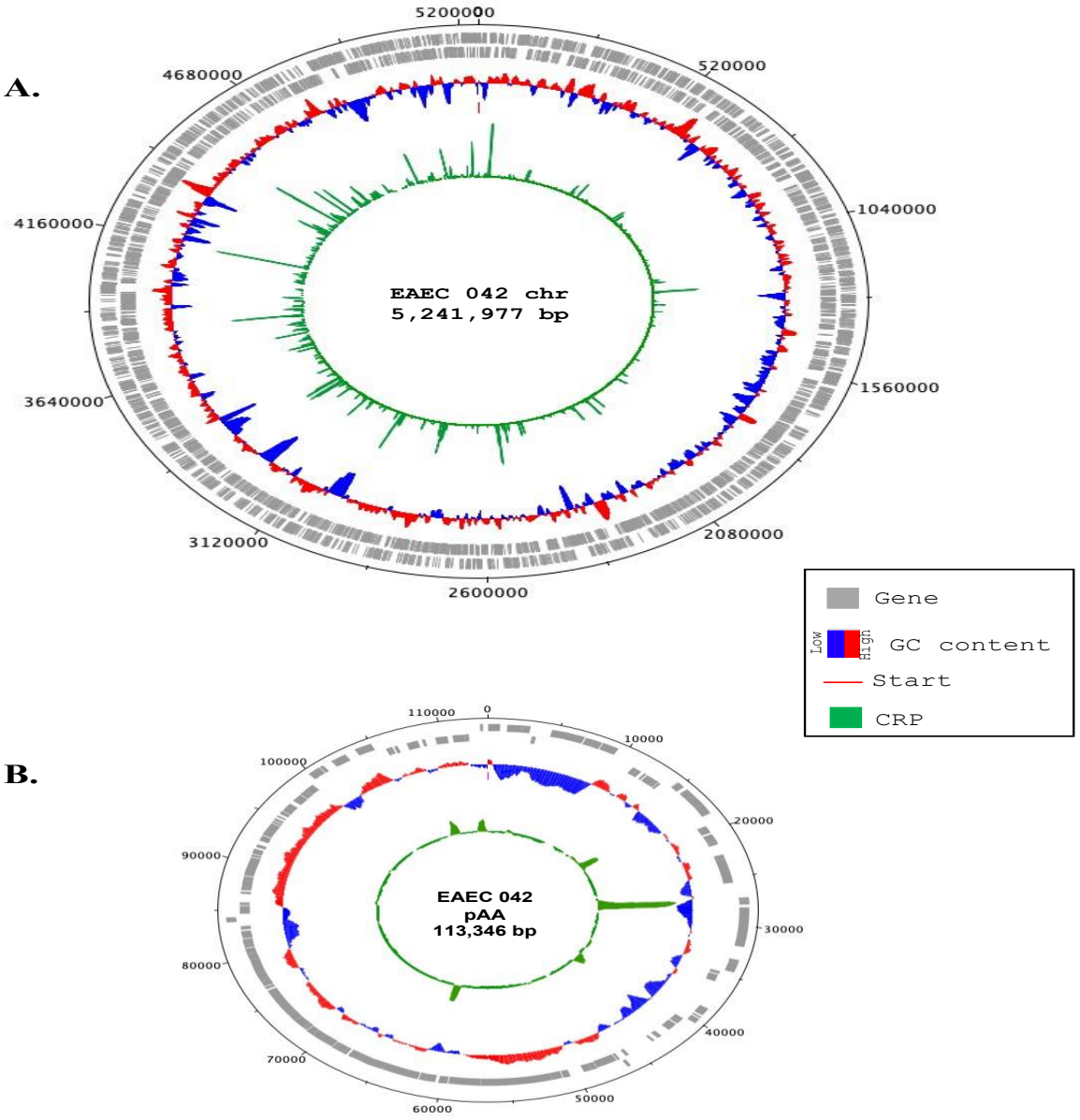


Figure 4.3: CRP binding across the EAEC 042 genome determined by ChIP-seq analysis

In the plots, **A** and **B**, genes are indicated with grey blocks located in the outer two tracks. GC content is shown in track coloured based on to GC percentage (red as higher than the average of GC content whereas blue represents a lower level than the average of GC content). Coordinate zero of the genomes is shown by a red line. The green tracks indicate the binding by CRP, determined from ChIP-seq data. The DNA plotter was obtained from the Artemis genome browser (Carver *et al.,* 2011).

nepI. The translation-associated protein category contains 10 genes, including *ettA*, *raiA* and *bipA*. The protein modifying enzyme category contains 5 genes, encoding a protease (*yegQ*), a protein phosphatase (*yedP*), a serine protease (*decD*) and two metalloproteases (*nlpD* and *mtfA*). 79 genes were categorized as 'other' which includes 43 genes with unknown function, 5 pseudogenes and the 31 EAEC 042-specific genes.

All targets bound by CRP in the EAEC 042 chromosome, from this experiment, are listed in Appendix B and are compared with the CRP targets in the *E. coli* K-12 strain MG1655 that are listed in the RegulonDB database. 93 of the genes common to both *E. coli* K-12 strain MG1655 and EAEC 042 were listed in RegulonDB database as targeted by CRP, whereas 160 genes are new in the CRP regulon, indicating that more investigation is needed here. However, for 69 genes, no comparison was able to be made with genes in the *E. coli* K-12 strain MG1655 as these genes encoded unknown proteins.

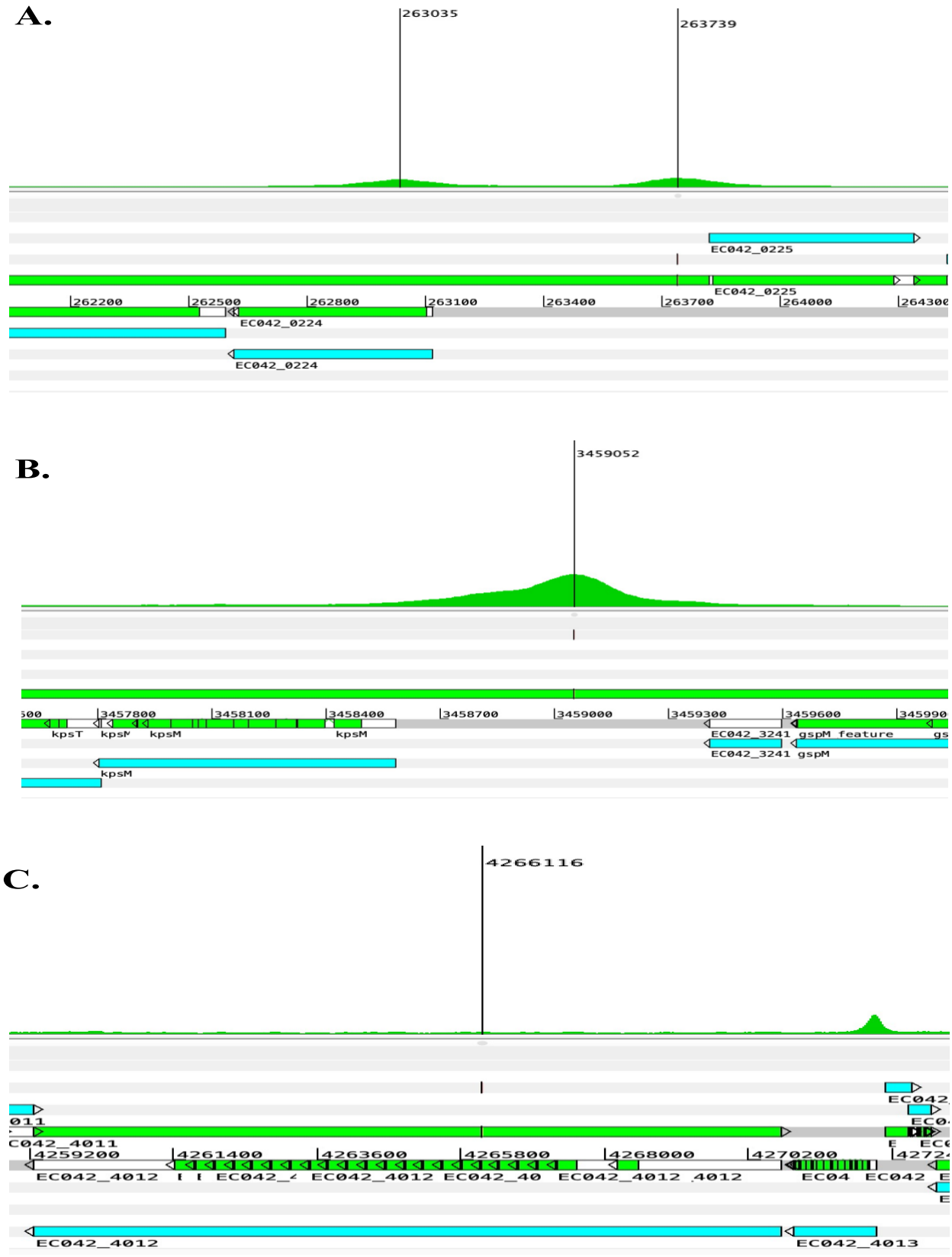


Figure 4.4: Binding of CRP at selected EAEC 042-specific genes (continued)

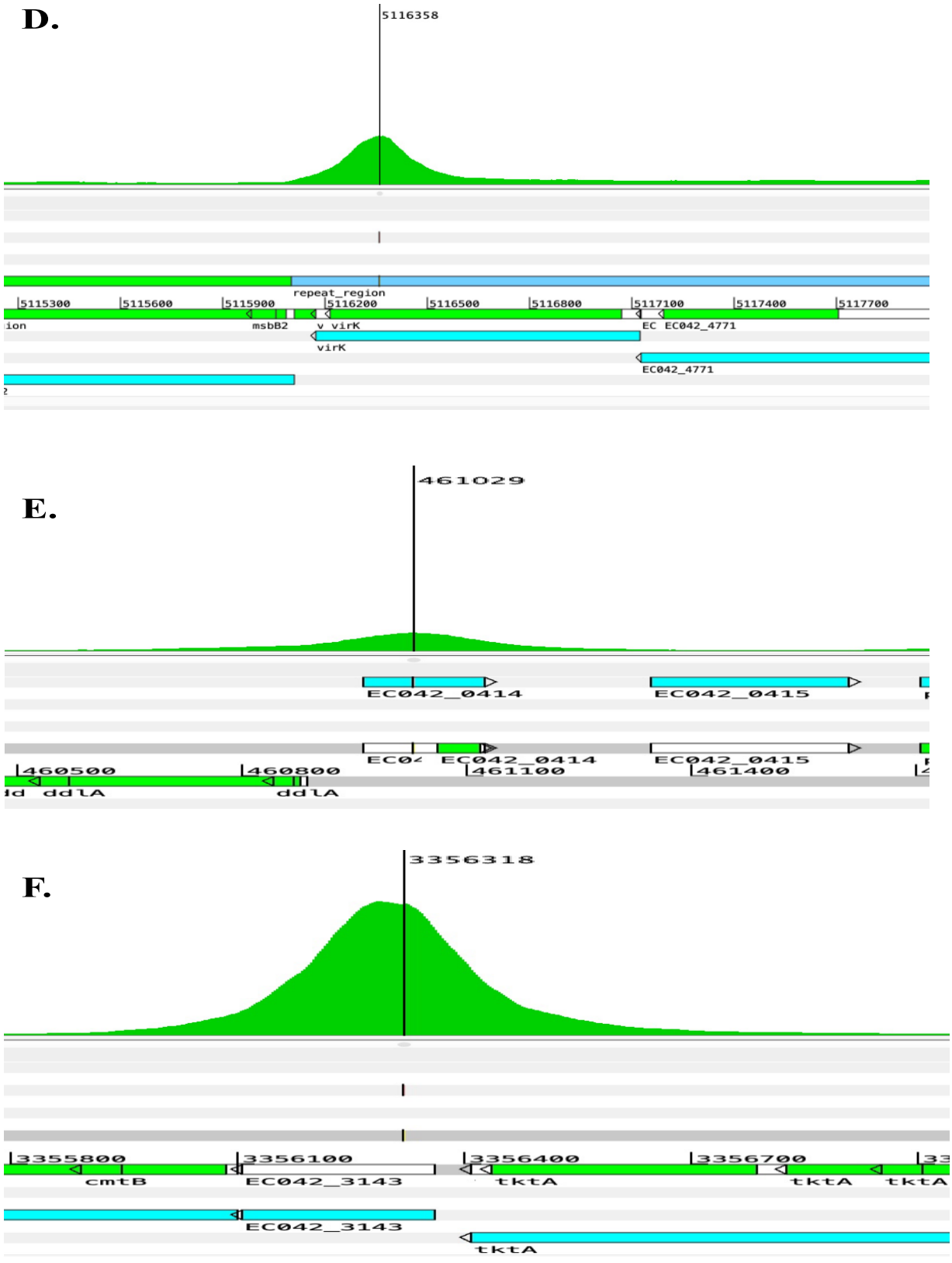


Figure 4.4: Binding of CRP at selected EAEC 042-specific genes (continued)

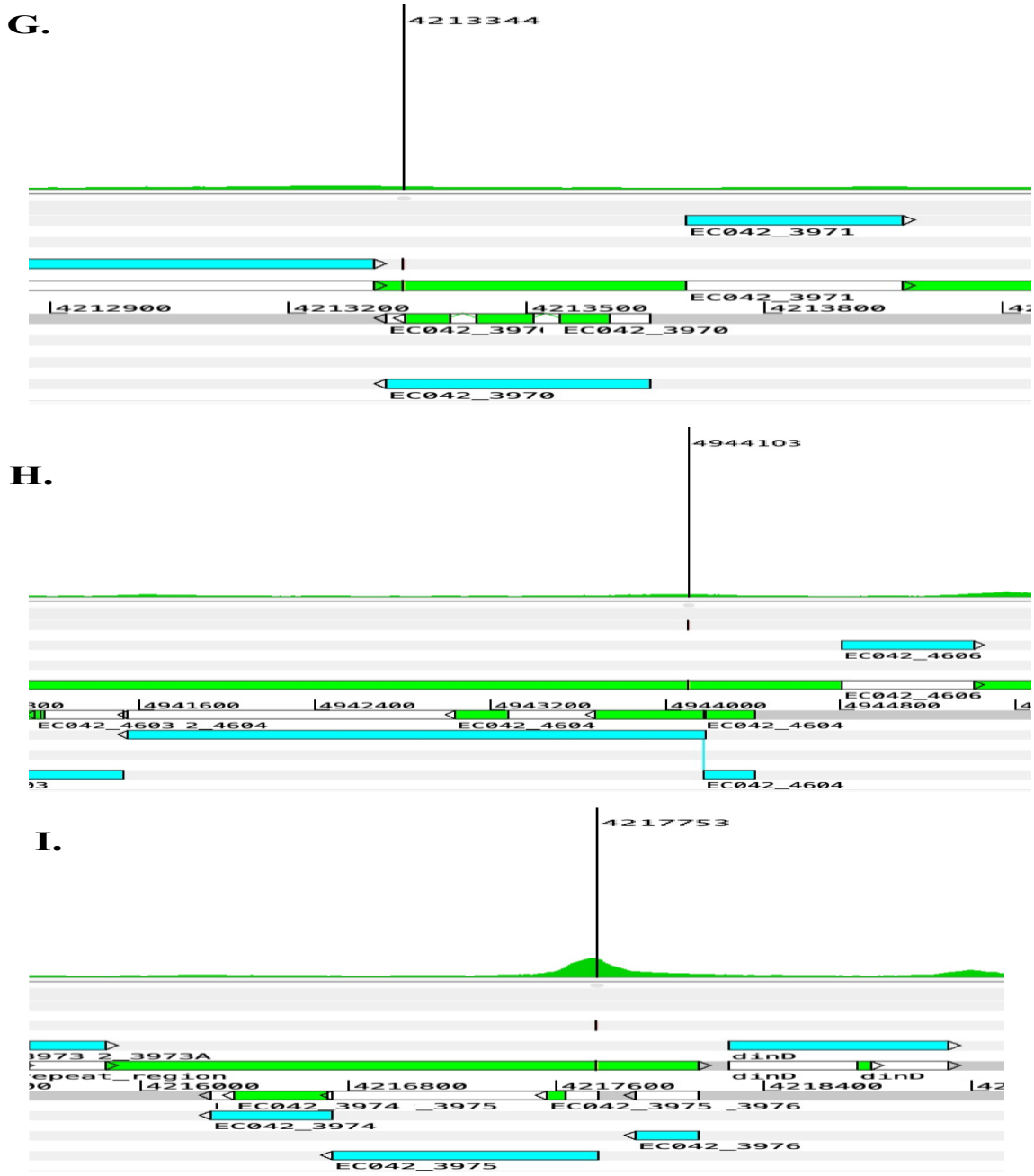


Figure 4.4: Binding of CRP at selected EAEC 042-specific genes

The figure shows the experimentally observed peaks of CRP binding at selected EAEC 042-specific loci, associated with: **A.** divergent genes (*EC042_0225 EC042_0224*), **B.** *kpsMII,* **C.** *EC042_4012,* **D.** *virK,* **E.** *EC042_0414,* **F.** *EC042_3143,* **G.** *EC042_3970,* **H.** *EC042_4604,* **I.** *EC042_3975*. Genes are shown as light blue blocks and CRP binding is shown as green peaks. Green blocks denote the position of the genes within the contig. Pictures were obtained from the Artemis genome browser (Carver *et al*., 2011).

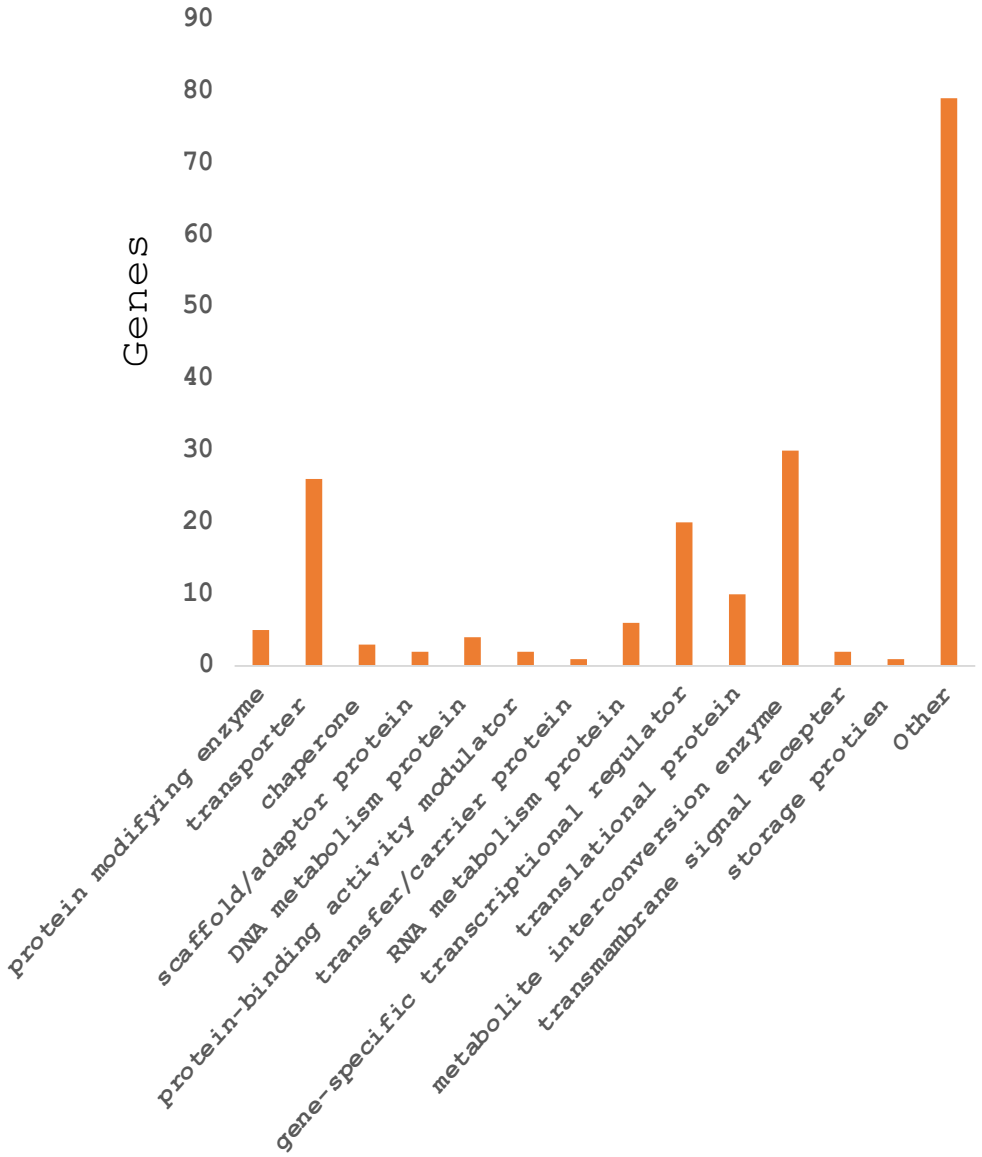


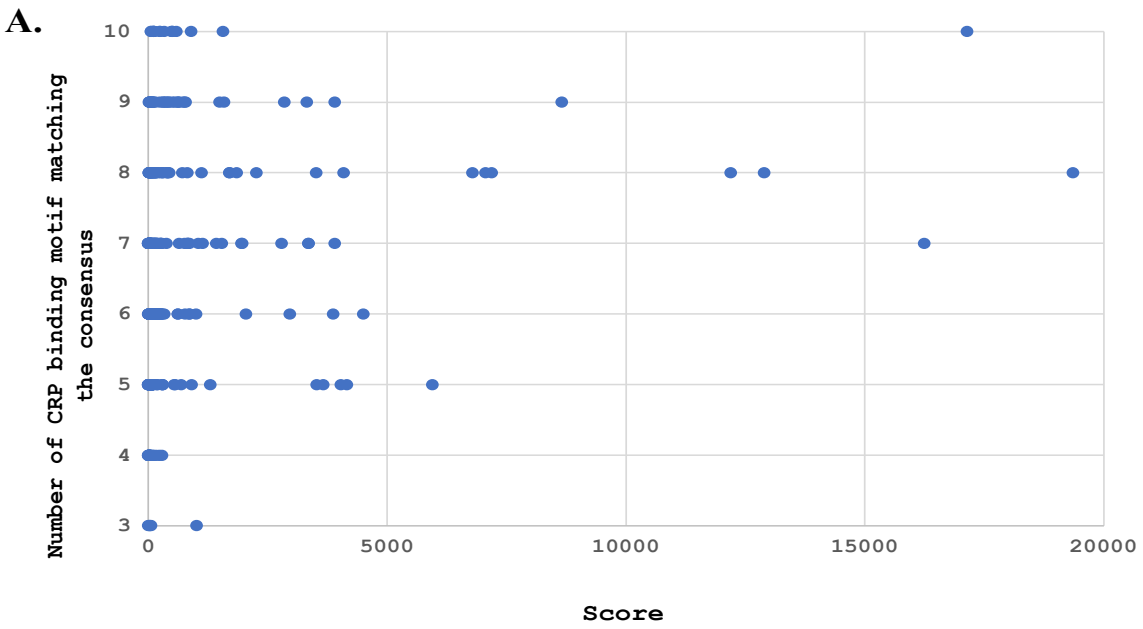
Figure 4.5: Gene categories of genes targeted by CRP in EAEC 042

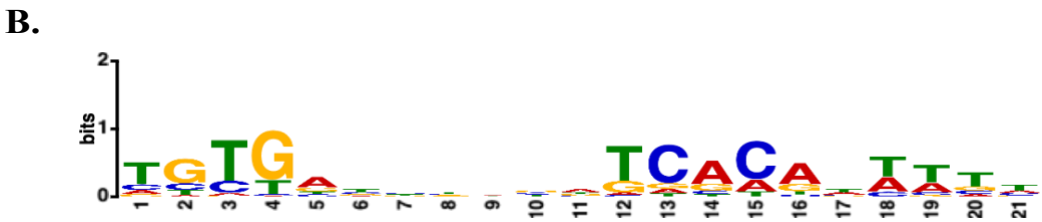
The figure shows the categories of genes that are targeted by CRP on the EAEC 042 chromosome, determined from ChIP-seq data. Categories are based on the general functions of genes. A total of 12 categories, shown on the x-axis, are correlated to the number of genes in each category showed on the y-axis. The 'Other' category denotes genes that encoded a protein with unknown function.

4.3.2 CRP binding site quality and location on the EAEC 042 chromosome

The purpose here was to study the relationship between the degree of CRP binding, judged by ChIP-seq analysis, and the 'quality' of the proposed DNA site for CRP at each location. Figure 4.6A shows the correlation between CRP ChIP-seq signal scores (*i.e.*, the number of reads) and the number of bases in each site that match to the known consensus DNA site for CRP (5`-**TGTGAN**$_6$**TCACA**-3`; Figure 4.6A). Surprisingly, the CRP ChIP-seq signals correlate poorly with the CRP target sequence consensus: the calculated correlated coefficient is negative, -0.12. Many CRP motifs, obtained from MEME SUITE, showed a highly conserved CRP motif (with *E*-value 2.2e-026), but many of the CRP binding sequences, marked towards the left of Figure 4.6A, failed to show any conserved CRP motif. For example, the highest score for a CRP ChIP-seq signal is associated with *fixA* (around 16000 reads) but the CRP binding site in the regulatory region of *fixA* is 5`-**TGT**TTTATAGA**TCAC**C-3` that corresponds to the consensus only at 7/10 positions. Note that a study by Buchet and colleagues demonstrated the involvement of CRP in the regulation of *fixA* (Buchet *et al*., 1998). Another striking example is the EAEC 042-specific gene *EC042_0225* gene that has a 10/10 CRP motif in its promoter region, 5`-**TGTGA**GCCGCA**TCACA**-3`, but the score for the reads betrays a weak signal (around 100 reads only) indicating this gene as an interesting subject for additional investigation.

The position of each CRP motif in intergenic and intragenic regions was calculated by its distance to the translation start site of the nearest genes. A total of 74 CRP motifs were found to locate within a coding region, indicating that CRP could be involved in long-distance regulation of the downstream gene, or alternatively, that CRP plays a role in chromosome organization. However, 248 DNA sites for CRP are positioned in non-coding intergenic regions, which suggests a potential regulatory role at these locations. Figure 4.6C illustrates the distribution of bound CRP between coding and non-coding regions, according to the ChIP data: 23% of sites are in intragenic regions, whilst 77% are found in intergenic regions. This distribution suggests that CRP is more involved in gene regulation than in DNA sculping (NAP) roles.





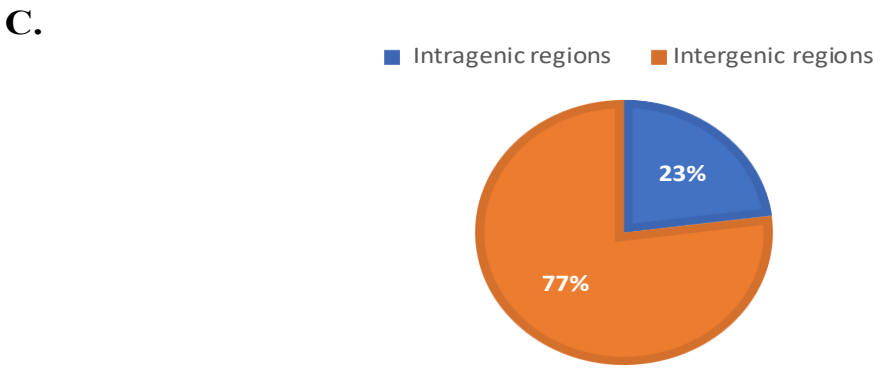


Figure 4.6: CRP target analysis

A. The panel illustrates the correlation of score reads with number of matches to the CRP consensus 5`**TGTGA**NNNNNN**TCACA**`3 (correlation coefficient= -0.1185289).

B. CRP binding sequence logo deduced from MEME software analysis of identified targets from CRP ChIP-seq on the EAEC 042 chromosome (Bailey *et al.*, 2009).

C. The pie chart represents the distribution of bound CRP between intergenic and intragenic regions on the EAEC 042 chromosome, deduced from ChIP-seq analysis.

4.3.3 CRP binding sites within the pAA plasmid of EAEC strain 042

The EAEC 042 pAA plasmid was included in the CRP ChIP-seq data to locate CRP binding sites on the virulence plasmid, using the reference of EAEC 042 pAA FN554767 from the NCBI database. A total of 10 CRP binding peaks were recorded by MACS2. As before, we applied the MEME SUITE software to find a consensus CRP motif in 101 bp windows centred on ChIP signals. Table 4.1 lists these CRP targets. Figure 4.7 illustrates the peaks using the Artemis genome browser (*pet*, *virK*, *aafA*) and Figure 4.8 shows the CRP motif found by MEME SUITE. Strikingly, the motif poorly resembled the usual conserved motif for CRP binding (*E* value 2.7e+005). The reason behind this might be that CRP has not evolved to be a major regulator of plasmid-encoded genes. However, there is one known promoter, targeted by CRP, on the EAEC 042 plasmid pAA, that drives expression of the *pet*, which was studied by Rossiter *et al.* (2011). Therefore, the CRP site in the *pet* promoter region was used as a positive control for the other nine CRP binding peaks.

Interestingly, the EAEC 042 genome encodes two versions of the *virK* gene, one in the chromosome, upstream of the *msbB2* gene, and the other on the plasmid, upstream of the EC042_pAA024 pseudogene. The ChIP-seq experiment identifies a CRP motif at both the chromosomal and plasmid *virK* locations, with similar sequences for the CRP motif: 5`-**TGTGG**TGACTG**GTACA**-3`. However, at each location, the DNA site for CRP is located in the coding region, indicating either a NAP role or long-distance regulation (or both). Another gene in the EAEC 042 plasmid that is associated with CRP binding is *aafA*, which encodes a fimbrial subunit. However, the CRP ChIP-seq signal profile in the *aafA* gene, obtained from the Artemis browser, showed a noisy background so this needs further study (Figure 4.7C).

Table 4.1: CRP targets identified by ChIP-seq analysis on the EAEC 042 plasmid.

Peak Centre[1]	Score[2]	Annotated genes[3]	*p*-value[4]	Matching sequence[5]	Distance to TSS[6]	*E. coli* K-12 Homologues[7]	Gene products[8]
112767	321	*repA*	7.83E-12	**CTTGA**GAATTG**TGACG**	-38	**In K-12**	Replication protein A
108769	293	*repA*	7.83E-12	**CTTGA**GAATTG**TGACG**	-35	**In K-12**	Replication protein A
62001	907	*EC042_RS26580*	3.46E-07	**AGACA**GTATTC**TGAAC**	-212	***traS***	Hypothetical protein
27988	2975	*pet*	5.92E-07	**CGAGA**GCATTG**TCACA**	-85	S.G.	Serine protease autotransporter toxin Pet
18257	387	*virK*	4.48E-06	**TGTGG**TGACTG**GTACA**	774	S.G.	Virulence protein
39765	81	*aafA*	7.72E-06	**GTGAA**TAATTG**TGATA**	113	S.G.	Aggregative adherence fimbria II major subunit aafA
86933	13	*EC042_RS26755*	1.11E-05	**CGTCA**GTATGG**CCTCT**	-109	***hok***	Type I toxin-antitoxin system Hok family toxin
50078	15	*EC042_RS29140*	1.36E-05	**ACGAT**GAATTC**TCAGA**	-143	pseudo	IS66 family insertion sequence element accessory protein TnpB
94581	23	*EC042_RS26810*	8.13E-05	**AGACA**TCCCCG**CGCCG**	-20	**In K-12**	Hypothetical protein
3987	13	*EC042_RS30500*	1.49E-04	**GACTA**AACTTC**CAGAG**	155	pseudo	Transposase

[1] Peak centre indicated by MACS2 and annotated to EAEC 042 plasmid pAA.

[2] Score represents the number of reads mapped to each position within the EAEC 042 genome.

[3] Annotated gene to an identified CRP site.

[4] *p*-value of CRP motif matches from MEME SUITE (Bailey *et al.*, 2009).

[5] Matching sequence obtained from MEME SUITE (Bailey *et al.*, 2009).

[6] Distance between an identified CRP site and the Translation Start Site (TSS) of annotated gene.

[7] *E. coli* K-12 homologues to EAEC 042 genes. S.G denotes an EAEC 042 specific genes. Genes in bold represent those listed for *E. coli* K-12 in RegulonDB (Santos-Zavaleta *et al.*, 2019). In K-12 indicates genes that are present in *E. coli* K-12 but with unknown functions.

[8] Products of encoding gene (Chaudhuri *et al.*, 2010).

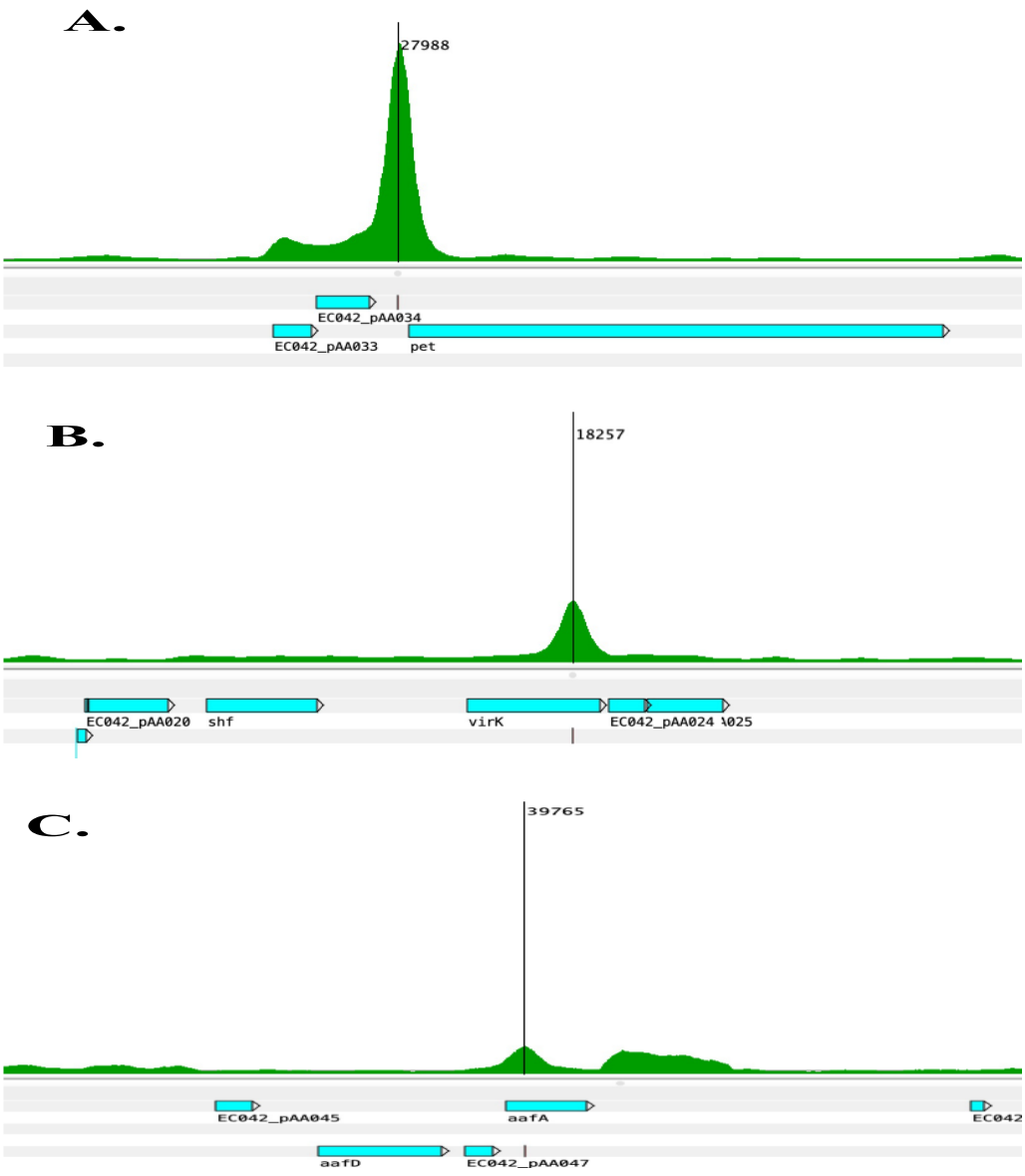


Figure 4.7: ChIP profiles CRP at selected EAEC 042 plasmid genes

The figure shows the CRP peaks at selected EAEC 042-plasmid genes: **A.** *pet,* **B.** *virK,* **C.** *aafA*. Genes are shown as light blue blocks and CRP binding is shown as green peaks. Green blocks denote the exact position of the genes within the annotation (obtained from the Artemis genome browser (Carver *et al*., 2011)).

Figure 4.8: CRP binding motif on EAEC 042 plasmid pAA

The logo represents the CRP consensus motif, deduced using MEME software (Bailey *et al.*, 2009) from DNA sequences at identified targets from CRP ChIP-seq analysis of the EAEC 042 pAA plasmid.

4.4 Discussion

The impact of CRP on bacterial biofilm formation is shown in Figure 4.2. EAEC strain 042DFB *Δcrp* formed a biofilm only when complemented with CRP (in pDCRP). This result, which evidences the involvement of the global transcription factor CRP in biofilm formation by the pathogenic EAEC strain 042, was unexpected. Building on this, I was able to show that defective CRP, CRP $AR1^{-}$ $AR2^{-}$, can be used as a tool to suppress biofilm formation in the wild type strain. This is because activating region 1 (AR1) and activing region 2 (AR2) are needed for activation by CRP (Busby, 2019).

Most of the data in this chapter focuses on a study of the role of CRP in EAEC 042 that aimed to understand whether novel CRP-regulated targets in the EAEC 042 genome had evolved. CRP ChIP-seq analysis of the EAEC 042 chromosome revealed 322 CRP sites related to 312 genes. Some of these CRP targets are common to *E. coli* K-12 and were reported by Grainger *et al.* (2005) (underlined written genes in Appendix B), whereas other targets are to our knowledge, new (genes labelled with S.G. in Appendix B) and are the subject for more investigation. Most CRP sites from the data were in non-coding regions and so the new targets may well be involved in CRP-dependent regulation. However, some CRP sites were in coding regions, suggesting either long-distance regulation or a NAP role for CRP.

Three genes involved in adhesion and autoaggregation like OmpC, OmpF, and OmpT, *lpxC* (encoding a protein associated with lipid A biosynthesis), and *slp* (encoding an outer-membrane lipoprotein induced after carbon starvation) were among the genes that showed increased expression in biofilms. *slp* and *ompC* are two of these genes that have recently been linked to the early stages of *E. coli* biofilm development on abiotic surfaces (Sauer, 2003). However, none of these genes were reported by CRP ChIP-seq, indicating indirect regulation by CRP of these genes, as the result in Figure 4.2 shows a positive impact of CRP in EAEC biofilm formation.

The targets from my CRP ChIP-seq experiments were compared with the targets from the previous study of Grainger *et al.* (2005) and the Regulon DB database for *E. coli* K-12 strain MG1655. A total of 291 targets for CRP binding were common between *E. coli* K-12 strain MG1655 and EAEC 042, whereas 31 genes from the CRP ChIP-seq analysis were exclusive to

EAEC 042, indicating the recruitment of new target genes into the CRP regulon. These new targets might contribute to EAEC 042 virulence, and this requires more investigation. Most of the CRP sites from CRP ChIP-seq analysis were recorded upstream of genes that encoded enzymes for cell metabolism indicating the important of these enzymes to nourish the cell and consume other carbon sources in environments with a low level of glucose or to compete with commensal microbes in the intestines (Shimada *et al.*, 2011). CRP sites in EAEC 042 plasmid pAA shows unconserved site as they encode the pathogenicity determinants, but we were able to record CRP binding peak in the regulatory region of *pet* which is supported by the work of Rossiter *et al.* (2011). However, no novel targets for CRP in the pAA plasmid were identified from the CRP ChIP-seq data.

Chapter 5

Investigation of some EAEC 042-specific genes targeted by CRP

5.1 Introduction

The EAEC 042-specific genes identified from CRP ChIP-seq data were investigated to check CRP binding and study the possible regulatory roles of CRP at these targets. EAEC 042-specific genes from Appendix B, can be classified into four groups depending on the gene product:

i. protein with unknown function, this group contain 14 genes.
ii. prophage protein, this group contains 4 genes.
iii. type VI secretion system protein, this group contain 3 genes.
iv. 10 genes encode 'various' products, some of which may well be involved in pathogenesis (Table 5.1).

5.2 Experiment design

In this Chapter, I selected 10 targets from Table 5.1 (underlined red-highlighted) for *in vitro* investigation using Electrophoretic Mobility Shift Assays (EMSA). The aim here is to study the CRP binding site *in vitro* and compare it with ChIP-seq score (as *in vivo* result). The 10 targets that I selected were: *kpsMII, EC042_0414, EC042_3143, EC042_0225, EC042_0224, EC042_4604, EC042_3975, virK, EC042_4012, EC042_3970*. The CRP binding site score for these targets show a range of values.

In Table 5.1 most of the CRP binding site located within genes, but 12 falls in intergenic regions. To test the role of CRP in regulation, I cloned putative promoter fragments into a *lacZ* vector and performed ß-galactosidase assays, only picking targets, if the CRP binding site was in an intergenic region with a good score of reads. It is important to clarify that the binding of CRP from ChIP-seq data in *EC042_3975* was too close to the TSS (-20), and any binding of TF near to TSS will inhibit the transcription, for this reason, *EC042_3975* was not included *in vivo* investigations. The selected CRP targets were *kpsMII*, *EC042_0225* and *EC042_0224* (divergent genes), and *EC042_0536* (bold typed genes in Table 5.1). The positive control, pRW50/*CC 41.5,* (provided from Dr. Douglas Browning) was used to calibrate the experiment.

Table 5.1: The EAEC 042-specific genes from CRP ChIP-seq data

Gene name	Gene Product	Score in ChIP-seq	Distance to Translation start	Matching Sequence
kpsMII	Polysialic acid transport permease protein	894	-479	**TGTGA**TTTATA**TCACA**
EC042_0414	Conserved hypothetical protein	312	65	**TGTGA**TCTTGC**GCACA**
EC042_3143	Hypothetical protein	6783	31	**CGTGT**TGTAGA**TCACA**
EC042_0536	Putative adhesin (not virulence)	168	-112	**TGTGA**TGTAAA**GCGCA**
EC042_0225	Putative type VI secretion system protein	111	-67	**TGTGA**GCCGCA**TCACA**
EC042_0224	Putative type VI secretion system protein	97	108	**ATTGC**CGGTGA**TCAAT**
EC042_4604	Putative helicase (pseudogene)	14	315	**TTTCG**CTGCCT**TCACA**
EC042_3975	Hypothetical protein	250	-20	**TGTTA**TCGTAT**ACACG**
virK	Virulence protein	827	774	**TGTGG**TGACTG**GTACA**
EC042_4012	Putative invasion '*air*'	4	4601	**CGTGA**CGTTTG**TCGCG**
EC042_4187	Hypothetical protein	35	1	**CGAGG**TGTTGA**TCACG**
EC042_3970	Putative prophage protein	4	301	**TGTGG**ATGAGA**GCAAA**
EC042_4606	Hypothetical protein	33	787	**TCTGT**GCTCCA**TCAAG**
EC042_2431	Putative prophage protein	81	-223	**TGTTT**TGTAGA**TGTAA**
aadA1	Aminoglycoside adenylyltransferase	46	61	**CGTTG**CTGGCC**GTACA**
EC042_3998	Putative phage immunity repressor protein	23	-45	**TGTGG**CGTTGA**TTGCG**
EC042_4746	Conserved hypothetical protein	13	-696	**TGCGG**TTTTGC**GGCAA**
EC042_0305	Conserved hypothetical protein	16	876	**TCCTA**TTATCG**GCACC**
EC042_0769	Hypothetical protein	39	-381	**TCTGA**GTTTGA**TGTTA**
EC042_4744	Conserved hypothetical protein	1	-1219	**GGTGA**ACTCCA**TCTTG**
EC042_3956	Putative prophage protein	33	435	**TATGA**CGCGCC**GAGCG**
EC042_3465	Hypothetical protein	20	-108	**CCCGT**CCTCGG**TACCA**
EC042_3075	Conserved hypothetical protein	24	-248	**CCTTC**GCCCGC**TCCAA**
tetR	Tetracycline repressor	45	108	**TCTGA**CGACAC**GCAAA**
EC042_4011	Transcriptional regulator	5	2663	**GGTTG**GAGCGA**TGCCA**
EC042_0363	Hypothetical protein	9	153	**TATGA**AAAAAA**TCATA**
EC042_4524	Putative type VI secretion protein	11	-137	**ATCGA**TATAAT**GCAAA**
EC042_2670	Conserved hypothetical protein	209	179	**GGAGG**TAGTCA**TCGCA**
EC042_3190	Conserved hypothetical protein	69	33	**ACTCA**TTCAGT**TCATG**

set1A	Enterotoxin 1	174	39	**ACTGA**CGGTTT**TCCCA**
EC042_2130	Putative prophage protein	95	769	**GGGTT**AGATCG**TCACT**

**Targets indicated by red-typed genes have been tested *in vitro*, whereas bold-typed genes have been tested *in vivo*.

5.3 CRP binding *in vitro* using Electrophoretic Mobility Shift Assays

To quantify the interaction of CRP with target sites at EAEC 042-specific genes *in vitro,* I used Electrophoretic Mobility Shift Assays (EMSAs). DNA fragments covering selected EAEC 042-specific targets were first amplified by PCR using chromosomal EAEC 042 DNA as template and primers listed in Table 2.4. Each PCR product was approximately 200 bp and flanked with EcoRI and HindIII restriction sites (*kpsMII, EC042_0414, EC042_3143, EC042_0225, EC042_0224, EC042_4604, EC042_3975, virK, EC042_4012, EC042_3970*). Each PCR product was than ligated into plasmid pSR, using the EcoRI and HindIII restriction sites, and the base sequence was checked. DNA fragments for EMSA experiments were generated using EcoRI and HindIII restriction enzymes and end-labelled with γ-^{32} PATP (10 µCi/µl). Labelled DNA fragments were incubated with different concentrations of purified CRP protein (200nM, 400nM, 800nM and 1.6 µM) at 37 $^{\circ}$C for 30 minutes, and a negative control was included in each experiment (no CRP). Then the mixture (labelled DNA fragments and CRP or the negative control) was run on a 6% (*w/v*) polyacrylamide gel supplemented with 250 nM cAMP. After electrophoresis, the gel was exposed to a phosphor-screen.

The results of EMSA experiments are shown in figure 5.1. A clear detectable shifted band was seen with *kpsMII, EC042_0225, EC042_4012, EC042_0414* and *EC042_3143,* indicating that CRP binds strongly at its target in these fragments. In contrast, no shifted bands were detected with *EC042_0224*. This could be explained by the poor fit of the suggested DNA site for CRP to the consensus, matching it at only 5/10 positions. For both the *virK* and *EC042_3975* fragments, CRP causes a very small shift at high concentration, consistent with a very weak binding target (at approximately 1.6 µM). For the *EC042_3970* fragment, CRP causes a shift only at 400nM, and for the *EC042_4604* fragment, the shift is seen at 800nM. These results show that fragments with the weaker sites required higher concentrations of CRP for binding to be evidenced. As expected, the CRP sites with a 10/10 match to the consensus, at the *kpsMII* and *EC042_ 0225* targets, showed a clear strong band shift.

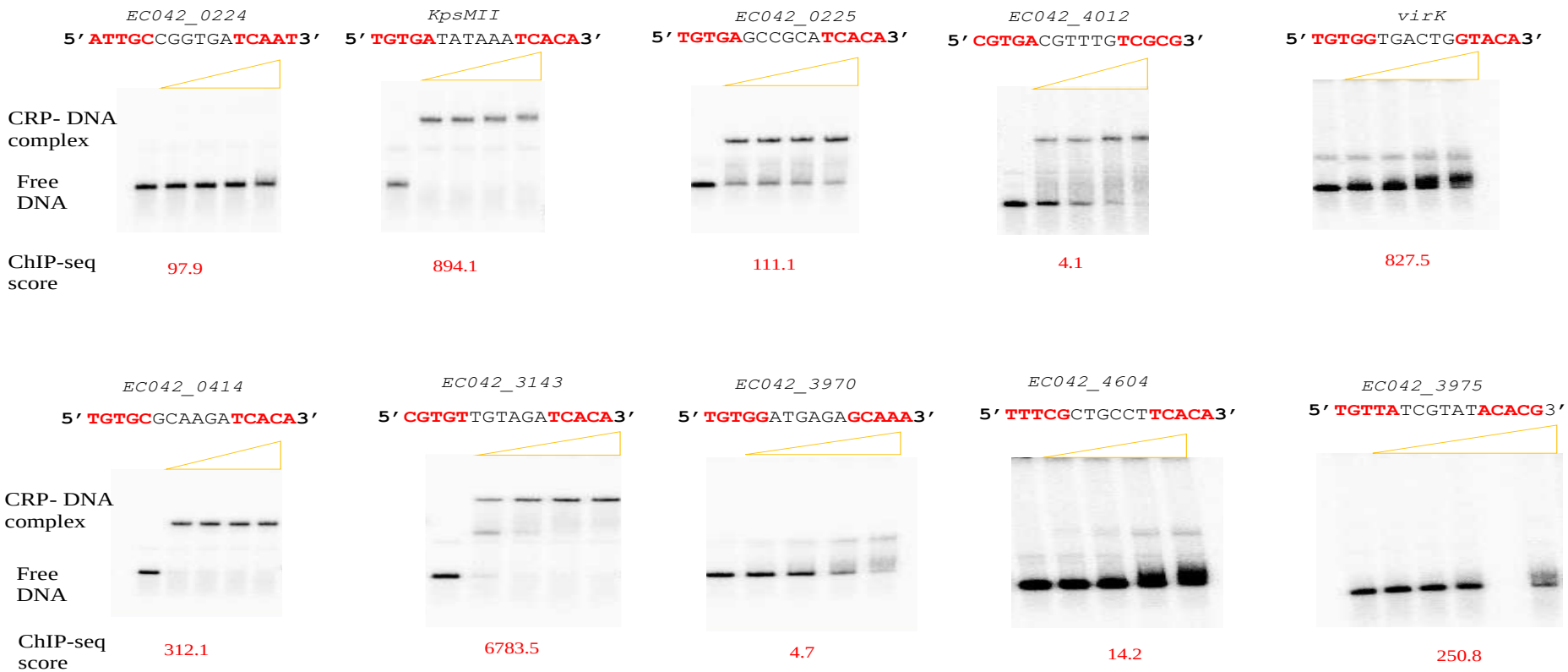


Figure 5.1: CRP binding *in vitro* at EAEC 042-specific genes

Radiolabelled EAEC 042-specific DNA fragments were incubated with purified CRP protein at 37 ^{o}C for 30 minutes. Mixtures were electrophoresed on 6% (w/v) polyacrylamide gel, in buffer supplemented with cAMP. Gels were fixed, then dried and exposed to a phosphor-screen. The CRP concentration in each panel started with 0 (no CRP), followed by 200nM, 400nM, 800nM and 1.6 µM. The CRP binding sequence for each target, obtained from MEME, is shown for each target. Bands resulting from binding of CRP are labelled as 'CRP-DNA complex'. The ChIP-seq score is noted in red typeface under each target.

5.4 *In vivo* investigation of CRP regulation at selected targets

Some CRP sites, found to be located at intergenic regions were investigated for the presence of CRP-regulated promoters. These sites are upstream of *kpsMII, EC042_0224*, *EC042_0225* and *EC042_0536,* suggesting a possible regulatory role for CRP. The location of the CRP site, as shown by ChIP-seq data, for *kpsMII* is -479 bp, for *EC042_0225*, it is -67, and for *EC042_0536*, it is -112. In the case of the CRP site at *EC042_0225*, a divergent gene, *EC042_0224*, was identified and is listed as specific gene to EAEC strain 042. Hence, I cloned the different promoter regions, into *lac* reporter plasmid, and tested for promoter activity by measuring ß-galactosidase activities.

5.4.1 The role of CRP at the *kpsMII* regulatory region

Amplified DNA fragments carrying of the *pkpsMII* regulatory region sequence were ligated into promoter-less *lacZ* reporter plasmid (pRW50) with help of EcoRI and HindIII sites. Recombinant plasmids pRW50/*pkpsMII* was then transformed separately into *E. coli* K-12 strains M182 Δ*lac* and *E. coli* K-12 strain M182 Δ*lac* Δ*crp*. Transformed cells were then grown in LB medium at 37˚C with shaking to mid-exponential phase, cells were lysed when growth reached OD_{650}= 0.4-0.6 and measured ß-galactosidase levels were taken as a proxy for promoter activity. The positive control, pRW50/CC 41.5, and negative control, pRW50, with no insert, were included in these experiments as controls.

Results are shown in Figure 5.3C. The DNA insert in pRW50/*pkpsMII* clearly carries one or more promoters and expression is inhibited 3-fold by CRP. Hence, CRP appears to act as a repressor here. Therefore, more investigation of the *pkpsMII* regulatory region is needed to understand the role of CRP (see Chapter 6).

>*pkpsMII* (527 bp)

```
                                                  500
GAATTCATAACACCATTTAAATGTGATATAAATCACAAATATGACTGTAAAGAGGGGGCTGTAG
                                450
ATATAAATAAGAAGTACACGAGTGAATTTTAAATAGGGAAATAGTTTCTCGGTGAACAATTTAT
     400                                                        350
TGGTAATCAATCGCGTGCGTTCTGGTTTGAAACAGATTGCAGGTATTGTTACGCATAAAATCCA
                                                       300
TGGGGGTATTATAATCAAGTAGTTAACAGTAATAACAAAAAATAATTTCCTGGGAAATTAACTC
                                    250
TACATTCTAAAACGGCCAGTTTTCTGAAATTACCAGAAGACCGAATACATTCATGCCTGAAGAA
                 200                                                150
TGGATTAGAAGAAACGAGACGGAAATAGATCTATTTATCCCTGCGGAAATAATTTCTGCTGAAA
                                                  100
TTTTTTGCGGCTATTAAAAAGGTCAAACCGTCTGAGTAAATTTTATCCAGTTACAAATAAGCAT
                                          50
TACCTCCAGTGTATTGGTAGCTGTTAAGCCAAGGGCGGTAGCGTACCTGAAGAGATTAGGATCA
                       1
CATCATCAAATGGCAAGAAGTAAGCTT
```

Figure 5.2: Base sequence of fragment carrying the EAEC 042 *kpsMII* regulatory region

The figure shows the base sequence of a DNA fragment carrying 527 bp upstream of the *kpsMII* gene. The CRP binding site is in bold bases and highlighted in light green. The red bases represent the start of the *kpsMII* gene coding region. The sequences highlighted with dark green denote EcoRI and HindIII restriction sites. The numbering of the fragment starts from the HindIII site.

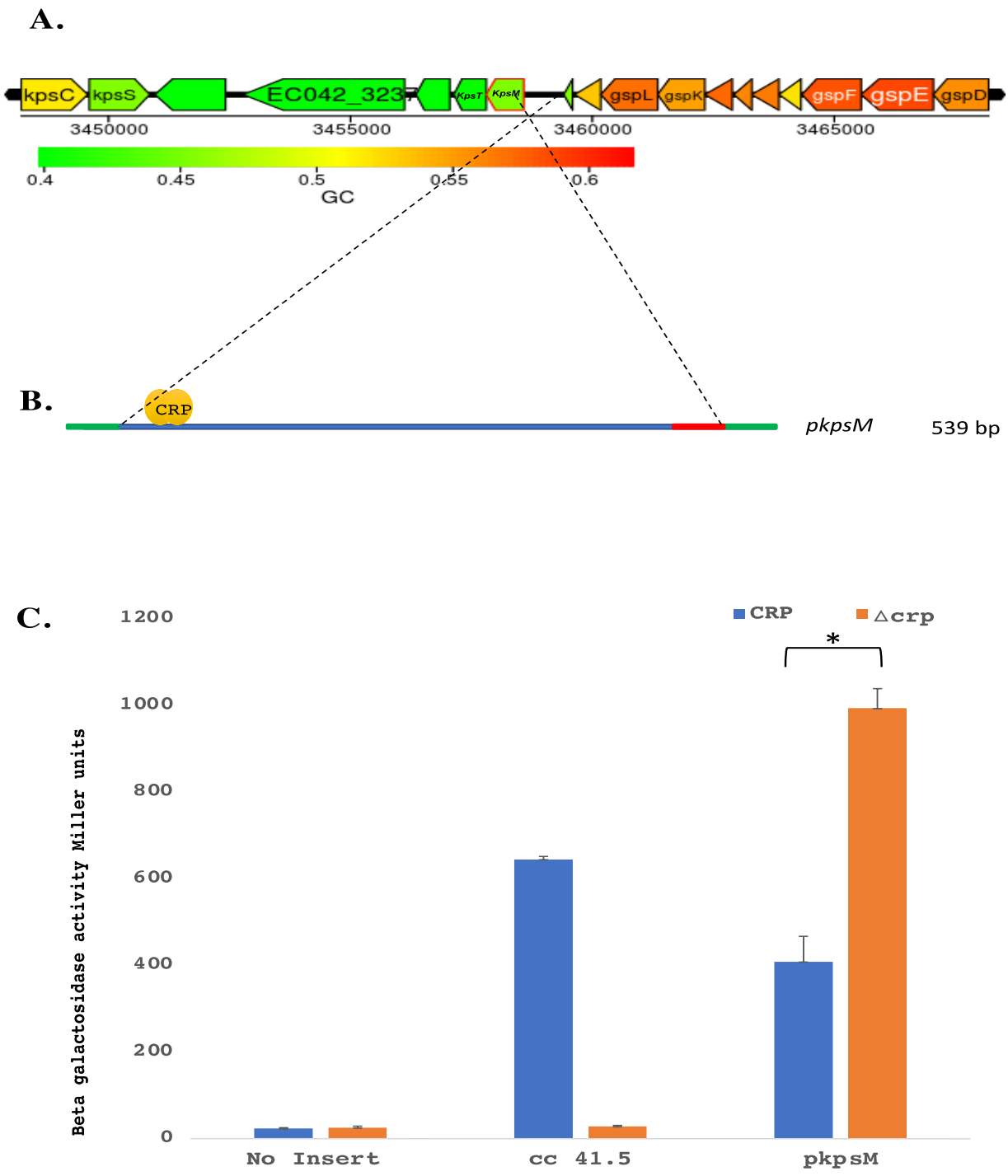


Figure 5.3: Assessing regulation by CRP at the *kpsMII* promoter region

A. The panel shows a diagrammatic representation of the position of the *kpsMII* gene in the EAEC 042 chromosome (Chaudhuri *et al.,* 2007).

B. Schematic of the *pkpsMII* promoter fragment. The blue line represents the regulatory region, and the red line represents the start codon of the gene. The CRP dimer is denoted in yellow, binding at its target. Restriction sites for EcoRI and HindIII are shown as green lines.

C. The Figure shows measured ß-galactosidase levels in *E. coli* K-12 strain M182 Δ*lac* (blue bars) and *E. coli* K-12 strain M182 Δ*lac* Δ*crp* (orange bars) carrying pRW50*/pkpsMII*, or pRW50/cc41.5, or pRW50 as a control. Each promoter activity indicates the average of three independent ß- galactosidase assays and the error bars denote the standard deviation (STDEV) and * indicates $P < 0.0001$ using a student's *t*-test.

5.4.2 The role of CRP at the *EC042_0225 / EC042_0224* intergenic region

CRP ChIP-seq data had shown an interesting CRP site between divergent genes that encoded two type VI secretion system genes (*EC042_0225*/ *EC042_0224*). Hence, this intergenic region was amplified by CRP using EAEC 042 isolated chromosomal DNA as template and primers listed in Table 2.4. The resulting fragments (shown in Figure 5.4) were ligated into promoter-less *lacZ* reporter plasmid, pRW50 (for *EC042_0225*) or pRW224 (for *EC042_0224*). The recombinant plasmids (pRW224/*p0224 VI* or pRW50/*p0225 VI*) were then transformed into *E. coli* K-12 strain M182 Δ*lac* and *E. coli* K-12 strain M182 Δ*lac* Δ*crp*, transformants were grown into LB medium at 37˚C with shaking, till the growth reach an OD_{650} between 0.4-0.6. Cultures were lysed and ß-galactosidase activities were taken as a proxy for promoter activity, with pRW50/cc41.5 and empty pRW50 being used as controls.

Results, shown in Figure 5.5C, indicate that the *p0225 VI* fragment failed to display any measurable promoter activity, not with standing the clear CRP binding, evidenced in Figure 5.1. Measured expression matches that found with the negative control, pRW50 with no insert. In contrast, the *p0224 VI* fragment showed activity, albeit low (less than 100 Miller units of ß-galactosidase was measured), that was impacted by CRP. At first sight the role of CRP here appears to be associated with chromosome organization.

>*p0224 VI* (163 bp)

```
                                150
GAATTCACGAACAACCCGTCGTTAATCCCCGACGGGTTAGAGTGTGATGCGGCTCACATCA
        100                                                  50
GGAACCACACCCTTTTCTTTATTCGAGTTACAGGCACACGTAATTAGCACCTACTGAAAAA
                                                          1
CTCTTTTATGGTGTTTAAATATATTTTTAGTCAATTTCAGAACACGGAAGCTT
```

>*p0225 VI* (169 bp)

```
                                150
GAATTCGAGTTTTTCAGTAGGTGCTAATTACGTGTGCCTGTAACTCGAATAAAGAAAAGGG
                100
TGTGGTTCCTGATGTGAGCCGCATCACACTCTAACCCGTCGGGGATTAACGACGGGTTGT
     50                                                    1
TCGTAAAACAGCAGTTGATAATTTCACAAGGAGTTCATAAATGCCAACCCCATGAAGCTT
```

Figure 5.4: Base sequence of fragments carrying the EAEC 042, EC042_0224 and EC042_0225 intergenic region

The figure shows the base sequence of a DNA fragments *p0224 VI*, carrying 163 bp upstream of the *EC042_0224* gene, *p0225 VI*, carrying 169 bp upstream of the *EC042_0225* gene. The consensus of CRP binding site is in bold highlighted in light green. The red bases denote part of the *EC042_0225* gene. Bases highlighted with dark green represent the EcoRI and HindIII restriction sites. The numbering of the fragment starts from HindIII site.

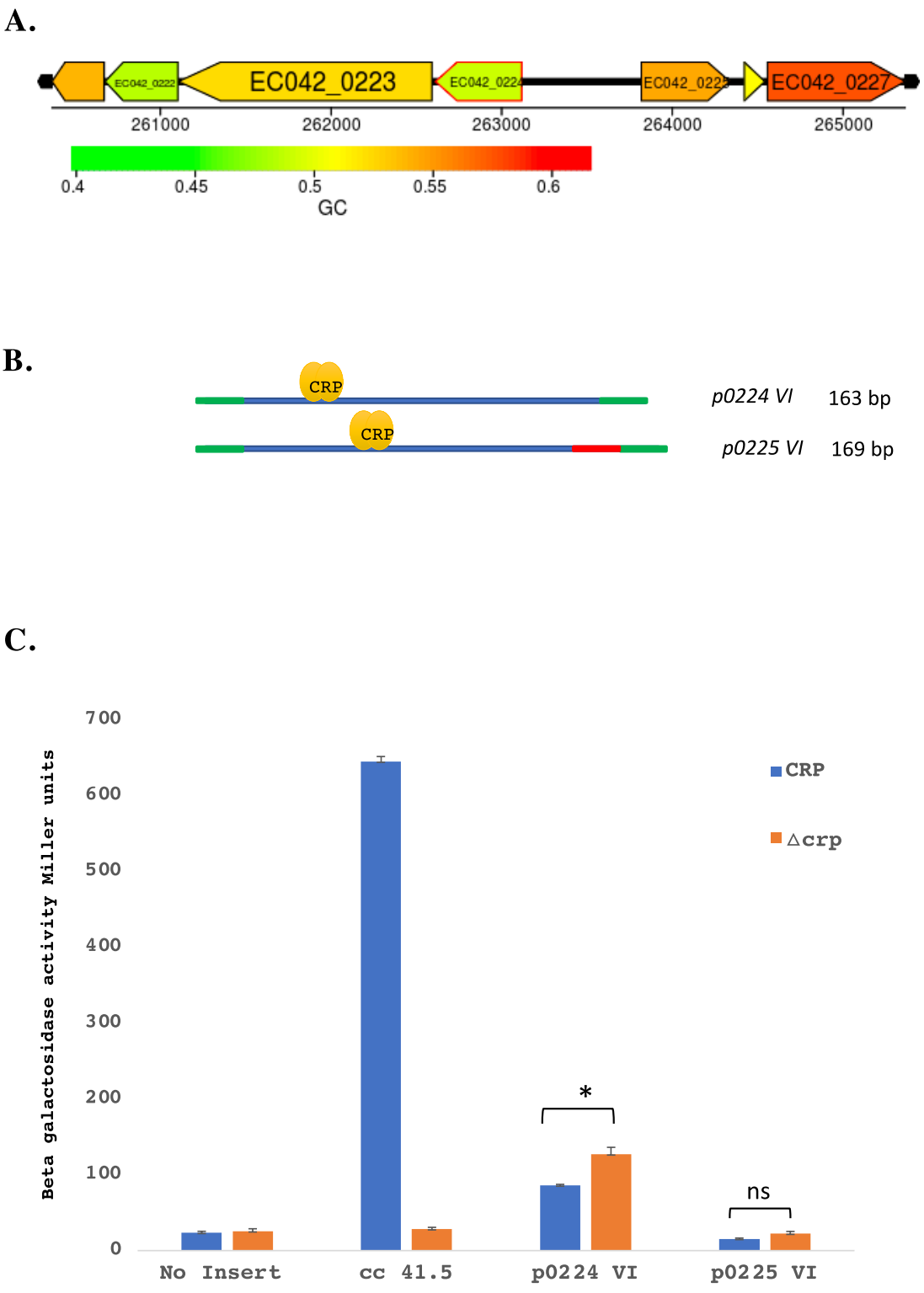


Figure 5.5: Promoter activity in the EC042_0224/EC042_0225 intergenic region

A. Diagrammatic representation of the *EC042_0224* (outlined with red) and *EC042_0225* (upstream of *EC042_0224)* genes in the EAEC 042 chromosome (Chaudhuri *et al.,* 2007).

B. The figure shows schematics of the *p0224* and *p0225* promoter fragments. The blue line represents the regulatory region, and the red line represents the start codon of the EC042_0225

gene. The CRP dimer is denoted in yellow, binding at its target. Restriction sites for EcoRI and HindIII are denoted as green lines.

C. The Figure shows measured ß-galactosidase levels in *E. coli* K-12 strain M182 Δ*lac* (blue bars) and *E. coli* K-12 strain M182 Δ*lac* Δ*crp* (orange bars) carrying the *lacZ* expression vector, pRW50 (negative control), or pRW50/cc41.5 (positive control) or pRW224/*p0224* or pRW50/*p0225*. Each promoter activity indicates the average of three independent ß-galactosidase assays and the error bars denote the standard deviation (STDEV). * Indicates $P < 0.001$ and **ns** not significant using a student's *t*-test.

5.4.3 The role of CRP at the EC042_0536 regulatory region

A target of interest from the CRP ChIP-seq data is *EC042_0536* gene, that encodes an EAEC-specific putative adhesin. This is an autotransporter adhesin, and the region from amino acids 615...717 suggests that it is like the *Yersinia* YadA anchor that was reported by Hoiczyk *et al.* (2000). The EcoRI-HindIII *p0536* fragment, shown in Figure 5.6, which carries sequence upstream from the EC042_0536 gene coding region, was amplified by PCR (primers are listed in Table 2.4). The fragment was ligated into pRW50 and resulting plasmid, pRW50/*p0536,* was then transformed into *E. coli* K-12 strain M182 Δ*lac* and *E. coli* K-12 strain M182 Δ*lac* Δ*crp*. Transformants were grown in LB medium at 37˚C, with shaking till the growth reached an OD_{650} between 0.4-0.6, cultures were then lysed, and the ß-galactosidase activities were taken as a proxy for promoter activity. Results are shown in Figure 5.7C. Cells that carry pRW50/*p0536* shows promoter activity that is higher than with pRW50/cc41.5. However, this activity is higher in cells where the *crp* gene has been knocked out, indicating that CRP acts as repressor here.

>*p0536* (168 bp)

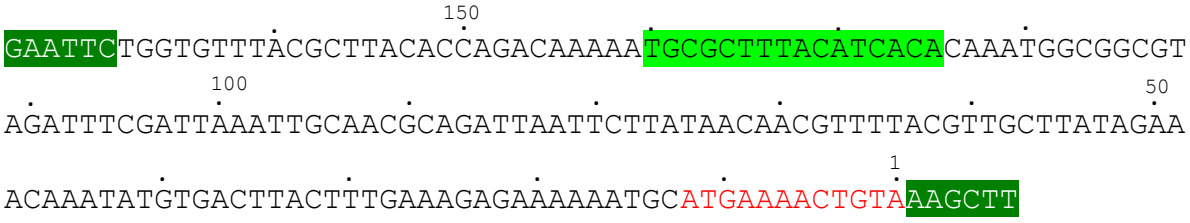


Figure 5.6: Bases sequence of the EAEC 042 *p0536* fragment

The figure shows the base sequence of *p0536*, 168 bp upstream of *EC042_0536* gene. The CRP binding site is denoted by bold bases and highlighted in light green. The red bases show the start of the *EC042_0536* gene coding sequence. The sequence highlighted with dark green denote the EcoRI and HindIII restriction sites. The numbering of the fragment starts from the HindIII site.

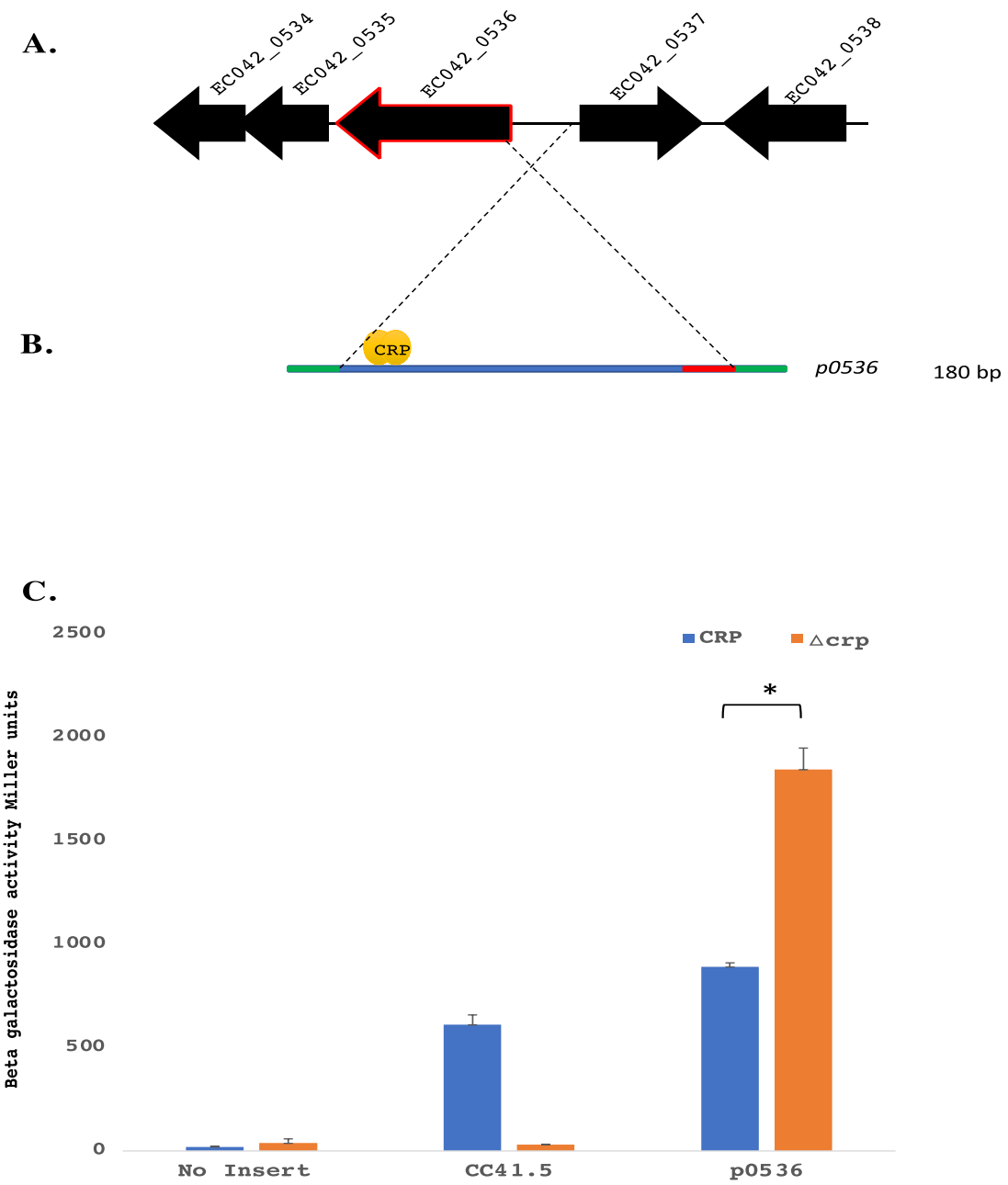


Figure 5.7: Assessing the regulation by CRP at the *EC042_0536* promoter region

A. The panel shows a diagrammatic representation of the *EC042_0536* gene in the EAEC 042 chromosome. *EC042_0536* shows as a black arrow with red outlined.

B. The figure illustrates the *p0536* promoter fragment. The blue line represents promoter region sequences. The red line represents the *EC042_0536* gene start codon. CRP dimer (yellow) are shown binding to a target site. The EcoRI and HindIII restriction sites are shown as green lines.

C. The Figure shows measured ß-galactosidase activities in *E. coli* K-12 strain M182 Δ*lac* (blue bars) and *E. coli* K-12 strain M182 Δ*lac* Δ*crp* (orange bars) carrying pRW50 (negative control) or pRW50/cc41.5 (positive control) or pRW50/*p0536*. Each promoter activity indicates the average of three independent ß- galactosidase assays and the error bars denote the standard deviation (STDEV) and * Indicates $P < 0.00001$ using a student's *t*-test.

5.5 Discussion

The interaction of CRP with sites near to EAEC 042-specific genes was checked here, both *in vitro* and *in vivo*. *In vitro* results in Figure 5.1 show strong binding of CRP when the CRP site in DNA has full consensus 10/10 (*EC042_ 0225, kpsMII*) or 9/10 (*EC042_ 0414)*. However, CRP consensus scores of less than 7/10 either required a higher concentrations of CRP protein like the result with *EC042_ 3970*, and *EC042_ 4604*, and *EC042_ 3975,* or *virK*. Or no binding shifts were detected such as the case of *EC042_ 0224*. Some of these results *in vitro* did not agree with CRP ChIP-seq results *in vivo,* such as the case of *virK* (Scored 827 reads) and *EC042_3975* (Scores 250 reads), where no detectable bands were recorded *in vitro*. Interestingly, *EC042_3143* scored high reads more than 6000 and shows a detectable binding from lower concentrations of CRP though CRP site within gene (consensens match 8/10) indicating a NAP role of CRP there or a long distance of promoter regulation. In the case of *EC042_ 4012* compared to *EC042_3970*, CRP showed detectable binding of CRP in *EC042_ 4012* at lower concentration in contrast to *EC042_3970*, though the results from ChIP-seq (Scored 4 reads) from these two targets were similar. These all indicate the need for a helper, and this helper could be another transcription factor.

The role of CRP at the intergenic regions of EAEC 042-specific genes has been studied *in vivo*. Results in Figure 5.3C show that CRP is involved in the regulation of the upstream region of *kpsMII* gene by repressing a promoter in its regulatory region. Therefore, the regulatory region of *kpsMII* was subject to further analysis (see Chapter 6). Another interesting reason for studying *kpsMII* promoter region is that this gene encodes virulence determinants (polysialic acid transport protein; Group 2 capsule) that might be involved in EAEC pathogenesis. Another regulatory region where CRP appears to repress is the *EC042_ 0536* promoter region that carries a strong promoter, and since *EC042_ 0536* encoded a putative adhesion that might be involved in EAEC 042 pathogenesis, further investigation is also needed in this case.

Finally, the divergent, *EC042_0225/EC042_0224,* genes were included here as the intergenic region carries a near consensus DNA site for CRP. However, this region showed only weak promoter activity, there was little regulation by CRP and no bidirectional promoter present. Conclusion in *EC042_0225/EC042_0224* case could be that CRP might play a NAP role there.

Chapter 6

CRP regulation at the EAEC strain 042 *kpsMII* promoter region

6.1 Introduction

Many bacteria express a polysaccharide capsule that is believed to facilitate interaction between the bacterial cell and its immediate environment. More than 70 different types of polysaccharide capsule have been studied in *E. coli* strains, and they are categorised into four groups depending on genetic correlations, serology, and biochemical information. The majority of isolated extraintestinal *E. coli* that cause disease express group 2 k antigens capsule (Whitfield and Roberts, 1999). Group 2 capsules are usually expressed only at 37°C, signalling the pathogens' presence inside the host rather than outside environments (Xue *et al.*, 2009). Pathogens benefit from their capsule as a protection from the host immune system, and also avoidance of desiccation during transmission between different environments (Evrard *et al.*, 2010; Wooster *et al.*, 2006; Ophir and Gutnick, 1994). Most group 2 capsule genetic loci contain three regions, with region one (contain *kpsFEDUCS* genes) and region three (contain *kpsMT* genes) being conserved among strains that encoded this type of group. Products of regions one and three play roles in maturation and exportation of the polysaccharide. In contrast, region two is serotype specific and encodes proteins that are involved in synthesis of the polymer (Whitfield and Roberts, 1999).

The regulation of region 3 expression was studied by Xue *et al.* (2009). They found that transcription regulation in this region involves H-NS, SlyA and a large 5'untranslated region (741 bp of 5`UTR) that is located between the transcription starting site +1 and the first gene of the operon, *kpsMII*. H-NS plays a crucial role on repressing the transcription of this region at temperature 20°C. SlyA promotes transcription independently of H-NS at 37°C, but H-NS is required for maximum induction. The temperature-dependent activation of region 3 transcription is thought to be influenced by the relative availability of H-NS and SlyA in the cell. Xue *et al.* concluded that the long 5'UTR moderates the number of transcripts that reach the *kpsMII* structural gene. This process is helped by the RfaH, an anti-termination factor, which after binding at its target site, helps RNAP to elongate and translate transcripts (Xue *et al.*, 2009; Stevens *et al.*, 1994).

The CRP ChIP-seq data, in Chapter 4, showed a strong peak located in the intergenic region upstream of the *kpsMII* gene. Section 4.1 of Chapter 5 describes my investigation using a DNA fragment (Figure 5.2) that lacks the Xue *et al.* (2009) promoter. My results (Figure 5.3C)

showed the existence of another promoter that is repressed by CRP. Here I attempt to investigate the presence of multiple promoter in the *kpsMII* regulatory region.

6.2 Investigation of *kpsMII* upstream DNA sequence

In EAEC strain 042, the intergenic region located upstream of the *kpsMII* gene region contains 827 bp with 67% AT bases (Figure 6.1A). The -10 element of the promoter identified by Xue *et al (*2009) is located at position -764 bp upstream of the *kpsMII* gene whereas the DNA site for CRP, deduced from ChIP-seq, is located at -512 bp (Figure 6.2). Here, I studied the possibility of multiple promoter elements in the *kpsMII* intergenic region, using nested deletions and point mutations. I studied the influence of CRP binding at the promoters using multi-round *in vitro* transcription assays. I also investigated the effects of temperature and different conditions. Finally, *kpsMII* regulatory region sequences from EAEC strain 042 were aligned with UPEC strain CFT073 and strain UTI89 to compare the binding site of CRP and to study the similarities and differences.

6.2.1 *kpsMII* cloning strategy

Differents *kpsMII* fragments used in this Chapter are illustrated in Figure 6.1C. The longest, an 839 bp *kpsMII* promoter region EcoRI-HindIII DNA fragment (Figure 6.2), including the start codon of the *kpsMII* gene, was constructed using PCR and primers listed in Table 2.4. The PCR product was ligated into promoter-less *lacZ* vector, pRW50, resulting in pRW50/*pkpsMII* WT. To investigate the promoter associated with the DNA site for CRP, a shorter 216 bp internal *kpsMII* promoter fragment was constructed using pRW50/*kpsMII* WT as template. Forward and reverse versions were produced (Figure 6.3B). These were cloned into pRW224 to study the orientation of any promoters adjacent to the CRP binding site. Starting with pRW50/*kpsMII* WT, nested deletions were used to define different promoters. Also, I constructed a shorter 440 bp fragment *kpsMII*, including the two possible promoters, for analysis with point mutations. This fragment, *kpsMII* 440, was also cloned upstream of a terminator in the small plasmid, pSR, for *in vitro* multi-round transcription assays.

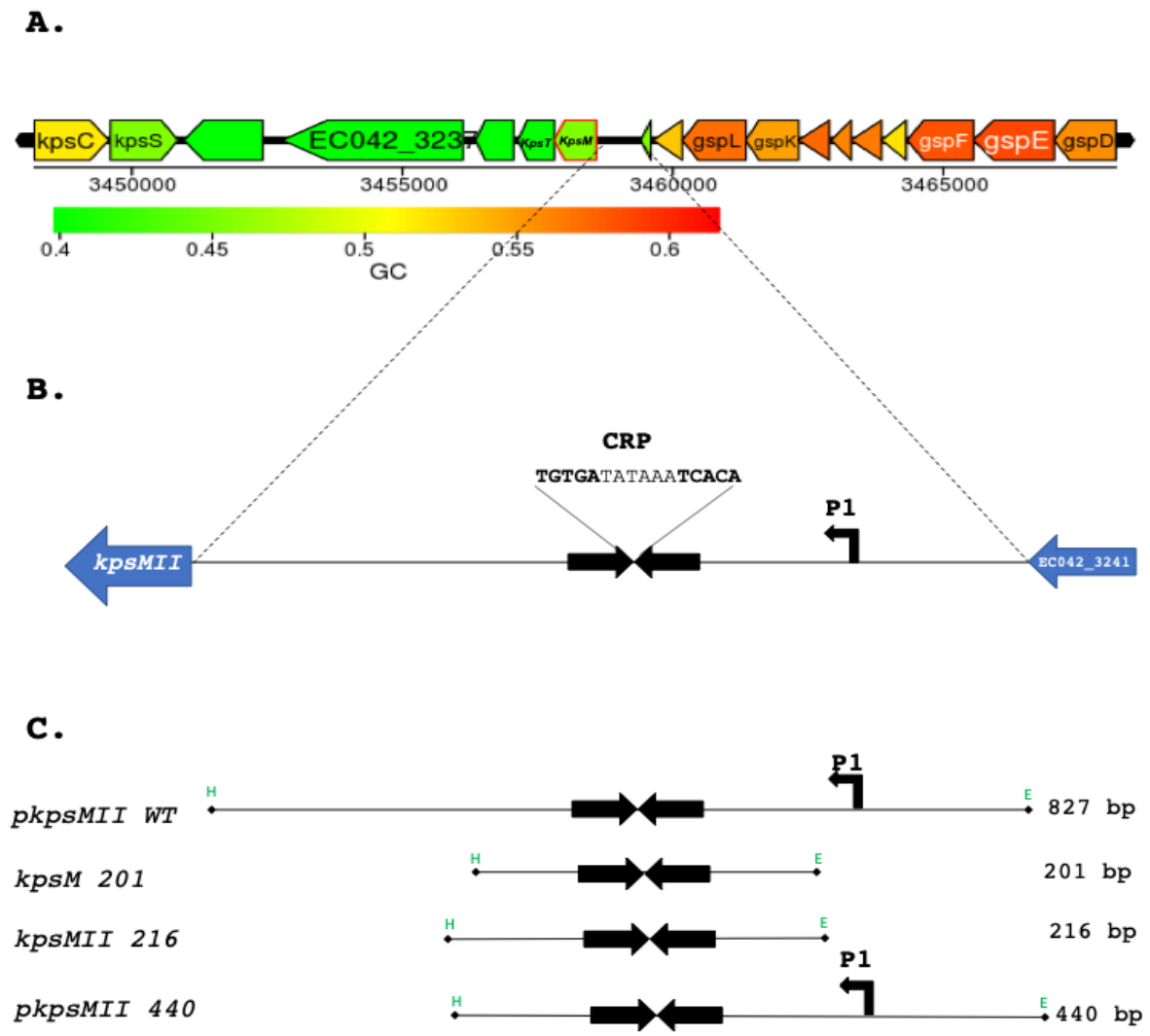


Figure 6.1: Chromosome context of *kpsMII* in EAEC 042

A. The panel shows a diagrammatic representation of the position of the *kpsMII* gene in EAEC 042. The *kpsMII* gene is outlined in red (Chaudhuri *et al.,* 2007).

B. The panel shows a schematic diagram of the *kpsMII* intergenic region. Blue arrows represent the gene orientations. The CRP binding site is shown as black arrows with the CRP consensus. The previously identified promoter by (Xue *et al.*, 2009) is shown as **P1** located upstream of the CRP binding site.

C. The panel shows a schematic diagram of the different *kpsMII* fragments used in experiments here: *pkpsMII* WT, *kpsM* 201, *kpsMII* 216 and *pkpsMII* 440. Size of fragments shown right side of panel. The CRP binding site is shown as black arrows. Restriction sites flanking the fragment are shown in green EcoRI (E) and HindIII (H). The previously identified promoter by (Xue *et al.*, 2009) is shown as P1.

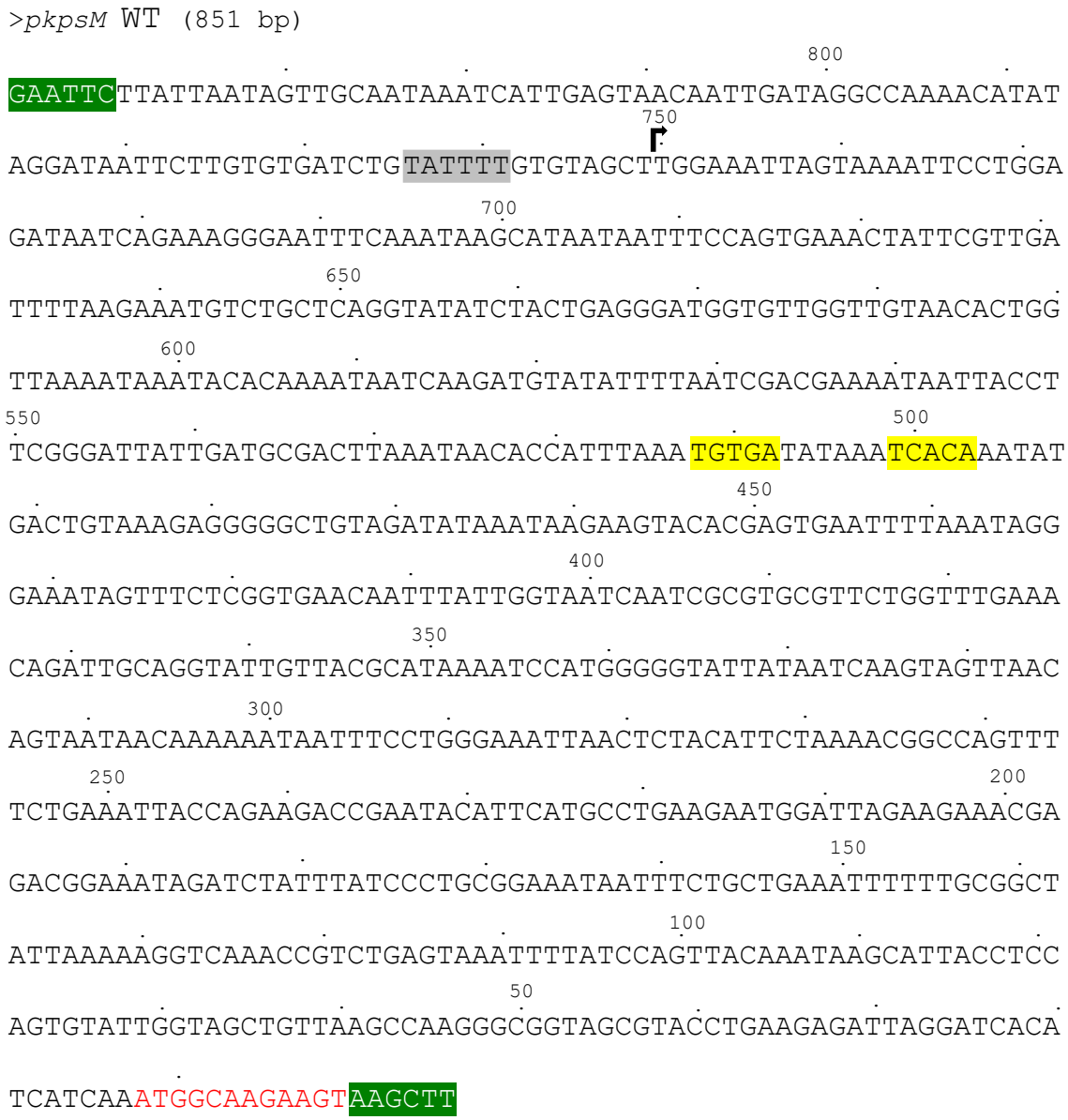


Figure 6.2: DNA sequence of the upstream of *kpsMII* promoter region from EAEC 042

The figure shows the DNA sequence of the 827 bp *kpsMII* WT fragment. The CRP binding site is highlighted in yellow. The DNA bases highlighted in grey represent the -10 element of the previously identified promoter (P1). The P1 transcript starting site +1 is illustrated with a black arrow at position 750 bp. *kpsMII* coding sequence is shown with red bases. The sequences highlighted with dark green denote the EcoRI and HindIII restriction sites. The numbering of the fragment starts from the HindIII site.

6.2.2 Orientation of the *kpsMII* second promoter

Results presented in Chapter 5 show promoter activity associated with the DNA sequence adjacent to the CRP binding site and hence the possibility of a second promoter distinct from that described by Xue *et al.* (2009). Hence a 216 bp segment of *kpsMII* intergenic region was cloned on an EcoRI and HindIII fragment in both orientations (Figure 6.3B). The resulting fragments, *kpsMII* 216 and *kpsMII* 216R, were cloned into *lacZ* promoter-less vector, pRW224, in order to measure the promoter activity. Recombinant plasmids were transformed then into *E. coli* K-12 strain M182 Δ*lac* and *E. coli* K-12 strain M182 Δ*lac* Δ*crp*. The transformed strains were grown in LB medium, incubated at 37˚C, with shaking to mid-logarithmic phase (OD_{650}=0.4-0.6). Cultures were lysed and ß-galactosidase activities were measured. The negative control, pRW224 with no insert, was used for promoter activity comparisons.

Data presented in Figure 6.3C show that ß-galactosidase expression was highest with *kpsMII* 216 in the Δ*crp* strain, indicating the presence of a CRP-repressed promoter. Lower, yet significant expression was observed with *kpsMII* 216R, suggesting a promoter that is divergent from the *kpsMII* 216 transcript and, hence, the *kpsMII* gene.

A.

>*pkpsM216* (216 bp)

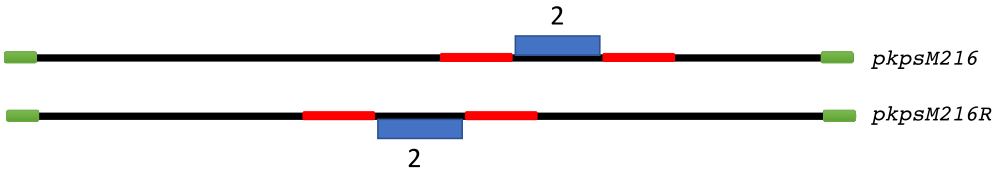
GAATTCTAACACTGGTTAAAATAAATACACAAAATAATCAAGATGTATATTTTAATCGACGAAA

ATAATTACCTTCGGGATTATTGATGCGACTTAAATAACACCATTTAAATGTGATATAAATCACA

AATATGACTGTAAAGAGGGGGCTGTAGATATAAATAAGAAGTACACGAGTGAATTTTAAATAGG

GAAATAGTTTCTCGGTGAACAAAGCTT

B.



C.

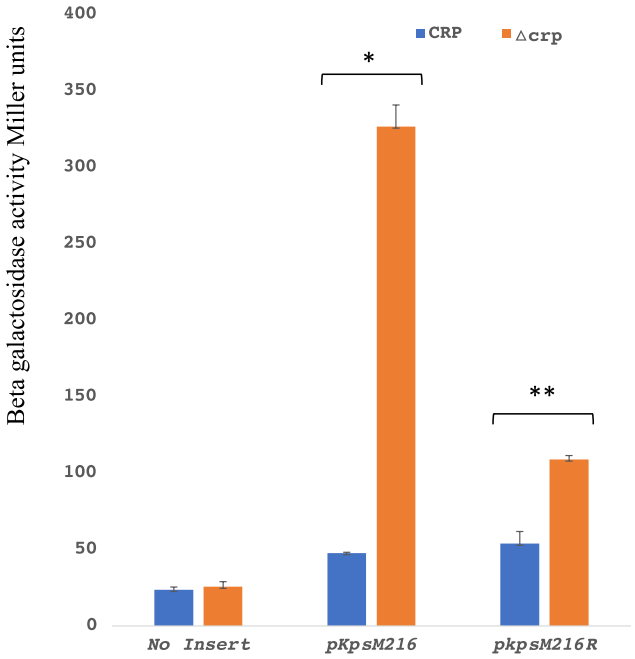


Figure 6.3: Promoter orientation of the *kpsMII* second promoter

A. The panel shows the base sequence of the *pkpsMII* 216 promoter fragment. The CRP binding site is shown in red and underlined bases. A possible second promoter -10 element is shown with blue bases and underlined within the CRP binding site. The sequences highlighted dark green denote flanking restriction sites (EcoRI and HindIII).

B. The panel shows schematic representations of the *pkpsMII 216* and *pkpsMII 216R* fragments. Red lines denote the CRP binding site and blue boxes represents the predicted -10 elements. Green lines represent flanking restriction sites.

C. The bar chart illustrates ß- galactosidase activites in *E. coli* K-12 strain M182 WT *Δ lac* (blue bars) and *E. coli* K-12 strain M182 *Δ lac Δcrp* (orange bars) carrying *pkpsMII 216* or *pkpsMII 216R* cloned in pRW224. No insert represents the negative control in the experiments (pRW224). Each activity is the average of three independent ß- galactosidase assays and the error bars denote the standard deviation (STDEV). * Indicates $P < 0.0000004$, and ** indicates $P < 0.0003$ using a student's *t*-test.

6.2.3 Location of *kpsMII* promoter determinants using nested deletions

To investigate the presence of multiple promoters in the *kpsMII* regulatory region, a series of nested deletions was constructed by deleting 100 bp each time. Starting with the pRW50/*kpsMII* WT (851 bp) fragment, I constructed five shorter fragments, using primers listed in Table 2.4: *kpsM* 751, *kpsM* 651 and *kpsM* 551, *kpsM* 451 and *kpsM* 351 fragments (Figure 6.4A). These were cloned separately into the *lacZ* vector, pRW50. The *kpsM* 851, *kpsM* 751, *kpsM* 651 and *kpsM* 551 fragments contain the DNA site for CRP binding, whereas it was lacking in the *kpsM* 451 bp and *kpsM* 351 bp fragments. The recombinant plasmids, including the pRW50/*kpsMII* WT (as positive control), were transformed into *E. coli* K-12 strain M182 Δ*lac* and *E. coli* K-12 strain M182 Δ*lac* Δ*crp.* Then, the transformed cells were grown in LB medium at 37˚C, with shaking to mid-exponential phase (OD_{650}=0.4-0.6). Cultures were then lysed for measurement of the ß-galactosidase levels. Here, the negative control, pRW50 with no insert, was used for comparison.

Figure 6.4B shows the result of the ß-galactosidase assays. Lysates of cells that carried plasmid with *kpsMII* WT 851 showed the highest activites compared with the other constructs indicating a powerful promoter located at the upstream end of *pkpsMII* WT 851. This is likely to be the promoter was previously reported by Xue *et al (*2009). The other plasmids carrying *kpsM* 751, *kpsM* 651 or *kpsM* 551) showed lower expression, that was optimal in cells with the *crp* deletion, consistent with the existence of a second promoter that is repressed by CRP. The other plasmids carrying *kpsM* 451 or *kpsM* 351, with no consensus DNA site for CRP showed the lowest ß-galactosidase levels. Analysing the DNA sequence around CRP site revealed a possible -10 element lying within the CRP consensus 5`-**TGTGA**TATAAA**TCACA**-3`. This could explain the repression of promoter activity by CRP by blocking RNAP engagement.

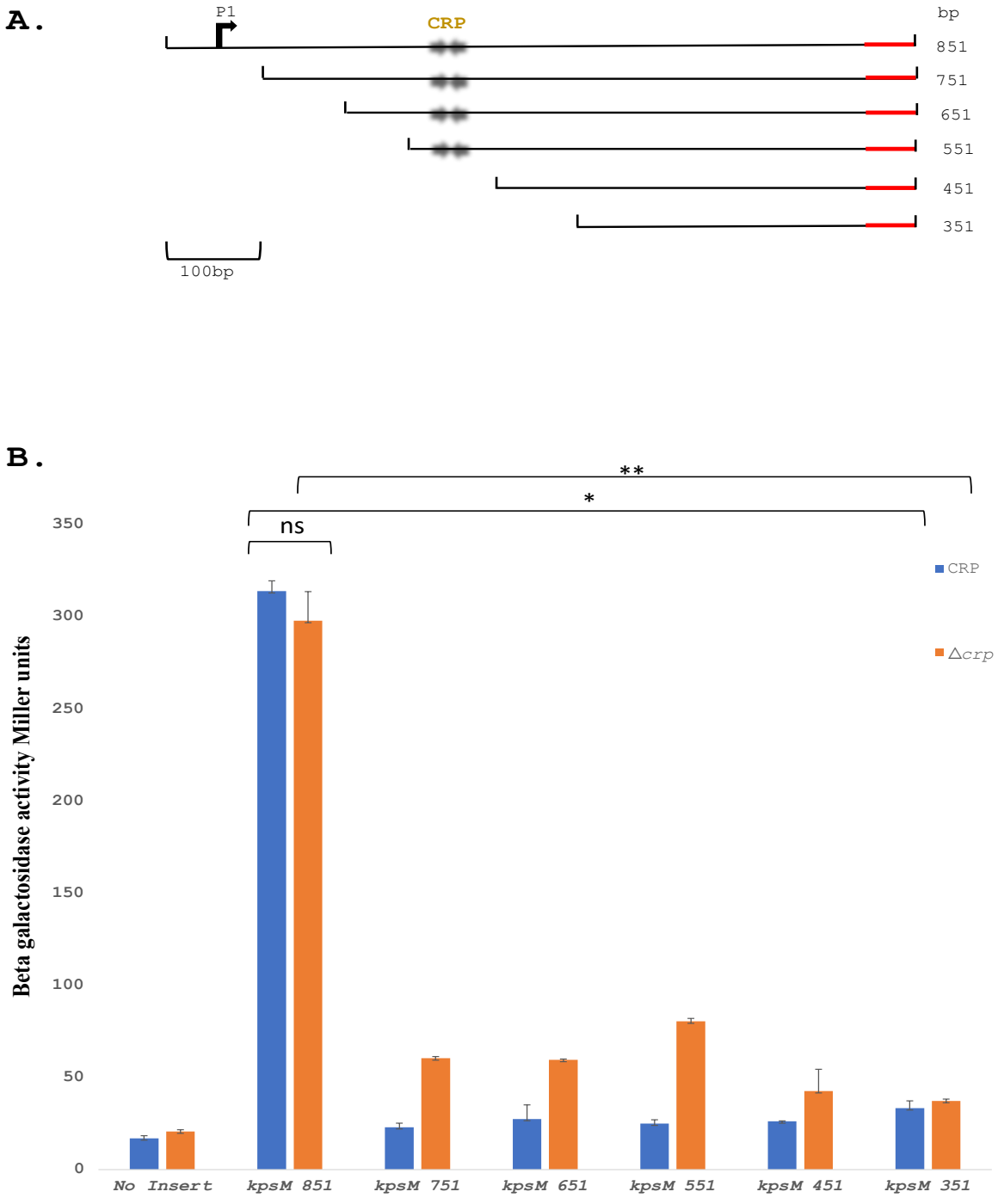


Figure 6.4: *kpsMII* regulatory region nested deletion analysis

A. The panel illustrates the promoter fragments of *kpsM* 851, *kpsM* 751, *kpsM* 651 and *kpsM* 551, *kpsM* 451 and *kpsM* 351. Gold arrows represent CRP binding sites. The black directional arrow indicates promoter **P1**. The red lines represent *kpsMII* gene coding sequence.

B. The bar chart illustrates measured ß- galactosidase activites in *E. coli* K-12 strain M182 WT *Δ lac* (blue bars) and *E. coli* K-12 strain M182 *Δ lac Δcrp* (orange bars) carrying different *kpsMII* intergenic fragments with nested deletions cloned in pRW50. No insert represents the negative control, empty pRW50. Each measureament is the average of three independent ß- galactosidase assays and the error bars represent the standard deviation (STDEV). * Indicates $P < 0.00000001$, ** Indicates $P < 0.000001$; and **ns** not significant using a student's *t*-test.

6.2.4 Identification of -10 element in the *kpsMII* promoter region by mutation analysis

To attempt to pinpoint the promoters in the *kpsMII* intergenic region, three derivatives of the *pkpsMII 440* fragment (Figure 6.5) were made by introducing point mutations in the -10 hexamer elements (shown in Figures 6.5 and 6.7A). A key point here is that I used a shortened version of the strating *kpsMII* with fragment with a size of 440 bp flanked with EcoRI and HindIII. The new fragments are *kpsM* p1KO (with knocking out the first promoter), *kpsM* p2KO (knocking out the second promoter), and *kpsM* p3KO (with knocking out both promoters). The different fragments were then cloned into *lacZ* reporter, pRW224 and transformed into *E. coli* K-12 strain M182 Δ*lac* and *E. coli* K-12 strain M182 Δ*lac* Δ*crp.* Transformed cells carrying the recombinant plasmids were grown in LB medium at 37˚C, with shaking to mid-logarithmic phase (OD_{650}=0.4-0.6). Cultures were lysed and ß-galactosidase levels were measured. A negative control, pRW224 with no insert, was included as a control.

Data in Figure 6.7B show that the starting fragment that carries two functional promoters gives high activity, but this is reduced nearly 3-fold in presence of *crp,* indicating a repression role for CRP here. Knocking out *p1* caused a major reduction of promoter activity, and again expression was reduced in the presence of *crp.* In the case of knocking out the second promoter, expression is lowered and, again, suppressed by *crp,* supporting the obstruction of RNAP from transcription by CRP. Optimal activity clearly needs both promoters here. Knocking out both -10 elements (pRW224/*kpsM* p3KO) led to a great reduction in promoter activity, as measured ß-galactosidase activity dropped to less than 100 Miller Units. These results support the existence of a second functional significant promoter in the *kpsMII* promoter region, and a role for CRP in regulation of *kpsMII* expression.

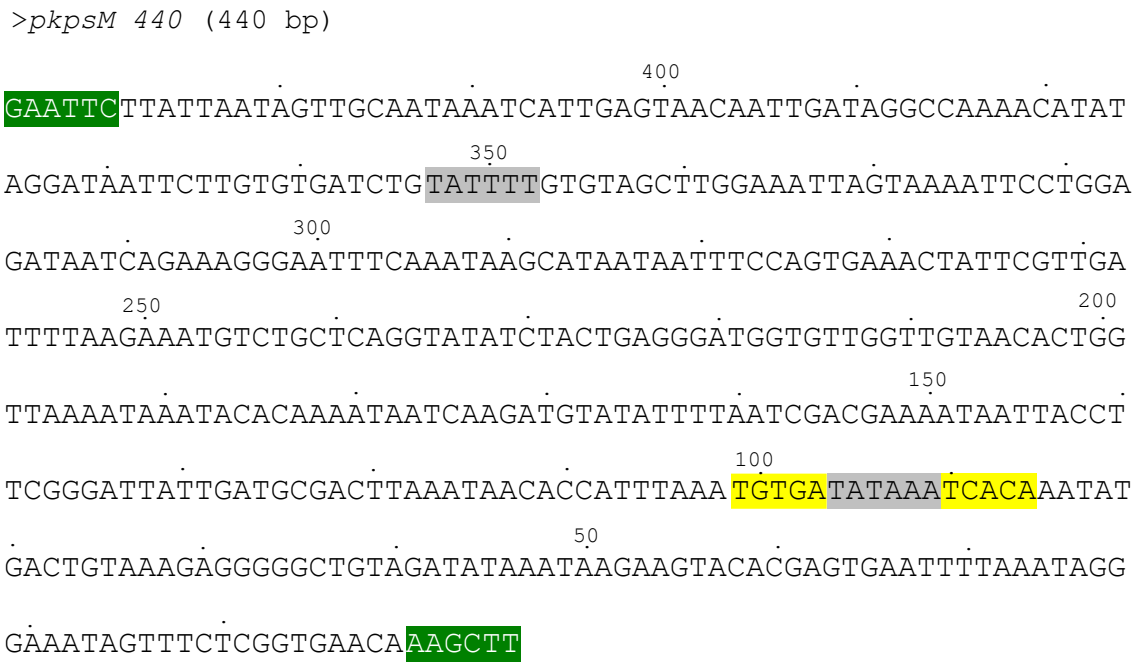


Figure 6.5: DNA sequence of the *kpsMII* 440 promoter fragment

The figure illustrates the DNA sequence of the *kpsMII* 440 promoter fragment, including two promoters P1 and P2. The CRP binding site is highlighted in yellow. The DNA bases highlighted in grey denote the suggested -10 element of the previously identified P1 promoter and the second promoter P2. The flanking EcoRI and HindIII restriction sites are shaded dark green. The numbering of the fragment starts from the HindIII site.

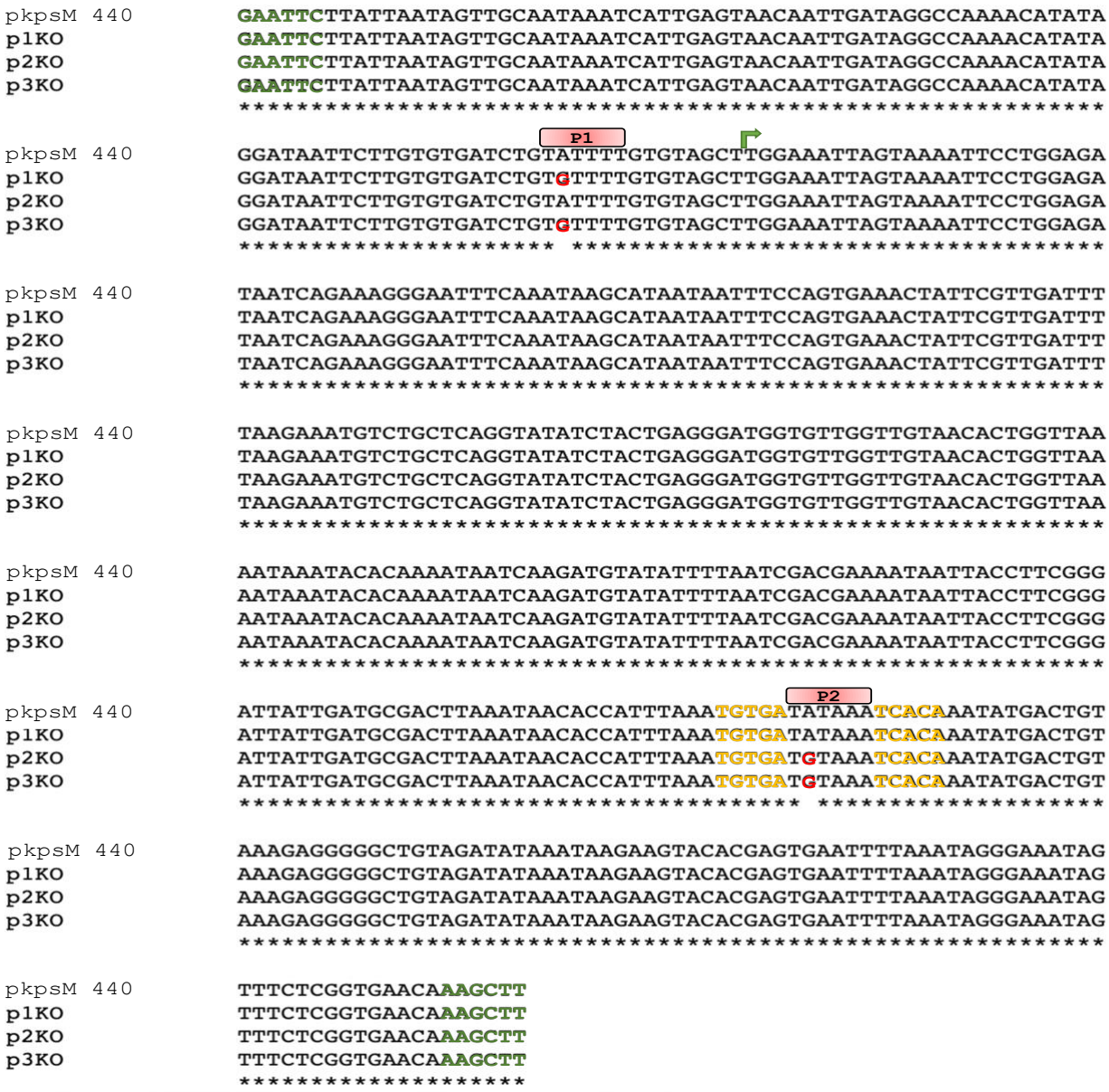


Figure 6.6: *kpsMII* derivatives with different -10 hexamer element mutations

The Figure presents base sequence alignments of *kpsMII* 440 fragments with two functional promoters (*pkpsM* 440), the upstream -10 element knock out (p1KO), the "both -10 elements" knock out (p3KO); and the downstream -10 element knock out (p2KO). The red-letter Gs mark the sites of mutation in the -10 elements and shaded red rectangles are located above each -10 element. The green directional arrow represents the transcript start site of *kpsM p1*. The gold DNA bases denote the CRP binding site. Green DNA bases indicate flanking restriction sites EcoRI (beginning of fragment) and HindIII (end of fragment). (Thompson *et al.*, 2003).

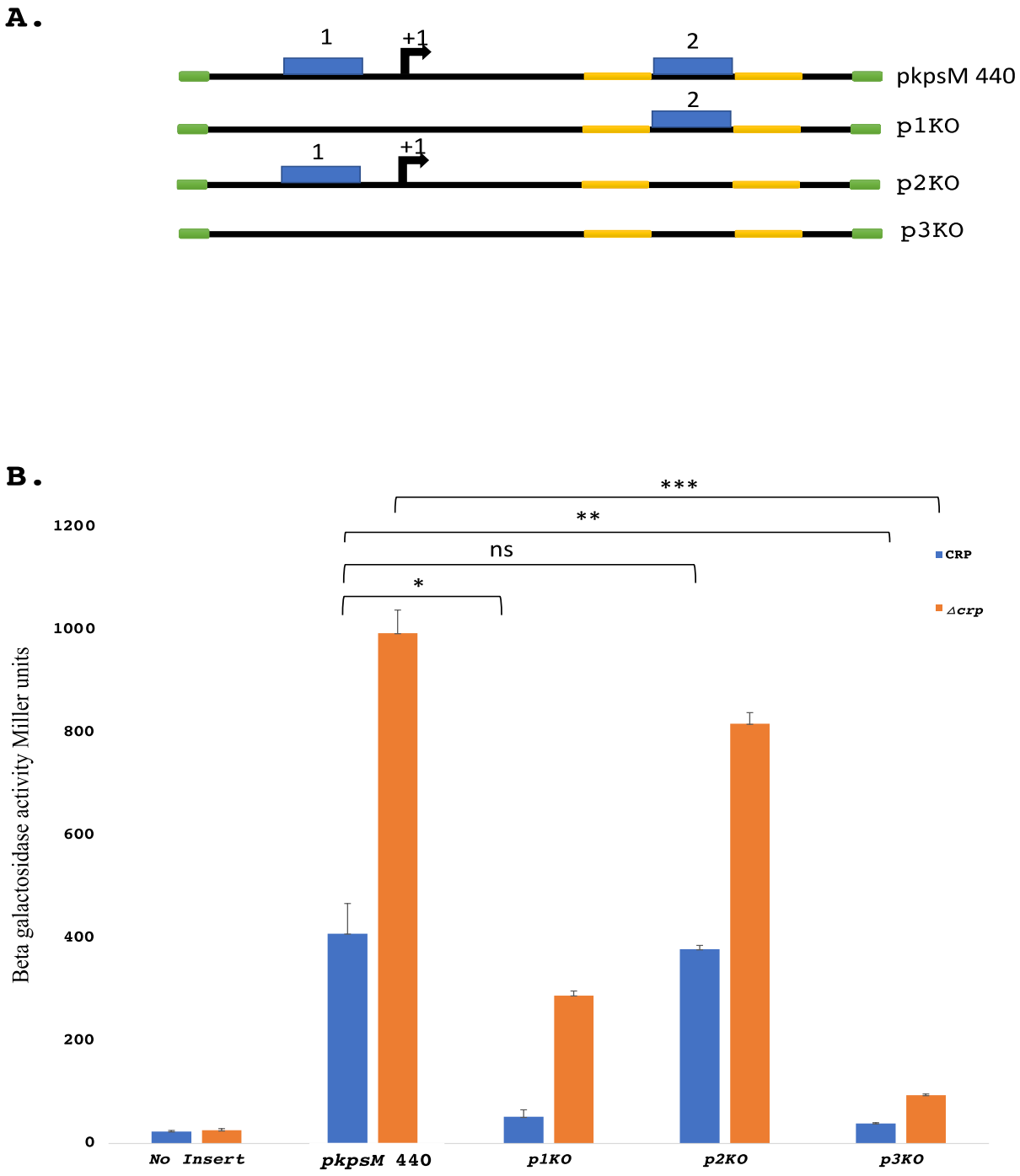


Figure 6.7: Mutational analysis of *kpsMII* promoters

A. The panel shows diagrams of the *kpsMII* promoter fragments pkpsM 440, p1KO, p2KO and p3KO. The blue bars represent the functional -10 element of *pkpsM* promoter. Gold lines represent the CRP binding sites and the black directional arrow represents +1 of *kpsMII* P1. Restriction sites (EcoRI and HindIII) are indicated by green lines.

B. The bar chart illustrates ß- galactosidase activites in *E. coli* K-12 strain M182 WT *Δ lac* (blue bars) and *E. coli* K-12 strain M182 *Δ lac Δcrp* (orange bars) carrying the different *kpsMII* 440 fragments cloned in pRW224. No insert represents the negative control with empty pRW224. Each measurement is the average of three independent ß-galactosidase assays and the error bars represent the standard deviation (STDEV). * Indicates $P < 0.0005$, ** Indicates $P < 0.003$, *** Indicates $P < 0.000004$; and **ns** not significant using a student's *t*-test.

6.2.5 Investigation of the *KpsM* promoter region DNA site for CRP

The *kpsMII* promoter region DNA site for CRP had been studied *in vitro* using Electrophoretic Mobility Shift Assays (EMSAs) (Figure 5.1). Here, a new 201 bp fragment was made with mutations that disrupt the CRP-binding consensus sequence (5`-**T<u>C</u>TGA**TATAAA**T<u>A</u>ACA**-3`) and compared with the same fragment with a functional DNA site for CRP (Figure 5.1, *kpsMII*). These fragments, *kpsM* 201 bp and *kpsM* 201CRPKO, again flanked by EcoRI and HindIII restriction sites, were ligated into the small plasmid pSR to check the DNA sequences. After that, the *kpsM* 201 and *kpsM* 201CRPKO fragments were end-labelled with labelled γ-^{32}P ATP (10 μCi/μl) and radiolabelled *kpsM* 201 and *kpsM* 201CRPKO fragments were incubated separately with different concentrations of purified CRP protein (200nM, 400nM, 800nM and 1.6 μM) at 37 ^{o}C for 30 minutes. The negative control in this experiment was "no CRP". Then the mixture (radiolabelled DNA fragments and CRP) was electrophoresed in a 6% (*w/v*) polyacrylamide gel supplemented with 250 nM cAMP. The gel was than dried and exposed to a phosphor-screen and visualized with a Bio-Rad Molecular Imager FX and analysed using Quantity One software (Bio-Rad). The results of this experiment are shown in Figure 6.8. As shown in Figure 5.1, CRP gives a clear band shift, even at the lowest concentration, and this is indicative of tight binding. This binding is stopped by the base changes in the *kpsM* 201CRPKO fragment. The result here underscores the crucial role of CRP consensus for CRP binding.

6.2.5 An *in vitro* study of the influence of CRP on transcription from the *kpsMII* second promoter

First, I used the EMSA assay to study the competition between CRP and RNAP for binding to the *pkpsM* 201 fragment. I followed the same procedure as in Section 6.2.5, but with adding *E. coli* holo-RNAP after fragments had been incubated with CRP. Figure 6.9A shows the binding of CRP and holo-RNAP to *kpsMII* 201 that contains only the P2 promoter. Clear bands appear as CRP binds to its site. However, after adding holo-RNAP to the mixture, most of the fragments remained occupied by CRP, leaving a small amount of fragment bound by holo-RNAP. This indicates that there is a competition between CRP and holo-RNAP binding the *kpsMII* P2 region. Another interesting point to show here is the faint band seen in Figure 6.9A with no holo-RNAP disappeared after the addition of holo-RNAP. This may indicate the

presence of a possible thired -10 element, likely to be at position 57 in the *kpsMII* 440 fragment (5`-TATAAA-3`: see Figure 6.5)

Next, I tried a multi-round *in vitro* transcription assay, using the *pkpsMII* 440 fragment cloned upstream of the *λoop* terminator in the small, supercoiled plasmid pSR. To preform the assay, pSR/*pkpsM* 440 was mixed with holo-RNAP and different concentrations of purified CRP protein (or no CRP as a negative control). After incubation with labelled NTPs, the reaction was stopped with stop solution and loaded directly into 8% (*w/v*) denaturing polyacrylamide gel. Then the gel was dried and exposed to a phosphor-screen and analysed. The transcript from *kpsMII* P1 appears at 250 nt as shown in Figure 6.9B. The transcription from *kpsMII* P2 was detected as an abundant 180 nt transcript in lane 1 (no CRP). However, the transcript from *kpsMII* P2 starts to decrease with increasing CRP concentration, indicating a transcription inhibition by blocking RNAP binding. Note that the transcription at *kpsMII* P1 was not affected by CRP binding. The P3 transcript, suggested from Figure 6.9A, also appeared in Figure 6.9B, as a transcript at 150 nt, supporting the presence of a possible third -10 element in the *kpsMII* intergenic region. Both, P2 and P3, showed an inhibition in transcription with the increase of CRP concentration, likely due to the near location of CRP site to these promoters.

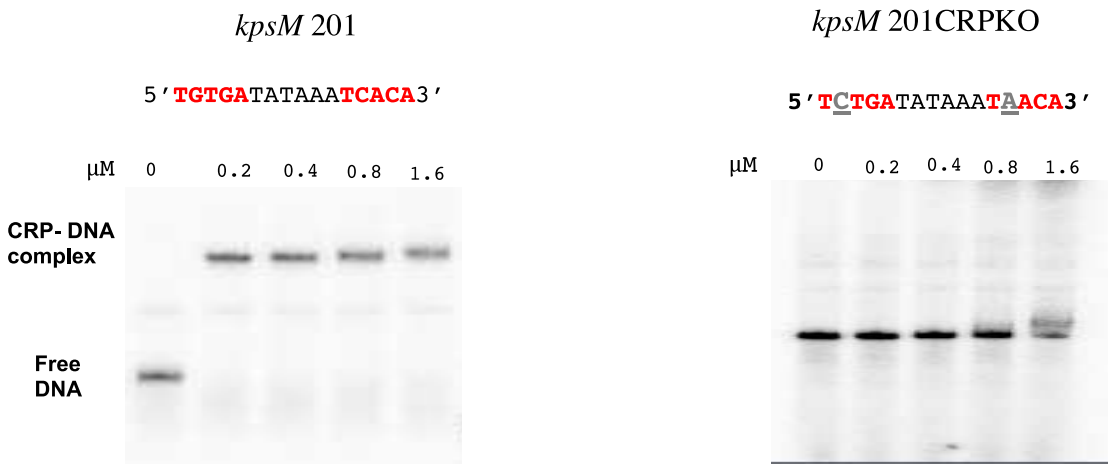


Figure 6.8: CRP binding *in vitro* studied by EMSA

Radiolabelled fragments (*kpsM* 201 and *kpsM* 201CRPKO) were incubated with purified CRP protein at 37 °C for 30 minutes. The mixture was then run on a 6% (*w/v*) polyacrylamide gel supplemented with cAMP, for 3 hours at 250 V. Gels were dried and exposed to a phosphor-screens. The CRP concentration in each panel starts with 0 (no CRP), 0.2 µM, 0.4 µM, 0.8 µM and 1.6 µM respectively. The base sequence of the CRP consensus is shown above the panels with red bold bases. For *kpsM* 201CRPKO, the grey bases denote the base changes that knockout CRP binding. The binding of CRP on DNA fragment labelled as 'CRP-DNA complex' and no binding was labelled with 'Free DNA'.

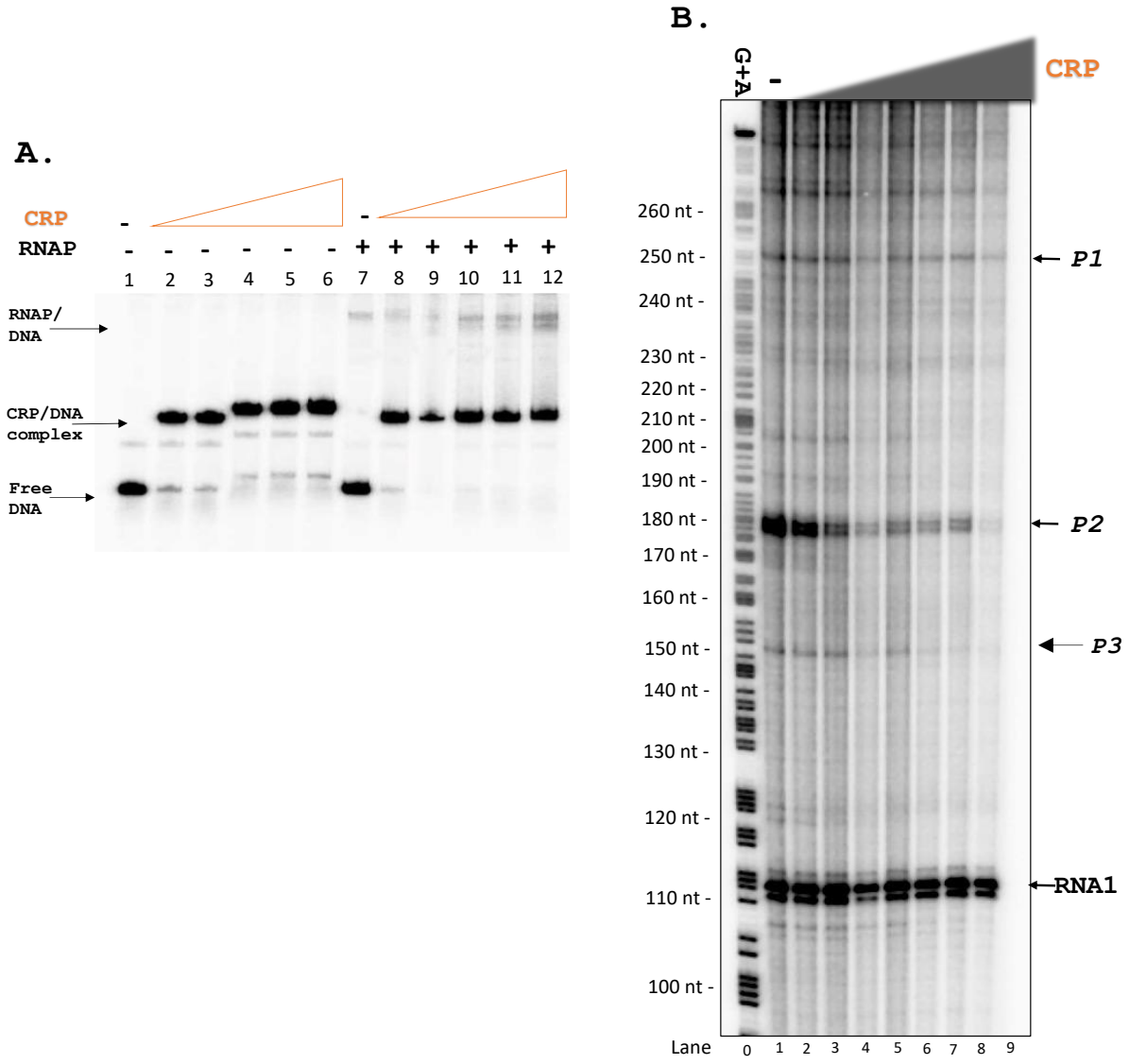


Figure 6.9: Promoter activity in the *kpsMII* intergenic region studied *in vitro*

A. The figure shows EMSA analysis of the effects of *E. coli* holo-RNAP binding to P2 at the *kpsMII* promoter region. The left part shows the radiolabelled *kpsMII* DNA fragments with no RNAP. CRP additions start from lane 2. The right part of figure shows the radiolabelled *kpsMII* DNA fragments mixed with *E. coli* holo-RNAP added into the mixture.

B. The figure shows transcripts resulting from *E. coli* holo-RNAP using the *kpsMII* promoter region, cloned in pSR plasmid. The RNA1 transcript, located at 110nt, represents the internal control for the experiment. The Maxam-Gillbert G+A ladder is shown in lane 0 and was used to calibrate transcripts starting at P1, P2 and P3 (as labelled). Lane 1 represents the control (no CRP). CRP concentrations start from lane 2 (0.1µM), lane 3 (0.2µM), lane 4 (0.4µM), lane 5 (0.6µM), lane 6 (0.8µM), lane 7 (1µM), lane 8 (2µM) and lane 9 (3µM).

6.2.7 Activity of the *kpsMII* second promoter at lower temperatures

The aim here was to assess expression from the second promoter (P2) in the *kpsMII* intergenic region at lower temperatures, since the first promoter was known to be repressed by H-NS at 20 ^{o}C (Xue *et al.*, 2009). pRW224/*kpsM* 440 that contains both functional promoters (P1 and P2) and the derivative pRW224/*kpsM* p2KO which has the second promoter knocked out (P2KO) were included in this experiment. Hence, these two recombinant plasmids were introduced into *E. coli* K-12 strain M182 Δ*lac* and *E. coli* K-12 strain M182 Δ*lac* Δ*crp*. The transformed cells were grown in LB medium and incubated at 37˚C or at 20˚C with shaking to mid-logarithmic phase (OD_{650}=0.4-0.6). Cultures were lysed to measure the ß-galactosidase activites. The negative control, pRW224 with no insert, was used here.

Figure 6.10 shows *kpsMII* promoter activities at two different temperatures (at 37 ^{o}C and 20 ^{o}C). Δ*crp* cells showed higher levels of ß-galactosidase expression, with more than 1000 Miller units at 37 ^{o}C, confirming that CRP blocks the second promoter from RNAP, and also interferes with RNAP that is transcribing from the first promoter. Cells grown at 20 ^{o}C showed much lower promoter activities: plasmids, pRW224/*kpsM* WT and pRW224/*kpsM* p2KO, both showed the same low level of low ß-galactosidase expression at 20 ^{o}C (only 100 Miller Units). These data suggest that the second promoter in the *kpsMII* intergenic region cannot be expressed at lower temperature and followed temperature regulation just like first promoter of *kpsMII* (P1) (Xue *et al.*, 2009).

6.2.8 *kpsMII* promoter activity in different nutrition conditions

The aim was to study the response of the *kpsMII* promoters to high glucose levels. Here, I used pRW224/*kpsM* 440 (containing both promoters) in *E. coli* K-12 strain M182 Δ*lac* and *E. coli* K-12 M182 Δ*lac* Δ*crp*. The level of ß-galactosidase expression was assessed during growth in LB or with LB supplemented with 0.4% (*w/v*) glucose. Cultures were incubated at 37˚C with shaking to mid-logarithmic phase (OD_{650}=0.4-0.6). pRW224 with no insert was used here as the negative control for promoter activity comparisons.

Figure 6.11 illustrates the activity of the *kpsMII* promoters in the different nutrition conditions. The results show the same pattern as previous results, with promoter activity boosted in the *crp* deleted strain. High levels of glucose are known to lower the cAMP signal in cells, but the

result here can be explained by the consensus nature of the DNA for CRP (Gaston *et al.*, 1988). Note however, that the repression factor seen due to *crp* is significantly less in the glucose-supplemented medium.

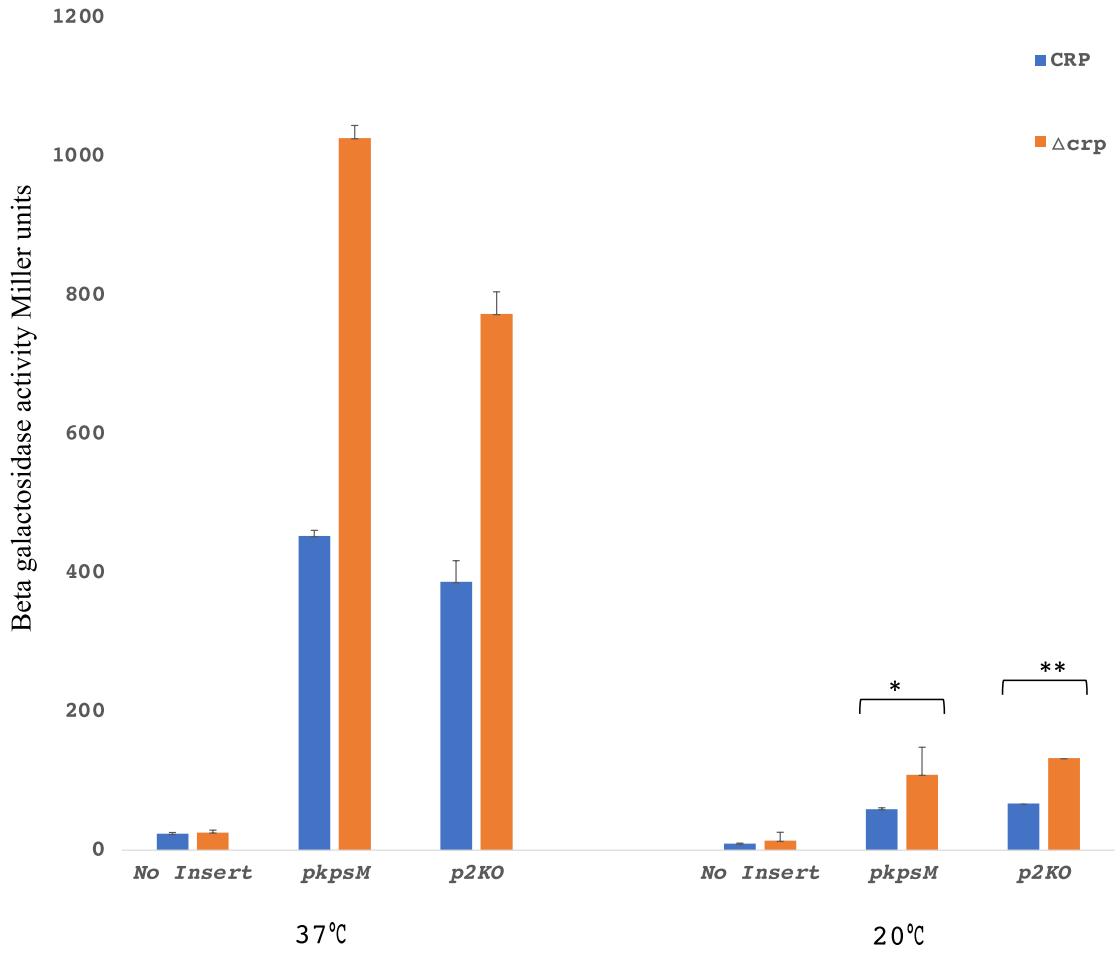


Figure 6.10: Temperature effects on expression from the *kpsMII* promoters and the effect of CRP

The bar chart illustrates measured expression from the *pkpsMII* 440 fragment and the p2KO derivative in *E. coli* K-12 strain M182 WT *Δ lac* (blue bars) and *E. coli* K-12 strain M182 *Δ lac Δcrp* (orange bars). The left-side panel concerns cells that were incubated at 37°C and the right-side panel concerns cells incubated at 20°C. 'No insert' denotes the negative control with empty pRW224. Each value indicates the average of three independent ß- galactosidase assays and the error bars represent the standard deviation (STDEV). * Indicates $P < 0.002$ and ** indicates $P < 0.04$ using a student's *t*-test.

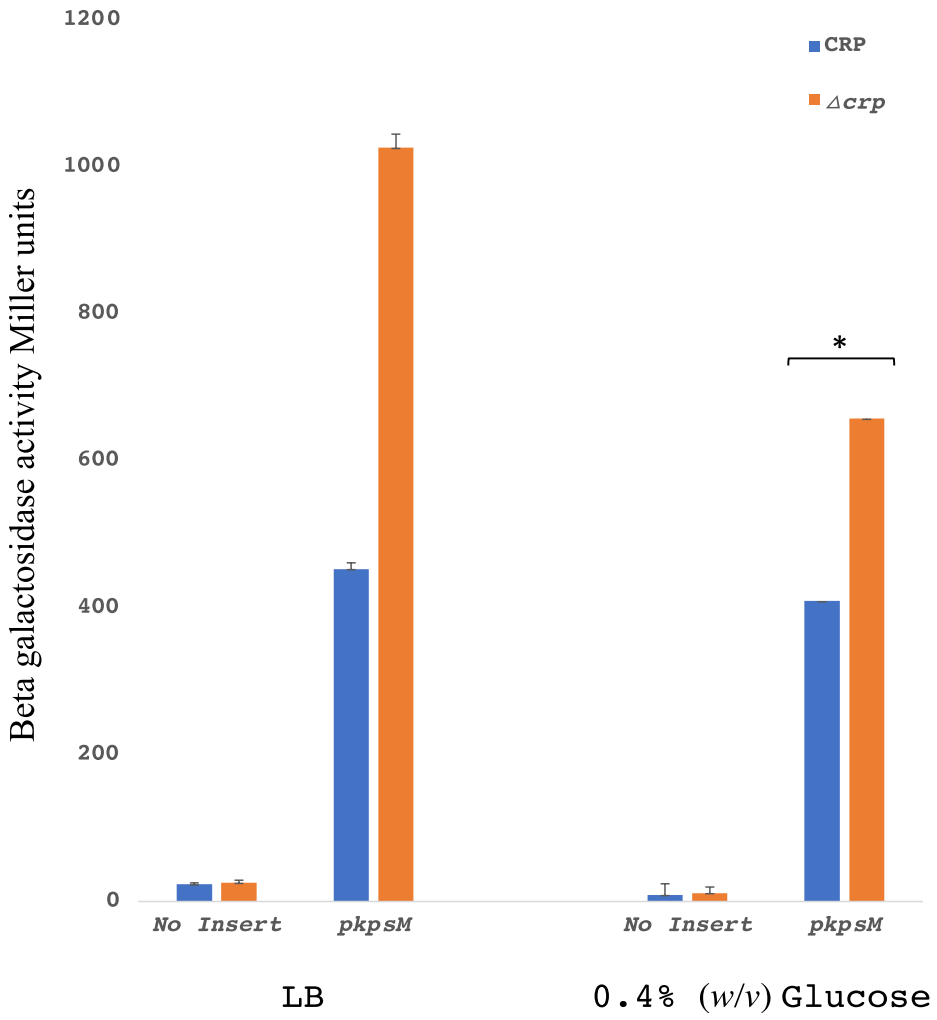


Figure 6.11: The effect of glucose on CRP-dependent repression of the *kpsMII* promoters

The bar chart illustrates ß- galactosidase expression driven by *kpsMII* promoters in *E. coli* K-12 strain M182 WT *Δ lac* (blue bars) and *E. coli* K-12 strain M182 *Δ lac Δcrp* (orange bars) carrying pRW224/*kpsM* 440. The left and right sides of the figure show data from experiments, respectively, without or with (*w/v*) glucose-supplemented LB media. 'No insert' represents the negative control using empty pRW224. Each measure indicates the average of three independent ß- galactosidase assays and the error bars represent the standard deviation (STDEV) and * Indicates $P < 0.000001$ using a student's *t*-test.

6.3 The *kpsMII* regulatory region: comparison between EAEC strain 042, UPEC strain CFT073 and UPEC strain UTI89

Following my previous comparisons of virulence factor genes between UPEC strain CFT073 and EAEC strain 042, the EAEC and UPEC *kpsMII* regulatory region sequences were aligned and are shown in Figure 6.12. The alignment shows that the P1 promoter -10 hexamer element is conserved. Interestingly, in CFT073, the -10 element of the second promoter is 5`-T**G**TAAA-3` which is predicted not to be functional, due to a natural mutation at position 2. In contrast, no such mutation is seen with the strain UPEC UTI89 strain. However, the CRP binding site in both UPEC strains is fully conserved. The alignment also shows a 116 bases of insertion, in the UPEC sequence (or a deletion, in the EAEC sequence) 256 bp downstream of the CRP binding site. As P2 should be functional in UPEC strain UTI89, it would be interesting to investigate how this affects mRNA translation and stability. Regardless to the insertion/deletion block, the *kpsMII* intergenic region from UPEC shows a high degree of conservation with the sequence from EAEC 042, with a similarity of 91% (approx. 45 mutations and two insertions in DNA sequence of *kpsMII* regulatory region from UPEC). Another point to mention in the alignment in Figure 6.12 is that *ops* element sequence for RfaH binding is conserved between the UPEC UTI89 and EAEC 042.

The long length of the 5'UTR in the *kpsMII* regulatory region increases the possibility of the existence of small regulatory RNAs (sRNA) in both UPEC and EAEC. The DNA sequences were applied to Pred^{sRNA} (Kumar *et al.,* 2021) and BLASTN analysis, however, no predictable sRNA sequence was recognised, suggesting that the existence of *trans*-acting sRNA here is unlikely.

```
kpsMI89          TTATTAATAGTTGTAATAAATCATTGAGTAACAATTGATAGGTCAAAACATATAGGATAA
kpsM042          TTATTAATAGTTGCAATAAATCATTGAGTAACAATTGATAGGCCAAAACATATAGGATAA
kpsM073          TTATTAATAGTTGTAATAAATCATTGAGTAACAATTGATAGGCCAAAACATATAGGATAA
                 ************* ***************************** ****************

kpsMI89          TTCTTATGTGATCTGTATTTTGTGTAGCCTGGAAATTAGTAAAATTACAGGAGATAATCA
kpsM042          TTCTTGTGTGATCTGTATTTTGTGTAGCTTGGAAATTAGTAAAATTCCTGGAGATAATCA
kpsM073          TTCTAGTGTGATCTGTATTTTGTGTAACCTGGAAGTTAGTAAAATTCCTGGAGATAATCA
                 ****  ******************** * ***** *********** * ***********

kpsMI89          GAAAGGGAATTTCAAATAAGCATAATAATTTCCAGGGAAACTATTCGTCGATTTTAAGAA
kpsM042          GAAAGGGAATTTCAAATAAGCATAATAATTTCCAGTGAAACTATTCGTTGATTTTAAGAA
kpsM073          GAAAGGGAATTTCAAATAAGCATAATAATTTCCAGTGAAACTATTCGTCGATTTTAAGAA
                 ********************************** ************ ************

kpsMI89          ATGTCTGCTCAGGTATATCTACTGAGGGATGGTGTTGGTTGTAACACTGGTTAAAATAAA
kpsM042          ATGTCTGCTCAGGTATATCTACTGAGGGATGGTGTTGGTTGTAACACTGGTTAAAATAAA
kpsM073          ATGTCTGCTCAGGTATATCTACTGAGGGATGGTGTTGGTTGTAACACTGGTTAAAATAAA
                 ************************************************************

kpsMI89          TACACAAAA-TAATCAAGATGTATATTTTAATCGACGAAAATAATTACCTTCGGGATTAT
kpsM042          TACACAAAA-TAATCAAGATGTATATTTTAATCGACGAAAATAATTACCTTCGGGATTAT
kpsM073          CACACAAAAATAATCAAGATGTATATTTTAATCGACAAAAATAATTACTCGCGGAATTAT
                  ******** ************************** ***********   *** *****

kpsMI89          TGATGCAACTTAAATAACACCATTTAAATGTGATATAAATCACAAATATGGCTGTAAAGA
kpsM042          TGATGCGACTTAAATAACACCATTTAAATGTGATATAAATCACAAATATGACTGTAAAGA
kpsM073          TGATGGGTGTTAAATAACACCGTTTAAATGTGATGTAAATCACAAATGTGACTGTAAAGA
                 *****    ************ ************ ************ ** *********

kpsMI89          GGGGGCTGTAGATATAAATAAGAAGTACACGAGTGAATTTTAAATAGGGAAATAGTTTCT
kpsM042          GGGGGCTGTAGATATAAATAAGAAGTACACGAGTGAATTTTAAATAGGGAAATAGTTTCT
kpsM073          GGGTGCTGTAGATATAAATAAGAAGTACATGAGTGAATTTTAAATAGGGAAATAGTTTCT
                 *** ************************ ******************************

kpsMI89          CGGTGAACAATTTATTGGTAATCAATCGCGTGCGTTCTGGTTTGAAACAGATTGCAAGTG
kpsM042          CGGTGAACAATTTATTGGTAATCAATCGCGTGCGTTCTGGTTTGAAACAGATTGCAGGTA
kpsM073          TGCTAAACAATTTATTGGTAATCAATCGCATGCGTTCTGGTTTGAGATAGATTACAGGTG
                  * * ************************ *************** * ***** ** **

kpsMI89          TTGTTACGCATAAAATCCATAGGGGATATTATAATCAAGTAGTTAACAGTAATAACAAAA
kpsM042          TTGTTACGCATAAAATCCATGGGGG-TATTATAATCAAGTAGTTAACAGTAATAACAAAA
kpsM073          TTACTACGCATAAAATTCATGGGTGGTATTATAATCAAGTAGTTAACAGTAATAGTGAAA
                 **  ************ *** ** * ***************************   ***

kpsMI89          AATAATTTC-----------------------------------------------------
kpsM042          AATAATTTC-----------------------------------------------------
kpsM073          AATAATTTCTTGCGAAATTAGCGGCCCGTTTTGAATAAGTGTGTTCCGGAGTAATCCAGA
                 *********

kpsMI89          ------------------------------------------------------------
kpsM042          ------------------------------------------------------------
kpsM073          ACGCATGCTGCTTTACATCTTTTTCAAGCCATATTGGTTAAGCAAAAAGAAGTAGAAATA


kpsMI89          -----CTGAGAAATTAACTCTGCATTCTAAAACGGCCAGTTTTCTGAAATTACCAGAAGA
kpsM042          -----CTGGGAAATTAACTCTACATTCTAAAACGGCCAGTTTTCTGAAATTACCAGAAGA
kpsM073          ATTTCCTGAGAAATTAACTCAACATTCTGAAACGGCCAGTTTTCTGAAATTACCAGAAGA
                      *** ***********  ****** ********************************

kpsMI89          CCGAATACATTCATGCCTGAAGAATGGATTAGAAGAAACGAGACGGAAATAGATCTATTT
kpsM042          CCGAATACATTCATGCCTGAAGAATGGATTAGAAGAAACGAGACGGAAATAGATCTATTT
kpsM073          CCGAATACATTCATGCCTGAAGAATGGATTAGAAGAAACGAGACGGAAATAGATCTATTT
                 ************************************************************
```

Figure 6.12: Base Sequence alignment of the UPEC CFT073, and EAEC 042, and UPEC strain UTI89 *kpsMII* regulatory regions

The figure shows the alignment of base sequence of the *kpsMII* regulatory region from UPEC UTI89 (*kpsMI89*); and EAEC 042 (*kpsM042*); and UPEC CFT073 (*kpsM073*). The two promoters -10 elements are highlighted with yellow shading and a key difference is underlined. Key bases for CRP binding are in gold typefaces. Triplets highlighted in light blue upstream of the CRP site indicate possible start codons, and *kpsMII* coding sequences are shown as underlined red bases. Green olive highlighted and dash underlined bases indicate the RfaH *ops* element from the Xue *et al.* (2009) study. Start markers denote identical base and dash mark means the deletion of the bases. (Thompson *et al.*, 2003).

6.4 Discussion

My results show that CRP plays a role in the *kpsMII* regulatory region, which, according to my data, contains at least two functional promoters and possibly a third weak promoter. The findings here are supported by the signal from CRP ChIP-seq in the *kpsMII* intergenic region (Chapter 4) and the comparable and continuing work reported by Xue *et al.* (2009) paper. However, Xue *et al.* (2009) did not test the *kpsMII* fragments in *Δ crp* cells. Interestingly, *kpsMII* P1 is clearly a more powerful than *kpsMII* P2, although the reasons for this are not evident from the base sequence. *kpsMII* P2 was optimally active only in the absence of CRP, indicating the repression mode of CRP at *kpsMII* P2. Not only does CRP affected the *kpsMII* P2 expression, but H-NS also affects the *kpsMII* P2 expression, as the promoter shows more activity after removing of H-NS binding sites in the *kpsM* 551 fragment (Figure 6.4B). Clearly more experiments are needed to understand the role, if any, of the long intergenic region before the *kpsMII* structural gene. And a further mutational analysis is needed in case of the third promoter (*kpsMII* P3). RfaH is important for virulence expression in *Vibrio vulnificus* (Garrett *et al.*, 2016), therefore, it is essential to study how can this antiterminator might influence the *kpsMII* P2.

It could be assumed that there might be a *trans-acting* sRNA embedded between the first promoter and the *kpsMII* gene, and the presence of the second promoter will interfere with this. It is possible that this sRNA is essential for the stability of *kpsMII* translation and the rest of the operon. However, Xue *et al.* (2009) had studied this possibility, but no a *trans-acting* sRNA was found there. Moreover, several more experiments of the role of CRP in *kpsMII* are needed to state if CRP affects the transcription of *kpsMII* indirectly.

Chapter 7

Final discussion and conclusions

7.1 Discussion

CRP was discovered as an activator for *lac* operon promoter and its mechanism of action was first reported by Malan *et al.* (1984). Following this, other promoters, beside the *lac* operon promoter, were found to be activated by CRP by different mechanisms (Busby and Ebright, 1999). Therefore, CRP was used as a model to study mechanisms of promoter activation, and it was found to function in two ways, known as Class I and Class II, to recruit RNAP to target promoters. However, studies also shown that CRP also can act as a repressor. The subsequent application of whole genome sequencing methods such as ChIP-seq showed that CRP binds to many targets besides what was discovered before. Also, CRP can bend DNA, and at many targets, this has no role in transcription activation or repression, leading to the idea that CRP can act as NAP at these targets.

The application of genomic methods has helped to study pathogen genomes. For example, in EAEC, a catalogue of genes under the control of the 'master' virulence regulator, AggR, was produced (Yasir *et al.*, 2019). It was assumed that all VFGs would be controlled by this bespoke regulator until the work of Rossiter *et al.* (2011) found that one of the virulence determinants, *pet*, was co-activated by the global regulators, CRP, and Fis. This could be explained by the need of SPATE for scavenge in nutritional starvation conditions. Other virulence determinants were found to be also co-activated by the CRP and Fis such as *sat* and *sigA* (Rossiter *et al.*, 2015). The outcome from these studies led to my project in which I asked if there are more virulence determinants that are dependent on CRP. Also, would the same pattern be seen in other pathogens?

The findings in Chapter three illustrates a comparison of the *pic* promoter region between EAEC and UPEC. Pathogens benefit from Pic during colonization of host intestines or bladder because of its mucinase activity (Flores-Sanchez *et al.*, 2020). The natural media of mucus is known to be nutritionally poor, and this creates the need for *pic* to be expressed in these circumstances. The results in Section 3.2.1 for the EAEC *pic* promoter and Section 3.3.1 for the UPEC *pic* promoter demonstrate CRP-dependence, as suggested from sequence analysis that showed a CRP binding site upstream of the *pic* gene in each case. My results show that the first base (+1) of the transcript differs between EAEC and UPEC. Thus, the EAEC *pic* promoter follows Class I CRP activation whilst the UPEC *pic* promoter follows Class II. My results suggest competing -10 element in *pic* promoter regions. Hence, my findings show an unusual

promoter architecture, with a single binding site for the transcriptional regulator, CRP, but two alternative downstream promoters, posing a challenge to incoming RNAP that has to make a choice of where to start.

Studies of the CRP consensus at the *pic* promoter in Section 3.4.4 underscored its importance in recruiting RNAP. Another significant finding was the impact of the length of 5`UTR on the translation of mRNA. The result in Section 3.4.6 showed that, when transcription of *lacZ* starts from the *pic p073* upstream promoter, translation of the *pic::lacZ* fusion gene was reduced, indicating that a challenging mRNA secondary structure occurs here which likely causes a decay in the transcript. In contrast, selection of the downstream promoter led to normal *pic::lacZ* translation. My findings suggest that *pic p073* is an ambiguous promoter where both Class I and Class II CRP-dependent activation is possible (Alhammadi *et al.*, 2022).

An unexpeted finding was that other regulators, like Fis and AggR, are involved in *pic* promoter regulation in EAEC and UPEC. My findings in Section 3.5.1 for Fis, and Section 3.5.2 for AggR, suggest a repressive role for both Fis and AggR, as their binding sites are downstream from the promoters. Furthermore, Fis levels increase in rapid growth conditions and promoters that are not needed in this condition *e.g., pic* will be repressed by Fis. Other work by previous PhD student, Muhammad Yasir, had shown that some CRP-dependent promoters repressed such as *pet* could be repressed by AggR (Yasir, unpublished). Here, I showed that AggR can also repress the *pic* promoter, which can be rationalised by the complementary roles of CRP and AggR as global and bespoke regulators, respectively. However, it is important to point out here that AggR is not present in UPEC (Wallace-Gadsden *et al.*, 2007). Togther my result lend support to the statement of "virulence determinants that are expressed in the early stage of infection time scale require a common or global regulator, until the bespoke regulator is expressed and drives the expression of the 'vicious' determinants. However, more studies and more examples are needed here to support this theory. The involvement of Pic in pathogen biofilm formation stage was illustrated by Liu *et al.* (2020). They found that the wild type EAEC could increase colonisation with differential in biofilm distribution and this was enhanced by mucus layer penetration driven by Pic. The findings of Liu *et al.* (2020) findings support the importance of Pic in colonisation and supported that Pic is expressed in the early stages of pathogenesis.

My research continued with an investigation of whether more EAEC strain 042 virulence genes beside *pet* and *pic*, have been recruited to the CRP regulon (Rossiter *et al.*, 2011; Alhammadi *et al.,* 2022). To address this, first, I tested the involvement of CRP in the formation of biofilm in EAEC 042. The result in Section 4.2 proves that CRP has a broad role. Therefore, guided by James Haycocks, I carried out CRP ChIP-seq on the EAEC 042 genome and the findings are catalogued in Chapter 4. We were able to find 322 CRP targets on the EAEC 042 chromosome, thirty-one of them were specific to this pathogen (Section 4.2.1) and only 10 CRP targets were found on the pAA virulence plasmid (Section 4.2.3). Many of these locations are positioned at promoter regions, which could indicate a role for CRP in transcription, but there were more sites where CRP appears to play a NAP role. In contrast to EAEC CRP ChIP-seq data, the distribution of CRP sites between intergenic and intragenic regions was approximately equal to that found in the *Mycobacterium tuberculosis* genome (Kahramanoglou *et al.*, 2014). Many of the CRP targets in the EAEC genome have been reported previously such as: *proP* (McLeod *et al.,* 2000), *galP* (Semsey *et al*., 2007), *gntK* (Izu *et al.,* 1997), *aaeR* (Raghavan *et al*., 2011); and others too (Appendix B; regulated by CRP).

Remarkably some apparently weak binding sites for CRP gave high read scores, such as: *cat* (4/10), *cspE* (3/10)*, hdfR* (4/10) and *fixA* (7/10). As expected, other CRP targets, with a complete match to the CRP consensuses, gave high score of reads. The best example was *kpsMII,* which is an interesting candidate for further investigation. DNA binding bases for CRP from ChIP-seq results showed a conserved bases 5`-TG**TG**AN$_6$**TC**A**C**A-3` (Section 4.2.2) which supported the importance of these bases in the interaction of CRP with the DNA template (reviewed in Lawson *et al.* 2004). The majority of regulatory CRP targets from my ChIP-seq data are involved in expression of metabolic enzymes, gene-specific transcriptional regulators, or transporters.

To try to understand the role of CRP at specific-EAEC 042 loci, I focused on CRP targets that are located in intergenic regions where the DNA site of CRP is not too close to the translation site of the gene. Most of my findings at new targets, such as *kpsMII* and *EC042_0536*, show a repression role for CRP. This repression mechanism is likely due to blocking the promoter elements from holo-RNAP binding, and this indicates that the corresponding gene products may be unnecessary when EAEC is faced with nutrient starvation mode. Some tested targets, like *EC042_0225* and *EC042_0224*, did not show any high promoter activity, indicating that CRP could play a NAP role in there.

In the following chapter, Chapter 6, I present the evidence from *in vivo* and *in vitro* investigations of the influence of CRP at the *kpsMII* regulatory region. Xue *et al.* (2009) previously reported the transcription regulation of *kpsMII*; however, the results of ChIP-seq detected a CRP binding site into the region after the reported promoter, indicating that CRP may have played a vital role in the expression of region three of group 2 capsule. The preliminary result in Section 5.4.1 indicated a repressive role for CRP in the *kpsMII* intergenic region. Taking into consideration the reported promoter of *kpsMII* P1 and the result from CRP ChIP-seq, the orientation of the repressed promoter by CRP in Section 5.4.1 needed to be investigated. The result of the previous point was shown in Section 6.2.2 and proved the existence of a second promoter oriented to the *kpsMII* gene (*kpsMII* P2).

Section 6.2.3 assessed the number of promoters that can drive the downstream gene expression, *kpsMII*, which concluded the existence of two promoters, showing the hexamer elements: 5'-TATTTT-3' (Xue *et al.*, 2009) and 5'-TATAAA-3' which are remarkably located between the CRP binding site 5'-TGTGATATAAATCACA-3'. In Section 6.2.4, *kpsMII* P1 showed powerful promoter activity, whereas *kpsMII* P2 showed weak promoter activity compared with *kpsMII* P1. This correlation is important as *kpsMII* is the first gene in the operon, and the promoter that drives the expression of region three will continue the translation to region two too (Xue *et al*., 2009; Stevens *et al*., 1994).

Thus, the repression role of CRP on the *kpsMII* second promoter may have contributed to the transcription of *kpsMII* in an indirect way that might interfere with the expression of sRNA. As most genes that encode outer membrane proteins are regulated by sRNA and are essential for mRNA translation stability (Johansen *et al.*, 2006). There are several significant differences in the alignment of the *kpsMII* promoter region between EAEC 042 and UPEC CFT073 and UPEC UTI89 strains, as shown in Chapter 6, Figure 6.12. The natural mutation in the *kpsMII* P2 in UPEC strain CFT073 (5'-TGTGATGTAAATCACA-3') but not in UPEC UTI89 due to conserved bases in P2 with EAEC 042 is one of the significant differences. Furthermore, the 116 bp, whose presence in UPEC UTI89 (Jia *et al*., 2017) raises the possibility that it plays a role in the *kpsMII* P2 expression.

7.2 Limitations

Although there were some good results were presented on this project, some limitations were recorded during the research, for example, the CRP ChIP-seq experiment in Chapter 4 the wild type strain, EAEC 042, was only used there and I did not investigate the EAEC 042 *Δcrp* strain. Because of the large quantities of data produced from the CRP ChIP-seq experiment, Appendix B, it was difficult to investigate all of the CRP targets, due to time limitations, and the possibility of false positive results remains. Another difficulty was ignorance of the involvement of other regulators at many CRP tragets: these could well have co-activator or co-repressor roles. Perhaps the most serious problems I faced it was the predictions of **CRP** binding sites from ChIP-seq data via the MEME SUITE software (Figure 4.6B). The selected motif by MEME was 5`-**TGTGA**YBBNBA**TCACA**WWWTT-3` where the bases:
Y=T or C,
B= G or C or T,
N= G or A or Tor C,
and W=A or T.
The presence of 5`-WWWTT-3` at the end of the motif did affect the chosen motif from the sequence window,101 bp, and in some cases gave the wrong motif, which was identified before in some studies, such as the case of *malT*. The proposed CRP DNA site in Chapon and Kolb's (1983) research was 11 bp upstream of where the MEME result suggested it.

7.3 Future research directions

The involvement of the global regulator, CRP, in the regulation of virulence genes had been reported in many papers before this project (Section 1.12), yet more virulence genes in CRP regulon need to be discovered from different strains. The outcomes from these researches can help to understand the behave of pathogens in serious situation like the lack of nutritions in the surrounding and how to cope with it.

Some suggestions could direct this project in the future are pointed here. An experiment could be considered in Chapter 4 for a comparison purposes is RNA-seq technique which can prove if there are any mRNA comes from the binding of the CRP on it site. Furthermore, the results in Chapter four of CRP ChIP-seq needs more investigations on the CRP new targets where the genes are common with *E. coli* K-12 to update CRP regulon network in RegulonDB database (Santos-Zavaleta *et al.*, 2019). Further research on the regulation of *kpsMII* is require to

determine if CRP involves at the post-transcriptional of *kpsMII* expression. Another investigation should be considered is testing the translation efficiency of *kpsMII* gene using the strain EAEC 042 with ±CRP background via reverse transcriptase experiments. Also, the promoter activity of *kpsMII* P2 from UPEC UTI89 need further investigation to compare it with *kpsMII* P2 from EAEC 042 and build a model suit these findings. Although Chapter five answered the regulation of CRP on EAEC specific genes in general, the question of CRP repression on *EC042_0536* remain to be answered.

7.4 Conclusions

This study helped to uncover how can pathogens cope the challenging circumstances by responding to the environment signals and exploit a global regulator to regulate the essential enzymes in early stages of infections of EAEC and UPEC.

The remarkable outcomes from this study are:

- Involvement of CRP to activate the transcription of some virulence factors like *pic*.
- Unusual architecture of the *pic* promoters from Enteroaggregative and Uropathogenic *Escherichia coli*.
- The remarkable of a conserved two promoters drive same gene from two different strains whereas one follow Class I activation and the other follows Class I/II activations.
- Most of CRP-dependent promoters that drive virulence genes are repressed by the master regulator of virulence genes in EAEC 042, AggR.
- CRP plays indirect regulation role on *kpsMII* regulatory region.
- The closer of binding site of a repressor to the -10 element of the promoter the more effect of repression will occur.
- The impact of the leader sequence lengths on the mRNA translation stability.

8.1 List of References

Afgan, E., Baker, D., Batut, B., Van Den Beek, M., Bouvier, D., Čech, M., Chilton, J., Clements, D., Coraor, N., Grüning, B.A. and Guerler, A., 2018. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. *Nucleic acids research*, 46(W1), pp. W537-W544.

Aiyar, S., Gourse, R. and Ross, W., 1998. Upstream A-tracts increase bacterial promoter activity through interactions with the RNA polymerase subunit. *Proceedings of the National Academy of Sciences*, 95(25), pp.14652-14657.

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K. and Walter, P., 2002. Genesis, modulation, and regeneration of skeletal muscle. In *Molecular Biology of the Cell. 4th edition*. Garland Science.

Alhammadi, M., Godfrey, R., Ingram, J., Singh, G., Bathurst, C., Busby, S. and Browning, D., 2022. Novel organisation and regulation of the *pic* promoter from enteroaggregative and uropathogenic *Escherichia coli*. *Virulence*, 13(1), pp.1393-1406.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J., 1990. Basic local alignment search tool. *Journal of molecular biology*, 215(3), pp.403-410.

Azam, A.T., Iwata, A., Nishimura, A., Ueda, S. and Ishihama, A., 1999. Growth phase-dependent variation in protein composition of the *Escherichia coli* nucleoid. *Journal of bacteriology*, 181(20), pp.6361-6370.

Baba, T., Ara, T., Hasegawa, M., Takai, Y., Okumura, Y., Baba, M., Datsenko, K.A., Tomita, M., Wanner, B.L. and Mori, H., 2006. Construction of *Escherichia coli* K-12 in-frame, single-gene knockout mutants: the Keio collection. *Molecular systems biology*, 2(1), pp.2006-0008.

Bailey, T.L., Boden, M., Buske, F.A., Frith, M., Grant, C.E., Clementi, L., Ren, J., Li, W.W. and Noble, W.S., 2009. MEME SUITE: tools for motif discovery and searching. *Nucleic acids research*, 37(suppl_2), pp. W202-W208.

Barker, M.M., Gaal, T., Josaitis, C.A. and Gourse, R.L., 2001. Mechanism of regulation of transcription initiation by ppGpp. I. Effects of ppGpp on transcription initiation *in vivo* and *in vitro*. *Journal of molecular biology*, 305(4), pp.673-688.

Baudry, B., Savarino, S.J., Vial, P., Kaper, J.B. and Levine, M.M., 1990. A sensitive and specific DNA probe to identify enteroaggregative *Escherichia coli*, a recently discovered diarrheal pathogen. *Journal of Infectious Diseases*, 161(6), pp.1249-1251.

Behrens, M., Sheikh, J. and Nataro, J., 2002. Regulation of the Overlapping *pic/set* Locus in *Shigella flexneri* and Enteroaggregative *Escherichia coli*. *Infection and Immunity*, 70(6), pp.2915-2925.

Bennett, V. and Gilligan, D.M., 1993. The spectrin-based membrane skeleton and micron-scale organization of the plasma membrane. *Annual review of cell biology*, 9(1), pp.27-66.

Bell, A., Gaston, K., Williams, R., Chapman, K., Kolb, A., Buc, H., Minchin, S., Williams, J. and Busby, S., 1990. Mutations that alter the ability of the *Escherichia coli* cyclic AMP receptor protein to activate transcription. *Nucleic acids research*, 18(24), pp.7243-7250.

Bertin, Y., Martin, C., Girardeau, J.P., Pohl, P. and Contrepois, M., 1998. Association of genes encoding P fimbriae, CS31A antigen and EAST 1 toxin among CNF1-producing *Escherichia coli* strains from cattle with septicemia and diarrhea. *FEMS microbiology letters*, 162(2), pp.235-239.

Bien, J., Sokolova, O. and Bozko, P., 2012. Role of uropathogenic *Escherichia coli* virulence factors in development of urinary tract infection and kidney damage. *International journal of nephrology*, 2012.

Blumer, C., Kleefeld, A., Lehnen, D., Heintz, M., Dobrindt, U., Nagy, G., Michaelis, K., Emödy, L., Polen, T., Rachel, R. and Wendisch, V.F., 2005. Regulation of type 1 fimbriae synthesis and biofilm formation by the transcriptional regulator LrhA of *Escherichia coli*. *Microbiology*, 151(10), pp.3287-3298.

Borukhov, S. and Nudler, E., 2008. RNA polymerase: the vehicle of transcription. *Trends in microbiology*, 16(3), pp.126-134.

Bradley, M.D., Beach, M.B., de Koning, A.J., Pratt, T.S. and Osuna, R., 2007. Effects of Fis on *Escherichia coli* gene expression during different growth stages. *Microbiology*, 153(9), pp.2922-2940.

Brown, N.L., Stoyanov, J.V., Kidd, S.P. and Hobman, J.L., 2003. The MerR family of transcriptional regulators. *FEMS microbiology reviews*, 27(2-3), pp.145-163.

Browning, D.F. and Busby, S.J., 2004. The regulation of bacterial transcription initiation. *Nature Reviews Microbiology*, 2(1), pp.57-65.

Browning, D.F. and Busby, S.J., 2016. Local and global regulation of transcription initiation in bacteria. *Nature Reviews Microbiology*, 14(10), pp.638-650.

Buchet, A., Eichler, K. and Mandrand-Berthelot, M., 1998. Regulation of the Carnitine Pathway in *Escherichia coli*: Investigation of the cai-fix Divergent Promoter Region. *Journal of Bacteriology*, 180(10), pp.2599-2608.

Busby, S.J., 2019. Transcription activation in bacteria: ancient and modern. *Microbiology*, 165(4), pp.386-395.

Busby, S. and Ebright, R.H., 1999. Transcription activation by catabolite activator protein (CAP). *Journal of molecular biology*, 293(2), pp.199-213.

Busby, S., Kotlarz, D. and Buc, H., 1983. Deletion mutagenesis of the *Escherichia coli* galactose operon promoter region. *Journal of Molecular Biology*, 167(2), pp.259-274.

Carver, T., Harris, S., Berriman, M., Parkhill, J. and McQuillan, J., 2011. Artemis: an integrated platform for visualization and analysis of high-throughput sequence-based experimental data. *Bioinformatics*, 28(4), pp.464-469.

Casadaban, M.J. and Cohen, S.N., 1980. Analysis of gene control signals by DNA fusion and cloning in *Escherichia coli*. *Journal of molecular biology*, 138(2), pp.179-207.

Chakraborty, A., Wang, D., Ebright, Y.W., Korlann, Y., Kortkhonjia, E., Kim, T., Chowdhury, S., Wigneshweraraj, S., Irschik, H., Jansen, R. and Nixon, B.T., 2012. Opening and closing of the bacterial RNA polymerase clamp. *Science*, 337(6094), pp.591-595.

Chapon, C. and Kolb, A., 1983. Action of CAP on the malT promoter *in vitro*. *Journal of bacteriology*, 156(3), pp.1135-1143.

Chaudhuri, R.R., Loman, N.J., Snyder, L.A., Bailey, C.M., Stekel, D.J. and Pallen, M.J., 2007. x BASE2: a comprehensive resource for comparative bacterial genomics. *Nucleic acids research*, 36(suppl_1), pp. D543-D546.

Chaudhuri, R.R., Sebaihia, M., Hobman, J.L., Webber, M.A., Leyton, D.L., Goldberg, M.D., Cunningham, A.F., Scott-Tucker, A., Ferguson, P.R., Thomas, C.M. and Frankel, G., 2010. Complete genome sequence and comparative metabolic profiling of the prototypical enteroaggregative *Escherichia coli* strain 042. *PloS one*, 5(1), p. e8801.

Chow, J., Tang, H. and Mazmanian, S.K., 2011. Pathobionts of the gastrointestinal microbiota and inflammatory disease. *Current opinion in immunology*, 23(4), pp.473-480.

Choy, H.E. and Adhya, S., 1992. Control of gal transcription through DNA looping: inhibition of the initial transcribing complex. *Proceedings of the National Academy of Sciences*, 89(23), pp.11264-11268.

Croxen, M.A. and Finlay, B.B., 2010. Molecular mechanisms of *Escherichia coli* pathogenicity. *Nature Reviews Microbiology*, 8(1), pp.26-38.

Cuthbertson, L., Kos, V. and Whitfield, C., 2010. ABC Transporters Involved in Export of Cell Surface Glycoconjugates. *Microbiology and Molecular Biology Reviews*, 74(3), pp.341-362.

Cuthbertson, L., Mainprize, I.L., Naismith, J.H. and Whitfield, C., 2009. Pivotal roles of the outer membrane polysaccharide export and polysaccharide copolymerase protein families in export of extracellular polysaccharides in gram-negative bacteria. *Microbiology and Molecular Biology Reviews*, 73(1), pp.155-177.

Dhiman, A. and Schleif, R., 2000. Recognition of overlapping nucleotides by AraC and the sigma subunit of RNA polymerase. *Journal of Bacteriology*, 182(18), pp.5076-5081.

Dudley, E.G., Thomson, N.R., Parkhill, J., Morin, N.P. and Nataro, J.P., 2006. Proteomic and microarray characterization of the AggR regulon identifies a pheU pathogenicity island in enteroaggregative *Escherichia coli. Molecular microbiology*, 61(5), pp.1267-1282.

Dufresne, K. and Daigle, F., 2021. Identification of Crp as a novel regulator of the Std fimbrial expression in *Salmonella*. *Microbiology*, 167(3), p.001022.

Ebright, R.H., 1993. Transcription activation at class I CAP-dependent promoters. *Molecular microbiology*, 8(5), pp.797-802.

Ebright, R.H. and Busby, S., 1995. The *Escherichia coli* RNA polymerase α subunit: structure and function. *Current opinion in genetics and development*, 5(2), pp.197-203.

El Mouali Benomar, Y., Gaviria Cantín, T.C., Sánchez-Romero, M.A., Gibert, M., Westermann, A.J., Vogel, J. and Balsalobre Parra, C., 2018. CRP-cAMP mediates silencing of Salmonella virulence at the post-transcriptional level. *PLoS Genetics,* 2018, vol. 14, num. 6, p. e1007401.

Elias Jr, W.P., Czeczulin, J.R., Henderson, I.R., Trabulsi, L.R. and Nataro, J.P., 1999. Organization of biogenesis genes for aggregative adherence fimbria II defines a virulence gene cluster in enteroaggregative *Escherichia coli*. *Journal of Bacteriology*, 181(6), pp.1779-1785.

Elias, W.P. and Navarro-Garcia, F., 2016. Enteroaggregative *Escherichia coli* (EAEC). In *Escherichia coli in the Americas* (pp. 27-57). Springer, Cham.

Emsley, P., McDermott, G., Charles, I.G., Fairweather, N.F. and Isaacs, N.W., 1994. Crystallographic characterization of pertactin, a membrane-associated protein from *Bordetella pertussis*. *Journal of molecular biology*, 235(2), pp.772-773.

Estrem, S., Gaal, T., Ross, W. and Gourse, R., 1998. Identification of an UP-element consensus sequence for bacterial promoters. *Proceedings of the National Academy of Sciences*, 95(17), pp.9761-9766.

Evrard, B., Balestrino, D., Dosgilbert, A., Bouya-Gachancard, J.L., Charbonnel, N., Forestier, C. and Tridon, A., 2010. Roles of capsule and lipopolysaccharide O antigen in interactions of human monocyte-derived dendritic cells and *Klebsiella pneumoniae*. *Infection and immunity*, 78(1), pp.210-219.

Farfan, M.J., Inman, K.G. and Nataro, J.P., 2008. The major pilin subunit of the AAF/II fimbriae from enteroaggregative *Escherichia coli* mediates binding to extracellular matrix proteins. *Infection and immunity*, 76(10), pp.4378-4384.

Feklistov, A., 2013. RNA polymerase: in search of promoters. *Annals of the New York Academy of Sciences*, 1293(1), pp.25-32.

Feuerbacher, L., Burgum, A. and Kolodrubetz, D., 2011. The cyclc-AMP receptor protein (CRP) regulon in *Aggregatibacter actinomycetemcomitans* includes leukotoxin. *Microbial Pathogenesis*, 51(3), pp.133-141.

Fischer, W., Schmidt, M., Jann, B. and Jann, K., 1982. Structure of the *Escherichia coli* K2 capsular antigen. Stereochemical configuration of the glycerophosphate and distribution of galactopyranosyl and galactofuranosyl residues. *Biochemistry*, 21(6), pp.1279-1284.

Flores-Sanchez, F., Chavez-Dueñas, L., Sanchez-Villamil, J. and Navarro-Garcia, F., 2020. Pic protein from enteroaggregative *E. coli* induces different mechanisms for its dual activity as a mucus secretagogue and a mucinase. *Frontiers in immunology*, 11, p.564953.

Foxman, B., 2014. Urinary tract infection syndromes: occurrence, recurrence, bacteriology, risk factors, and disease burden. *Infectious Disease Clinics*, *28*(1), pp.1-13.

Gallant, J., Irr, J. and Cashel, M., 1971. The mechanism of amino acid control of guanylate and adenylate biosynthesis. *Journal of Biological Chemistry*, 246(18), pp.5812-5816.

Garrett, S.B., Garrison-Schilling, K.L., Cooke, J.T. and Pettis, G.S., 2016. Capsular polysaccharide production and serum survival of *Vibrio vulnificus* are dependent on antitermination control by RfaH. *FEBS letters*, 590(24), pp.4564-4572.

Gaston, K., Chan, B., Kolb, A., Fox, J. and Busby, S., 1988. Alterations in the binding site of the cyclic AMP receptor protein at the *Escherichia coli* galactose operon regulatory region. *Biochemical Journal*, 253(3), pp.809-818.

Goeddel, D.V., 1990. Gene expression technology.

Görke, B. and Stülke, J., 2008. Carbon catabolite repression in bacteria: many ways to make the most out of nutrients. *Nature Reviews Microbiology*, 6(8), pp.613-624.

Grainger, D. and Busby, S., 2008. Methods for studying global patterns of DNA binding by bacterial transcription factors and RNA polymerase. *Biochemical Society Transactions*, 36(4), pp.754-757.

Grainger, D.C., Hurd, D., Harrison, M., Holdstock, J. and Busby, S.J., 2005. Studies of the distribution of *Escherichia coli* cAMP-receptor protein and RNA polymerase along the *E. coli* chromosome. *Proceedings of the National Academy of Sciences*, 102(49), pp.17693-17698.

Grant, J. and Stothard, P., 2008. The CGView Server: a comparative genomics tool for circular genomes. *Nucleic Acids Research,* 36(2), pp. 181-184.

Griesenbeck, J., Tschochner, H. and Grohmann, D., 2017. Structure and function of RNA polymerases and the transcription machineries. *Macromolecular Protein Complexes*, pp.225-270.

Gutiérrez-Jiménez, J., Arciniega, I. and Navarro-García, F., 2008. The serine protease motif of Pic mediates a dose-dependent mucolytic activity after binding to sugar constituents of the mucin substrate. *Microbial pathogenesis*, 45(2), pp.115-123.

Hacker, J. and Blum-Oehler, G., 2007. In appreciation of Theodor Escherich. *Nature Reviews Microbiology*, 5(12), pp.902-902.

Hardin, J., Bertoni, G.P. and Kleinsmith, L.J., 2017. *Becker's World of the Cell, eBook*. Pearson Higher Ed.

Harrington, S.M., Sheikh, J., Henderson, I.R., Ruiz-Perez, F., Cohen, P.S. and Nataro, J.P., 2009. The Pic protease of enteroaggregative *Escherichia coli* promotes intestinal colonization and growth in the presence of mucin. *Infection and immunity*, 77(6), pp.2465-2473.

Henderson, I.R., Czeczulin, J., Eslava, C., Noriega, F. and Nataro, J.P., 1999. Characterization of Pic, a secreted protease of *Shigella flexneri* and enteroaggregative *Escherichia coli*. *Infection and immunity*, 67(11), pp.5587-5596.

Henderson, I.R., Navarro-Garcia, F., Desvaux, M., Fernandez, R.C. and Ala'Aldeen, D., 2004. Type V protein secretion pathway: the autotransporter story. *Microbiology and molecular biology reviews*, 68(4), pp.692-744.

Henderson, I.R., Navarro-Garcia, F. and Nataro, J.P., 1998. The great escape: structure and function of the autotransporter proteins. *Trends in microbiology*, 6(9), pp.370-378.

Hoiczyk, E., 2000. Structure and sequence analysis of *Yersinia* YadA and Moraxella UspAs reveal a novel class of adhesins. *The EMBO Journal*, 19(22), pp.5989-5999.

Islam, M.S., Pallen, M.J. and Busby, S.J., 2011. A cryptic promoter in the LEE1 regulatory region of enterohaemorrhagic *Escherichia coli*: promoter specificity in AT-rich gene regulatory regions. *Biochemical Journal*, 436(3), pp.681-686.

Itoh, Y., Nagano, I., Kunishima, M. and Ezaki, T., 1997. Laboratory investigation of enteroaggregative Escherichia coli O untypeable: H10 associated with a massive outbreak of gastrointestinal illness. *Journal of clinical microbiology*, 35(10), pp.2546-2550.

Iwanaga, M., Song, T., Higa, N., Toma, C., Nakasone, N. and Kakinohana, S., 2002. Enteroaggregative *Escherichia coli*: incidence in Japan and usefulness of the clump-formation test. *Journal of infection and chemotherapy*, 8(4), pp.345-348.

Izu, H., Adachi, O. and Yamada, M., 1997. Gene organization and transcriptional regulation of the *gntRKU* operon involved in gluconate uptake and catabolism of *Escherichia coli*. *Journal of molecular biology*, 267(4), pp.778-793.

Jia, J., King, J.E., Goldrick, M.C., Aldawood, E. and Roberts, I.S., 2017. Three tandem promoters, together with IHF, regulate growth phase dependent expression of the *Escherichia coli* kps capsule gene cluster. *Scientific reports*, 7(1), pp.1-11.

Jofré, M.R., Rodríguez, L.M., Villagra, N.A., Hidalgo, A.A., Mora, G.C. and Fuentes, J.A., 2014. RpoS integrates CRP, Fis, and PhoP signaling pathways to control *Salmonella Typhi* hlyE expression. *BMC microbiology*, 14(1), pp.1-12.

Johansen, J., Rasmussen, A.A., Overgaard, M. and Valentin-Hansen, P., 2006. Conserved small non-coding RNAs that belong to the σE regulon: role in down-regulation of outer membrane proteins. *Journal of molecular biology*, 364(1), pp.1-8.

Johansson, M.E., Larsson, J.M.H. and Hansson, G.C., 2011. The two mucus layers of colon are organized by the MUC2 mucin, whereas the outer layer is a legislator of host–microbial interactions. *Proceedings of the national academy of sciences*, 108(supplement_1), pp.4659-4665.

Kahramanoglou, C., Cortes, T., Matange, N., Hunt, D.M., Visweswariah, S.S., Young, D.B. and Buxton, R.S., 2014. Genomic mapping of cAMP receptor protein (CRPMt) in *Mycobacterium tuberculosis*: relation to transcriptional start sites and the role of CRPMt as a transcription factor. *Nucleic acids research*, 42(13), pp.8320-8329.

Kalia, D., Merey, G., Nakayama, S., Zheng, Y., Zhou, J., Luo, Y., Guo, M., Roembke, B.T. and Sintim, H.O., 2013. Nucleotide, c-di-GMP, c-di-AMP, cGMP, cAMP, (p) ppGpp signaling in bacteria and implications in pathogenesis. *Chemical Society Reviews*, 42(1), pp.305-341.

Kaper, J.B., Nataro, J.P. and Mobley, H.L., 2004. Pathogenic *Escherichia coli*. *Nature reviews microbiology*, 2(2), pp.123-140.

Kim, T.J., Chauhan, S., Motin, V.L., Goh, E.B., Igo, M.M. and Young, G.M., 2007. Direct transcriptional control of the plasminogen activator gene of *Yersinia pestis* by the cyclic AMP receptor protein. *Journal of bacteriology*, 189(24), pp.8890-8900.

Kim, J. and Khan, W., 2013. Goblet Cells and Mucins: Role in Innate Defense in Enteric Infections. *Pathogens*, 2(1), pp.55-70.

Klein, R.D., and Hultgren, S.J., 2020. Urinary tract infections: microbial pathogenesis, host–pathogen interactions and new treatment strategies. *Nature Reviews Microbiology*, 18(4), pp.211-226.

Kolb, A., Kotlarz, D., Kusano, S. and Ishihama, A., 1995. Selectivity of the *Escherichia coli* RNA polymerase Eσ 38 for overlapping promoters and ability to support CRP activation. *Nucleic acids research*, 23(5), pp.819-826.

Kröncke, K., Golecki, J. and Jann, K., 1990. Further electron microscopic studies on the expression of *Escherichia coli* group II capsules. *Journal of Bacteriology*, 172(6), pp.3469-3472.

Kumar, K., Chakraborty, A. and Chakrabarti, S., 2021. PresRAT: a server for identification of bacterial small-RNA sequences and their targets with probable binding region. *RNA biology*, 18(8), pp.1152-1159.

Kumari, S., Beatty, C.M., Browning, D.F., Busby, S.J., Simel, E.J., Hovel-Miner, G. and Wolfe, A.J., 2000. Regulation of acetyl coenzyme A synthetase in *Escherichia coli*. *Journal of bacteriology*, 182(15), pp.4173-4179.

Lawrence, J.G. and Roth, J.R., 1996. Evolution of coenzyme B12 synthesis among enteric bacteria: evidence for loss and reacquisition of a multigene complex. *Genetics*, 142(1), pp.11-24.

Lawson, C.L., Swigon, D., Murakami, K.S., Darst, S.A., Berman, H.M. and Ebright, R.H., 2004. Catabolite activator protein: DNA binding and transcription activation. *Current opinion in structural biology*, 14(1), pp.10-20.

Lee, D.J., Minchin, S.D. and Busby, S.J., 2012. Activating transcription in bacteria. *Annual review of microbiology*, 66(1), pp.125-152.

Liu, L., Saitz-Rojas, W., Smith, R., Gonyar, L., In, J.G., Kovbasnjuk, O., Zachos, N.C., Donowitz, M., Nataro, J.P. and Ruiz-Perez, F., 2020. Mucus layer modeling of human colonoids during infection with enteroaggragative *E. coli*. *Scientific reports*, 10(1), pp.1-14.

Lodge, J., Fear, J., Busby, S., Gunasekaran, P. and Kamini, N.R., 1992. Broad host range plasmids carrying the *Escherichia coli* lactose and galactose operons. *FEMS microbiology letters*, 95(2-3), pp.271-276.

Lybecker, M., Bilusic, I. and Raghavan, R., 2014. Pervasive transcription: detecting functional RNAs in bacteria. *Transcription*, 5(4), p. e944039.

Malan, T.P., Kolb, A., Buc, H. and McClure, W.R., 1984. Mechanism of CRP-cAMP activation of *lac* operon transcription initiation activation of the P1 promoter. *Journal of molecular biology*, 180(4), pp.881-909.

Mallik, P., Paul, B.J., Rutherford, S.T., Gourse, R.L. and Osuna, R., 2006. DksA is required for growth phase-dependent regulation, growth rate-dependent control, and stringent control of fis expression in *Escherichia coli*. *Journal of bacteriology*, 188(16), pp.5775-5782.

Manneh-Roussel, J., Haycocks, J.R., Magán, A., Perez-Soto, N., Voelz, K., Camilli, A., Krachler, A.M. and Grainger, D.C., 2018. cAMP receptor protein controls *Vibrio cholerae* gene expression in response to host colonization. *MBio*, 9(4), pp. e00966-18.

Mathew, R. and Chatterji, D., 2006. The evolving story of the omega subunit of bacterial RNA polymerase. *Trends in microbiology*, 14(10), pp.450-455.

McLeod, S.M., Xu, J. and Johnson, R.C., 2000. Coactivation of the RpoS-dependent proP P2 promoter by Fis and cyclic AMP receptor protein. *Journal of Bacteriology*, 182(15), pp.4180-4187.

Mejía-Almonte, C., Busby, S.J., Wade, J.T., van Helden, J., Arkin, A.P., Stormo, G.D., Eilbeck, K., Palsson, B.O., Galagan, J.E. and Collado-Vides, J., 2020. Redefining fundamental concepts of transcription initiation in bacteria. *Nature Reviews Genetics*, 21(11), pp.699-714.

Mi, H., Muruganujan, A., Ebert, D., Huang, X. and Thomas, P.D., 2019. PANTHER version 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. *Nucleic acids research*, 47(D1), pp. D419-D426.

Morin, N., Santiago, A.E., Ernst, R.K., Guillot, S.J. and Nataro, J.P., 2013. Characterization of the AggR regulon in enteroaggregative *Escherichia coli*. *Infection and immunity*, 81(1), pp.122-132.

Moxon, R., Bayliss, C. and Hood, D., 2006. Bacterial contingency loci: the role of simple sequence DNA repeats in bacterial adaptation. *Annu. Rev. Genet.*, 40, pp.307-333.

Müller, C.M., Åberg, A., Straseviçiene, J., Emődy, L., Uhlin, B.E. and Balsalobre, C., 2009. Type 1 fimbriae, a colonization factor of uropathogenic *Escherichia coli*, are controlled by the metabolic sensor CRP-cAMP. *PLoS pathogens*, 5(2), p. e1000303.

Müller, J., Oehler, S. and Müller-Hill, B., 1996. Repression of *lac* Promoter as a Function of Distance, Phase and Quality of an Auxiliary *lac* Operator. *Journal of molecular biology*, 257(1), pp.21-29.

Murakami, K.S., 2015. Structural biology of bacterial RNA polymerase. *Biomolecules*, 5(2), pp.848-864.

Nataro, J.P., 2005. Enteroaggregative *Escherichia coli* pathogenesis. *Current opinion in gastroenterology*, 21(1), pp.4-8.

Nataro, J.P. and Kaper, J.B., 1998. Diarrheagenic *escherichia coli*. *Clinical microbiology reviews*, 11(1), pp.142-201.

Nataro, J.P., Kaper, J.B., Robins-Browne, R.O.Y., Prado, V., Vial, P. and Levine, M.M., 1987. Patterns of adherence of diarrheagenic *Escherichia coli* to HEp-2 cells. *The Pediatric infectious disease journal*, 6(9), pp.829-831.

Nataro, J.P., Yikang, D., Cookson, S., Cravioto, A., Savarino, S.J., Guers, L.D., Levine, M.M. and Tacket, C.O., 1995. Heterogeneity of enteroaggregative *Escherichia coli* virulence demonstrated. *Journal of Infectious Diseases*, 171(2), pp.465-468.

Navarro-García, F., Canizalez-Roman, A., Burlingame, K.E., Teter, K. and Vidal, J.E., 2007. Pet, a non-AB toxin, is transported and translocated into epithelial cells by a retrograde trafficking pathway. *Infection and immunity*, 75(5), pp.2101-2109.

Navarro-García, F., Canizalez-Roman, A., Luna, J., Sears, C. and Nataro, J.P., 2001. Plasmid-encoded toxin of enteroaggregative *Escherichia coli* is internalized by epithelial cells. *Infection and Immunity*, 69(2), pp.1053-1060.

Navarro-Garcia, F. and Elias, W.P., 2011. Autotransporters and virulence of enteroaggregative *E. coli*. *Gut microbes*, 2(1), pp.13-24.

Navarro-Garcia, F., Gutierrez-Jimenez, J., Garcia-Tovar, C., Castro, L.A., Salazar-Gonzalez, H. and Cordova, V., 2010. Pic, an autotransporter protein secreted by different pathogens in the Enterobacteriaceae family, is a potent mucus secretagogue. *Infection and immunity*, 78(10), pp.4101-4109.

Nickels, B.E., Dove, S.L., Murakami, K.S., Darst, S.A. and Hochschild, A., 2002. Protein–protein and protein–DNA interactions of σ70 region 4 involved in transcription activation by λcI. *Journal of molecular biology*, 324(1), pp.17-34.

Nishino, K., Senda, Y. and Yamaguchi, A., 2008. CRP Regulator Modulates Multidrug Resistance of *Escherichia coli* by Repressing the mdtEF Multidrug Efflux Genes. *The Journal of Antibiotics*, 61(3), pp.120-127.

Okeke I. N., Lamikanra A., Czeczulin J., Dubovsky F., Kaper J. B., and Nataro J. P., 2000. Heterogeneous virulence of enteroaggregative *Escherichia coli* strains isolated from children in Southwest Nigeria. *Journal of Infectious Diseases*, vol. 181, no. 1, pp. 252–260.

Ophir, T. and Gutnick, D.L., 1994. A role for exopolysaccharides in the protection of microorganisms from desiccation. *Applied and environmental microbiology*, 60(2), pp.740-745.

Osek, J., 2003. Detection of the enteroaggregative *Escherichia coli* heat-stable enterotoxin 1 (EAST1) gene and its relationship with fimbrial and enterotoxin markers in *E. coli* isolates from pigs with diarrhoea. *Veterinary microbiology*, 91(1), pp.65-72.

Parham, N.J., Srinivasan, U., Desvaux, M., Foxman, B., Marrs, C.F. and Henderson, I.R., 2004. PicU, a second serine protease autotransporter of uropathogenic *Escherichia coli*. *FEMS microbiology letters*, 230(1), pp.73-83.

Paul, B.J., Barker, M.M., Ross, W., Schneider, D.A., Webb, C., Foster, J.W. and Gourse, R.L., 2004. DksA: a critical component of the transcription initiation machinery that potentiates the regulation of rRNA promoters by ppGpp and the initiating NTP. *Cell*, 118(3), pp.311-322.

Paul, B.J., Berkmen, M.B. and Gourse, R.L., 2005. DksA potentiates direct activation of amino acid promoters by ppGpp. *Proceedings of the National Academy of Sciences*, 102(22), pp.7823-7828.

Perederina, A., Svetlov, V., Vassylyeva, M.N., Tahirov, T.H., Yokoyama, S., Artsimovitch, I. and Vassylyev, D.G., 2004. Regulation through the secondary channel—structural framework for ppGpp-DksA synergism during transcription. *Cell*, 118(3), pp.297-309.

Petersen, S. and Young, G.M., 2002. Essential role for cyclic AMP and its receptor protein in *Yersinia enterocolitica* virulence. *Infection and immunity*, 70(7), pp.3665-3672.

Pigeon, R.P. and Silver, R.P., 1994. Topological and mutational analysis of KpsM, the hydrophobic component of the ABC-transporter involved in the export of polysialic acid in *Escherichia coli* K1. *Molecular microbiology*, 14(5), pp.871-881.

Potrykus, K. and Cashel, M., 2008. (p) ppGpp: still magical? *Annu. Rev. Microbiol.*, 62, pp.35-51.

Raghavan, R., Sage, A. and Ochman, H., 2011. Genome-wide identification of transcription start sites yields a novel thermosensing RNA and new cyclic AMP receptor protein-regulated genes in *Escherichia coli*. *Journal of bacteriology*, 193(11), pp.2871-2874.

Rhodius, V.A., West, D.M., Webster, C.L., Busby, S.J. and Savery, N.J., 1997. Transcription activation at class II CRP-dependent promoters: the role of different activating regions. *Nucleic acids research*, 25(2), pp.326-332.

Rickman, L., Scott, C., Hunt, D.M., Hutchinson, T., Menéndez, M.C., Whalan, R., Hinds, J., Colston, M.J., Green, J. and Buxton, R.S., 2005. A member of the cAMP receptor protein family of transcription regulators in *Mycobacterium tuberculosis* is required for virulence in mice and controls transcription of the rpfA gene coding for a resuscitation promoting factor. *Molecular microbiology*, 56(5), pp.1274-1286.

Rivetti, C., Guthold, M. and Bustamante, C., 1999. Wrapping of DNA around the *E. coli* RNA polymerase open promoter complex. *The EMBO journal*, 18(16), pp.4464-4475.

Roberts, I., 1996. The biochemistry and genetics of capsular polysaccharide production in bacteria. *Annual review of microbiology*, 50(1), pp.285-315.

Rodriguez-Siek, K.E., Giddings, C.W., Doetkott, C., Johnson, T.J. and Nolan, L.K., 2005. Characterizing the APEC pathotype. *Veterinary research*, 36(2), pp.241-256.

Rossiter, A.E., Browning, D.F., Leyton, D.L., Johnson, M.D., Godfrey, R.E., Wardius, C.A., Desvaux, M., Cunningham, A.F., Ruiz-Perez, F., Nataro, J.P. and Busby, S.J., 2011. Transcription of the plasmid-encoded toxin gene from Enteroaggregative *Escherichia coli* is regulated by a novel co-activation mechanism involving CRP and Fis. *Molecular microbiology*, 81(1), pp.179-191.

Rossiter, A.E., Godfrey, R.E., Connolly, J.A., Busby, S.J., Henderson, I.R. and Browning, D.F., 2015. Expression of different bacterial cytotoxins is controlled by two global transcription factors, CRP and Fis, that co-operate in a shared-recruitment mechanism. *Biochemical Journal*, 466(2), pp.323-335.

Russell, P.J. and Gordey, K., 2002. *IGenetics* (No. QH430 R87). San Francisco: Benjamin Cummings.

Russo, T.A., Liang, Y. and Cross, A.S., 1994. The presence of K54 capsular polysaccharide increases the pathogenicity of *Escherichia coli in vivo*. *Journal of Infectious Diseases*, 169(1), pp.112-118.

Santos-Zavaleta, A., Salgado, H., Gama-Castro, S., Sánchez-Pérez, M., Gómez-Romero, L., Ledezma-Tejeida, D., García-Sotelo, J.S., Alquicira-Hernández, K., Muñiz-Rascado, L.J., Peña-Loredo, P. and Ishida-Gutiérrez, C., 2019. RegulonDB v 10.5: tackling challenges to unify classic and high throughput knowledge of gene regulation in *E. coli* K-12. *Nucleic acids research*, 47(D1), pp. D212-D220.

Sauer, K., 2003. The genomics and proteomics of biofilm formation. *Genome biology*, *4*(6), pp.1-5.

Sarantuya, J., Nishi, J., Wakimoto, N., Erdene, S., Nataro, J.P., Sheikh, J., Iwashita, M., Manago, K., Tokuda, K., Yoshinaga, M. and Miyata, K., 2004. Typical enteroaggregative *Escherichia coli* is the most prevalent pathotype among *E. coli* strains causing diarrhea in Mongolian children. *Journal of Clinical Microbiology*, 42(1), pp.133-139.

Savery, N.J., Rhodius, V.A., Wing, H.J., and Busby, S.J., 1995. Transcription activation at *Escherichia coli* promoters dependent on the cyclic AMP receptor protein: Effects of binding sequences for the RNA polymerase α-subunit. *Biochemical Journal*, 309(1), pp.77-83.

Scavia, G., Staffolani, M., Fisichella, S., Striano, G., Colletta, S., Ferri, G., Escher, M., Minelli, F. and Caprioli, A., 2008. Enteroaggregative *Escherichia coli* associated with a foodborne outbreak of gastroenteritis. *Journal of medical microbiology*, 57(9), pp.1141-1146.

Semsey, S., Krishna, S., Sneppen, K. and Adhya, S., 2007. Signal integration in the galactose network of *Escherichia coli*. *Molecular microbiology*, 65(2), pp.465-476.

Sheikh, J., Czeczulin, J.R., Harrington, S., Hicks, S., Henderson, I.R., Le Bouguénec, C., Gounon, P., Phillips, A. and Nataro, J.P., 2002. A novel dispersin protein in enteroaggregative *Escherichia coli. The Journal of clinical investigation*, 110(9), pp.1329-1337.

Shimada, T., Fujita, N., Yamamoto, K. and Ishihama, A., 2011. Novel roles of cAMP receptor protein (CRP) in regulation of transport and metabolism of carbon sources. *PloS one*, 6(6), p. e20081.

Shin, M., Kang, S., Hyun, S.J., Fujita, N., Ishihama, A., Valentin-Hansen, P. and Choy, H.E., 2001. Repression of deoP2 in *Escherichia coli* by CytR: conversion of a transcription activator into a repressor. *The EMBO journal*, 20(19), pp.5392-5399.

Simon, R.U.P.A.P., Priefer, U. and Pühler, A., 1983. A broad host range mobilization system for *in vivo* genetic engineering: transposon mutagenesis in gram negative bacteria. *Bio/technology*, 1(9), pp.784-791.

Singh, S.S., Singh, N., Bonocora, R.P., Fitzgerald, D.M., Wade, J.T. and Grainger, D.C., 2014. Widespread suppression of intragenic transcription initiation by H-NS. *Genes & development*, 28(3), pp.214-219.

Smith, H.R., Cheasty, T. and Rowe, B., 1997. Enteroaggregative *Escherichia coli* and outbreaks of gastroenteritis in UK. *The Lancet*, 350(9080), pp.814-815.

Specian, R.D. and Oliver, M.G., 1991. Functional biology of intestinal goblet cells. *American Journal of Physiology-Cell Physiology*, *260*(2), pp.C183-C193.

Steenbergen, S. and Vimr, E., 2008. Biosynthesis of the *Escherichia coli* K1 group 2 polysialic acid capsule occurs within a protected cytoplasmic compartment. *Molecular Microbiology*, 68(5), pp.1252-1267.

Stevens, M.P., Hänfling, P., Jann, B., Jann, K. and Roberts, I.S., 1994. Regulation of *Escherichia coli* K5 capsular polysaccharide expression: evidence for involvement of RfaH in the expression of group II capsules. *FEMS microbiology letters*, 124(1), pp.93-98.

Sun, W., Roland, K.L., Kuang, X., Branger, C.G. and Curtiss III, R., 2010. *Yersinia pestis* with regulated delayed attenuation as a vaccine candidate to induce protective immunity against plague. *Infection and immunity*, 78(3), pp.1304-1313.

Terlizzi, M.E., Gribaudo, G. and Maffei, M.E., 2017. UroPathogenic *Escherichia coli* (UPEC) infections: virulence factors, bladder responses, antibiotic, and non-antibiotic antimicrobial strategies. *Frontiers in microbiology*, 8, p.1566.

Thompson, J.D., Gibson, T.J. and Higgins, D.G., 2003. Multiple sequence alignment using ClustalW and ClustalX. *Current protocols in bioinformatics*, (1), pp.2-3.

Vimr, E.R., Kalivoda, K.A., Deszo, E.L. and Steenbergen, S.M., 2004. Diversity of microbial sialic acid metabolism. *Microbiology and molecular biology reviews*, 68(1), pp.132-153.

Walker, K. and Osuna, R., 2002. Factors Affecting Start Site Selection at the *Escherichia coli fis* Promoter. *Journal of Bacteriology*, 184(17), pp.4783-4791.

Wallace-Gadsden, F., Johnson, J.R., Wain, J. and Okeke, I.N., 2007. Enteroaggregative *Escherichia coli* related to uropathogenic clonal group A. *Emerging infectious diseases*, 13(5), p.757.

Wallecha, A., Munster, V., Correnti, J., Chan, T. and Van der Woude, M., 2002. Dam-and OxyR-dependent phase variation of agn43: essential elements and evidence for a new role of DNA methylation. *Journal of bacteriology*, 184(12), pp.3338-3347.

Welch, R.A., Burland, V., Plunkett III, G., Redford, P., Roesch, P., Rasko, D., Buckles, E.L., Liou, S.R., Boutin, A., Hackett, J. and Stroud, D., 2002. Extensive mosaic structure revealed by the complete genome sequence of uropathogenic Escherichia coli. *Proceedings of the National Academy of Sciences*, *99*(26), pp.17020-17024.

Wells, T., Totsika, M. and Schembri, M., 2010. Autotransporters of *Escherichia coli*: a sequence-based characterization. *Microbiology*, 156(8), pp.2459-2469.

Wells, T., Tree, J., Ulett, G. and Schembri, M. (2007). Autotransporter proteins: novel targets at the bacterial cell surface. *FEMS Microbiology Letters*, 274(2), pp.163-172.

West, D., Williams, R., Rhodius, V., Bell, A., Sharma, N., Zou, C., Fujita, N., Ishihama, A. and Busby, S., 1993. Interactions between the *Escherichia coli* cyclic AMP receptor protein and RNA polymerase at class II promoters. *Molecular microbiology*, 10(4), pp.789-797.

Whitfield, C., 2006. Biosynthesis and assembly of capsular polysaccharides in *Escherichia coli*. *Annual review of biochemistry*, 75(1), pp.39-68.

Whitfield, C. and Roberts, I.S., 1999. Structure, assembly, and regulation of expression of capsules in *Escherichia coli*. *Molecular microbiology*, 31(5), pp.1307-1319.

Wiles, T.J., Kulesus, R.R. and Mulvey, M.A., 2008. Origins and virulence mechanisms of uropathogenic *Escherichia coli*. *Experimental and molecular pathology*, 85(1), pp.11-19.

Willis, L., Stupak, J., Richards, M., Lowary, T., Li, J. and Whitfield, C., 2013. Conserved glycolipid termini in capsular polysaccharides synthesized by ATP-binding cassette transporter-dependent pathways in Gram-negative pathogens. *Proceedings of the National Academy of Sciences*, 110(19), pp.7868-7873.

Willis, L.M. and Whitfield, C., 2013. Structure, biosynthesis, and function of bacterial capsular polysaccharides synthesized by ABC transporter-dependent pathways. *Carbohydrate research*, 378, pp.35-44.

Won, H.S., Lee, Y.S., Lee, S.H. and Lee, B.J., 2009. Structural overview on the allosteric activation of cyclic AMP receptor protein. *Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics*, 1794(9), pp.1299-1308.

Wooster, D.G., Maruvada, R., Blom, A.M. and Prasadarao, N.V., 2006. Logarithmic phase *Escherichia coli* K1 efficiently avoids serum killing by promoting C4bp-mediated C3b and C4b degradation. *Immunology*, 117(4), pp.482-493.

Xue, P., Corbett, D., Goldrick, M., Naylor, C. and Roberts, I.S., 2009. Regulation of expression of the region 3 promoter of the *Escherichia coli* K5 capsule gene cluster involves H-NS, SlyA, and a large 5′ untranslated region. *Journal of bacteriology*, 191(6), pp.1838-1846.

Xue, J., Tan, B., Yang, S., Luo, M., Xia, H., Zhang, X., Zhou, X., Yang, X., Yang, R., Li, Y. and Qiu, J., 2016. Influence of cAMP receptor protein (CRP) on bacterial virulence and transcriptional regulation of allS by CRP in *Klebsiella pneumoniae*. *Gene*, 593(1), pp.28-33.

Yamamoto, T. and Nakazawa, M., 1997. Detection and sequences of the enteroaggregative *Escherichia coli* heat-stable enterotoxin 1 gene in enterotoxigenic *E. coli* strains isolated from piglets and calves with diarrhea. *Journal of clinical microbiology*, 35(1), pp.223-227.

Yasir, M., Icke, C., Abdelwahab, R., Haycocks, J., Godfrey, R., Sazinas, P., Pallen, M., Henderson, I., Busby, S. and Browning, D., 2019. Organization and architecture of AggR-dependent promoters from enteroaggregative *Escherichia coli*. *Molecular Microbiology*, 111(2), pp.534-551.

Yen, M.R., Peabody, C.R., Partovi, S.M., Zhai, Y., Tseng, Y.H. and Saier Jr, M.H., 2002. Protein-translocating outer membrane porins of Gram-negative bacteria. *Biochimica et Biophysica Acta (BBA)-Biomembranes*, 1562(1-2), pp.6-31.

Yerushalmi, G., Nadler, C., Berdichevski, T. and Rosenshine, I., 2008. Mutational analysis of the locus of enterocyte effacement-encoded regulator (Ler) of enteropathogenic *Escherichia coli*. *Journal of bacteriology*, 190(23), pp.7808-7818.

Zhan, L., Han, Y., Yang, L., Geng, J., Li, Y., Gao, H., Guo, Z., Fan, W., Li, G., Zhang, L., Qin, C., Zhou, D. and Yang, R., 2008. The Cyclic AMP Receptor Protein, CRP, Is Required for Both Virulence and Expression of the Minimal CRP Regulon in *Yersinia pestis* Biovar microtus. *Infection and Immunity*, 76(11), pp.5028-5037.

Zheng, D., Constantinidou, C., Hobman, J.L. and Minchin, S.D., 2004. Identification of the CRP regulon using *in vitro* and *in vivo* transcriptional profiling. *Nucleic acids research*, 32(19), pp.5874-5893.

Zhang, P., Wu, G., Zhang, D. and Lai, W.F., 2023. Mechanisms and strategies to enhance penetration during intravesical drug therapy for bladder cancer. *Journal of Controlled Release*, 354, pp.69-79.

Zhou, Y.U., Pendergrast, P.S., Bell, A., Williams, R., Busby, S. and Ebright, R.H., 1994. The functional subunit of a dimeric transcription activator protein depends on promoter architecture. *The EMBO journal*, 13(19), pp.4549-4557.

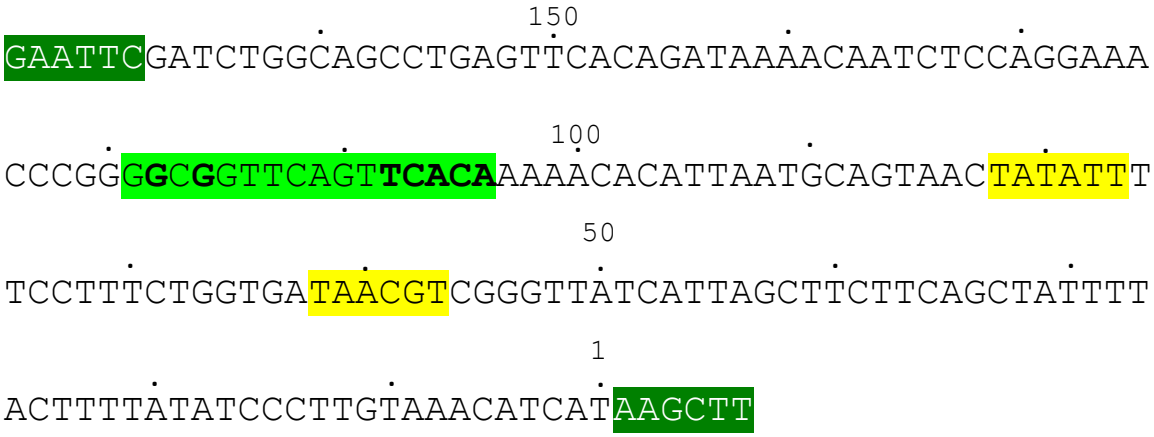
Zong, B., Liu, W., Zhang, Y., Wang, X., Chen, H. and Tan, C., 2016. Effect of kpsM on the virulence of porcine extraintestinal pathogenic *Escherichia coli*. *FEMS Microbiology Letters*, 363(21).

Appendix

Appendix A

A.

> *pic p042* WT



B.

> *pic p042* UKO

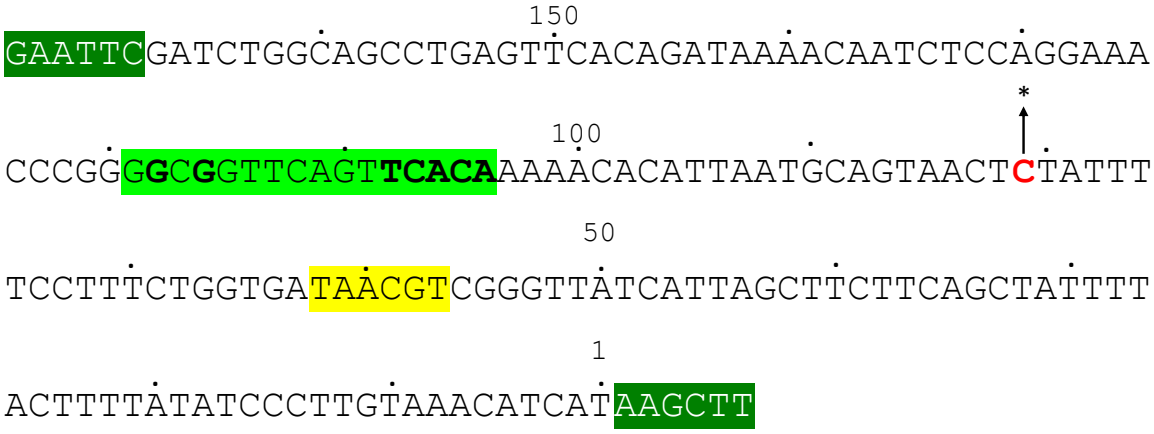
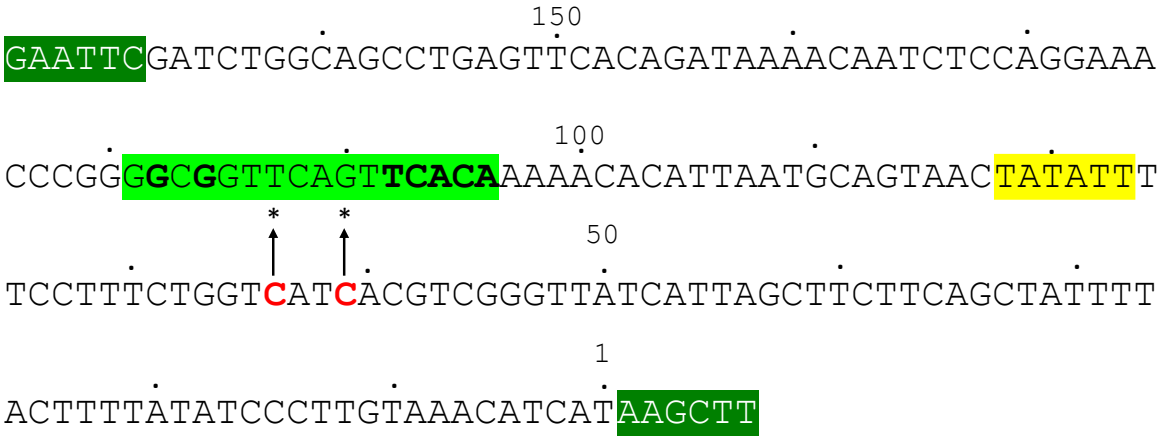


Figure 1: DNA base sequence of *pic p042* promoter fragment WT and derivatives (continued)

C.

> *pic p042* DKO



D.

> *pic p042* BKO

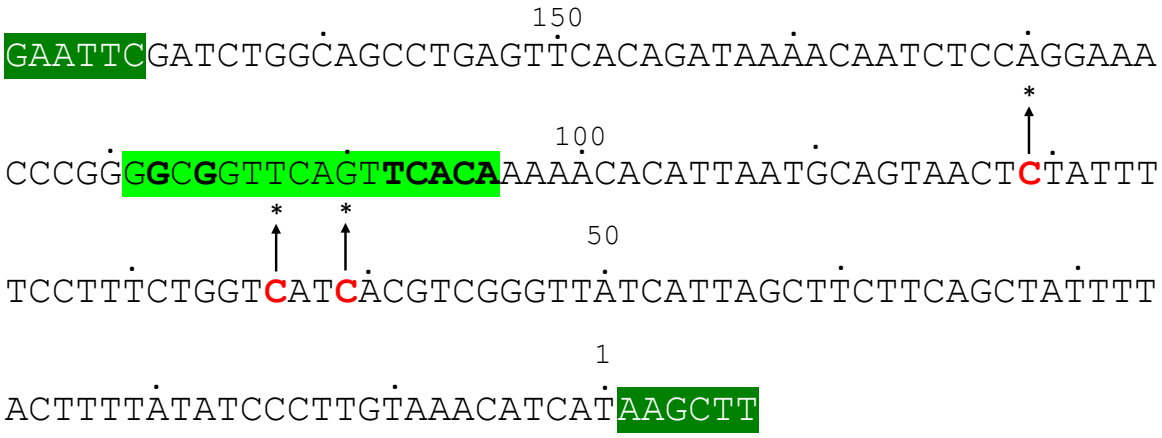


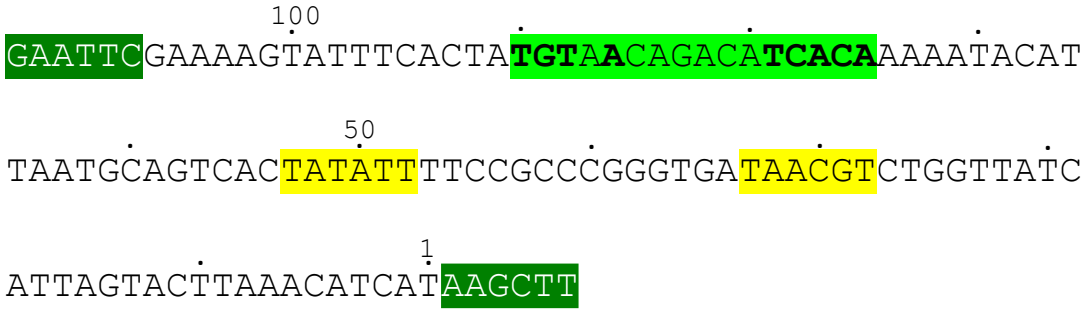
Figure 1: DNA base sequence of *pic p042* promoter fragment WT and derivatives

The panels show the DNA sequence of the *pic p042* promoter and its derivatives: (**A**) *pic p042* WT promoter fragment, (**B**) *pic p042* UKO promoter fragment, (**C**) *pic p042* DKO promoter fragment, and (**D**) *pic p042* BKO promoter fragment. 167 base pairs upstream of the *pic* gene from EAEC strain 042. The CRP binding site is shown in bold bases and is highlighted in light green. The arrows show the point of mutation in the upstream promoter (A ➔ C) at position 81 bp; and in the downstream promoter (A ➔ C) at position 61 bp; and the point of mutation

in the extended -10 element (G ➔ C) at position 64 bp. The sequence highlighted with dark green represents the restriction sites for EcoRI and HindIII. The numbering of the fragment starts from the HindIII site.

A.

>*pic p073* WT



B.

>*pic p073* UKO

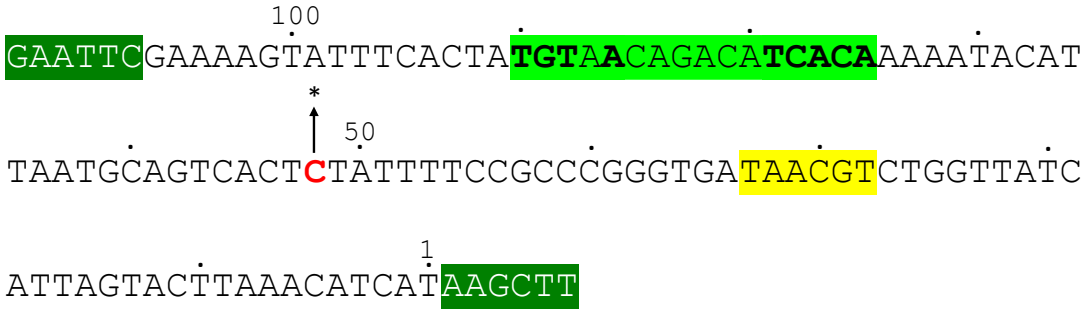
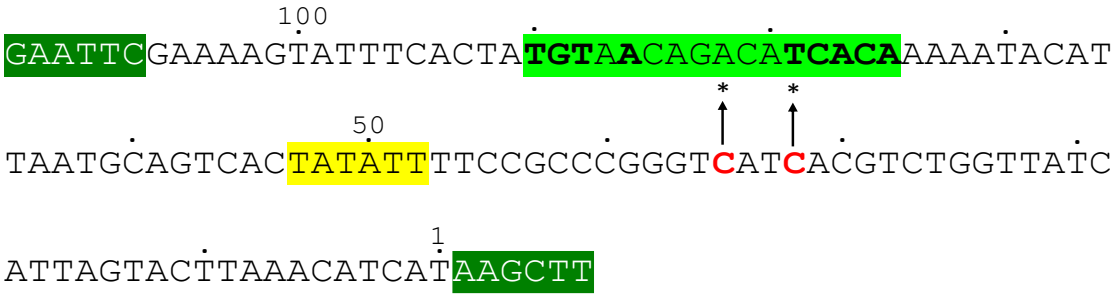


Figure 2: DNA base sequence of *pic p073* promoter fragment WT and derivatives (continued)

C.

>*pic p073* DKO



D.

>*pic p073* BKO

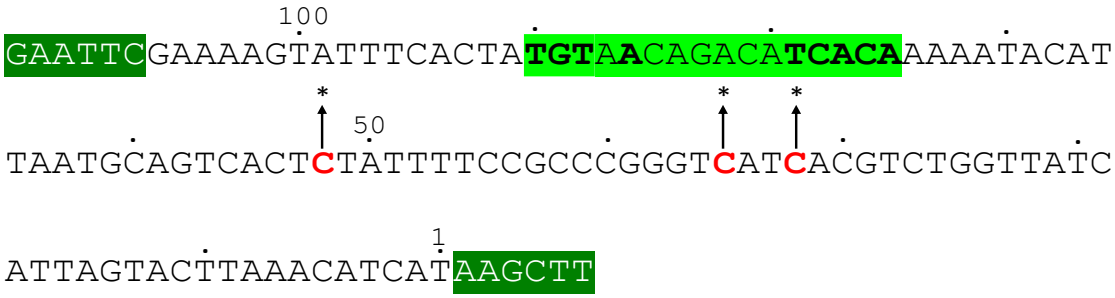


Figure 2: DNA base sequence of *pic p073* promoter fragment WT and derivatives

The panels show the base sequence of the *pic p073* promoter and its derivatives: (A) *pic p073* WT promoter fragment, (B) *pic p073* UKO promoter fragment (C) *pic p073* DKO promoter fragment, and (D) *pic p073* BKO promoter fragment. 107 base pairs upstream of the *pic* gene from UPEC strain CFT073. The CRP binding site is in bold bases and highlighted in light green. The arrows show the point of mutation in the upstream promoter (A ➔ C) at position 52 bp; and in the downstream promoter (A ➔ C) at position 32 bp; and the point mutation in the extended -10 element (G ➔ C) at position 35 bp. The sequence highlighted with dark green represents the restriction sites for EcoRI and HindIII. The numbering of the fragment starts from the HindIII site.

A.

>*kpsMII* (201 bp)

```
                                                  150
GAATTCTAATTACCTTCGGGATTATTGATGCGACTTAAATAACACCATTTAAATGTGATATA
                                        100
AATCACAAATATGACTGTAAAGAGGGGGCTGTAGATATAAATAAGAAGTACACGAGTGAATT
                          50                                  10
TTAAATAGGGAAATAGTTTCTCGGTGAACAATTTATTGGTAATCAATCGCGTGCGTTCTGGT

TTGAAACAGAAGCTT
```

B.

>*kpsMII* CRPKO (201 bp)

```
                                                  150
GAATTCTAATTACCTTCGGGATTATTGATGCGACTTAAATAACACCATTTAAATCTGATATA
                                        100
AATAACACAATATGACTGTAAAGAGGGGGCTGTAGATATAAATAAGAAGTACACGAGTGAATT
                          50                                  10
TTAAATAGGGAAATAGTTTCTCGGTGAACAATTTATTGGTAATCAATCGCGTGCGTTCTGGT

TTGAAACAGAAGCTT
```

C.

>*EC042_0414* (171 bp)

```
                150                                         
GAATTCATGTTGCAATCTTCTGCTGACAAAGCGTGCAACGTACTGGTGAAGAAAGTGCGTTA
  100                                                   50
TCTCAATGATGTGCGCAAGATCACAAAAATGATGAACGGGAAGCTAATTTATTCCTGGCTTG
                        10
  .         .         .
TATGGCCATGCAGTGAGTTTTTTCTCTTAATTATAAGTTAAAAGCTT
```

D.

>*EC042_3143* (191bp)

```
                150                                         
GAATTCATGTCCGTTTGCGGACAAGCAATAGATAAGGCGTGTTGTAGATCACAAATATTTAT
          100                                               
  .         .         .         .         .         .
ATGCAATAAATATCAATTATGTAACATGCATCACGATATGCGTATTGACATTTGTTGTTATA
          50                                            10
  .         .         .         .         .         .
TCTATAACTCAATGTTATATAAGAAATTAATAATTCACTGTTTTCAAAACACCGGTTTCCCA

AGCTT
```

Figure 3: DNA base sequence of EAEC 042 specific genes regulatory regions (continued)

E.

>*EC042_3644* (192 bp)

GAATTCTTTAACATTAATGCCAAAAACCGGGGAAACCCGGTTTTTTTATCCTCTTTTTGTAA
TAATCCGAAAAAATGTGAAGCGCCTCGCCGTTTCCACATCTCGCAGTTGCCACTTATCTTCA
TTTTTATGAAGATATAAATATTCATTTTATTGAAAATTTACTATGCGTCGAACGTTTATCAA

AAGCTT

F.

>*EC042_0224* (200 bp)

GAATTCATGAGCAAAATGAACAACAATGGCGGCAGTGTCGCGCCGAAAGAAAGAATCAGCGT
TCGGTATACACCAAAAGTTGATGGGGTGGCTGCTGATATCGAATTGCCGCTGAATTTATTGA
TCACCGGCAATCTGAAAGGGAAACCAGATAACACGCCGTTGGACGAGCGAACTGCCATTGCT

ATCAACCGAAGCTT

G.

>*EC042_4012* (192 bp)

GAATTCGGTGGAGATCTCTGTCACCAGCCAGACGGCGGGCACCAGTGCGGTCACTGCCAGCA
TTAACAGCAGCAGCCAGAGCCGGAACGTGACGTTTGTCGCGGATGTCAGGACGGCGAAGATA
GCGGATCTGGTCGTTAGCCAGGATAACGCGGTGGCAGACGGTTCGACGGCGAACACGCTGCG

AAGCTT

H.

>*EC042_3975* (192 bp)

GAATTCAGATTCGGTTTTTCAGACCCCCATCCTAAACCAGCATGAATCAGACTAAAAGTTCA
CTTACCCATACTGGCATTAACAATGTTATCGTATACACGTTATTGCGTCTTTATAGGTGTTA
ACATGGCTAAAGATTTAACAACCTCTGCTCATGACAGACAAAATATTTTGAATAACTCTTAT

AAGCTT

Figure 3: DNA base sequence of EAEC 042 specific genes regulatory regions (continued)

I.

>*EC042_3970* (186 bp)

GAATTCCAATAGCTAATTTTTTTGGCGCACCATATGAGGTAGGTTATCTATTCTCTGTTTTT
TGTCTTGTTATTGCCCTTCTTTACCTGTGGATGAGAGCAAAGAACAAAAATTAGCCATCCGT
GGCTAACCAGTTATTGAGCTTCTTTTTGTGATGCCTGTACGACTCTGGTGAGCGATAAGCTT

J.

>*EC042_4604* (192 bp)

GAATTCAACTATTGAGGCCAGCCTGATTTTGGCGCAAAATGGGCTGAACGCCGCCGAAAACT
GTTGCTGATTGCACCCGCAACGCTACGCAAGCAGTGGAGCCAGGAGCTGGAAGAGAAATTTT
CGCTGCCTTCACAAATTATTGAAGCGAAGATCTTCAATCAAATGGTTAAAGAGGGGAGCGAT

AAGCTT

K.

>*virK* (192 bp)

GAATTCTCATGTTTTCCGGCAATTGAGATACTGGTATCAGAAACGCAAAACCTTTGTTGCGG
TCTATAGTGATTTCTGGGAGTCTGTAGCAGGAAAAACCTGTGGTGACTGGTACAGATTACCA
ACCCAGGTGATTCGAAAGCCGCTAAGCGATATAGCCAGTAAAAAACGCTCTGAGTACCGAAA

AAGCTT

L.

>*EC042_0225* (145 bp)

GAATTCAACTCGAATAAAGAAAAGGGTGTGGTTCCTGATGTGAGCCGCATCACACTCTAACC
CGTCGGGGATTAACGACGGGTTGTTCGTAAAACAGCAGTTGATAATTTCACAAGGAGTTCAT
AAATGCCAACCCCATAAGCTT

Figure 3: DNA base sequence of EAEC 042 specific genes regulatory regions

The panel shows the DNA sequence of **A.** *kpsMII,* **B.** *kpsMII CRPKO,* **C.** *EC042_0414,* **D.** *EC042_3143,* **E.** *EC042_3644,* **F.** *EC042_0224,* **G.** *EC042_4012* **H.** *EC042_3975,* **I.** *EC042_3970,* **J.** *EC042_4604,* **K.** *virK***;** and **L.** *EC042_0225*. Black bases indicate promoter fragments whereas the red bases indicate the intragenic regions. CRP binding sites identified by CRP ChIP-seq are shown as yellow highlighted bases. The sequences highlighted with dark

green represent the restriction sites for EcoRI and HindIII. The numbering of each fragment starts from the downstream HindIII site.

A.

>*pkpsM*216 (216bp)

```
                200
GAATTCTAACACTGGTTAAAATAAATACACAAAATAATCAAGATGTATATTTTAATCGACGA
150                                                 100
AAATAATTACCTTCGGGATTATTGATGCGACTTAAATAACACCATTTAAATGTGATATAAAT
                                                50
CACAAATATGACTGTAAAGAGGGGGCTGTAGATATAAATAAGAAGTACACGAGTGAATTTTA
                    10
AATAGGGAAATAGTTTCTCGGTGAACAAAGCTT
```

B.

> *pkpsM*216R (216bp)

```
                200
GAATTCTGTTCACCGAGAAACTATTTCCCTATTTAAAATTCACTCGTGTACTTCTTATTTAT
150                                                 100
ATCTACAGCCCCCTCTTTACAGTCATATTTGTGATTTATATCACATTTAAATGGTGTTATTT
                                                50
AAGTCGCATCAATAATCCCGAAGGTAATTATTTTCGTCGATTAAAATATACATCTTGATTAT
                    10
TTTGTGTATTTATTTTAACCAGTGTTAAAGCTT
```

Figure 4: DNA sequence of the *pkpsMII*216 forward and reverse fragments

The figure illustrates the DNA sequences of 216 bp of *pkpsMII* DNA around the CRP binding site that was located by CRP ChIP-seq, which is shown in yellow with bold bases. A. *pkpsM*216 oriented to the *kpsM*II gene. B. *pkpsM*216R complement sequence of A. The sequences highlighted with dark green represents the restriction sites EcoRI and HindIII. The numbering of the fragment starts from the HindIII site.

A.

>*pkpsM* WT (851 bp)

```
                                                          800
GAATTCTTATTAATAGTTGCAATAAATCATTGAGTAACAATTGATAGGCCAAAACATAT
                                                    750
AGGATAATTCTTGTGTGATCTGTATTTTGTGTAGCTTGGAAATTAGTAAAATTCCTGGA
                                          700
GATAATCAGAAAGGGAATTTCAAATAAGCATAATAATTTCCAGTGAAACTATTCGTTGA
                                650
TTTTAAGAAATGTCTGCTCAGGTATATCTACTGAGGGATGGTGTTGGTTGTAACACTGG
                          600
TTAAAATAAATACACAAAATAATCAAGATGTATATTTTAATCGACGAAAATAATTACCT
550                                                       500
TCGGGATTATTGATGCGACTTAAATAACACCATTTAAATGTGATATAAATCACAAATAT
                                                450
GACTGTAAAGAGGGGGCTGTAGATATAAATAAGAAGTACACGAGTGAATTTTAAATAGG
                                      400
GAAATAGTTTCTCGGTGAACAATTTATTGGTAATCAATCGCGTGCGTTCTGGTTTGAAA
                                350
CAGATTGCAGGTATTGTTACGCATAAAATCCATGGGGGTATTATAATCAAGTAGTTAAC
                          300
AGTAATAACAAAAAATAATTTCCTGGGAAATTAACTCTACATTCTAAAACGGCCAGTTT
      250                                                 200
TCTGAAATTACCAGAAGACCGAATACATTCATGCCTGAAGAATGGATTAGAAGAAACGA
                                                150
GACGGAAATAGATCTATTTATCCCTGCGGAAATAATTTCTGCTGAAATTTTTTGCGGCT
                                      100
ATTAAAAAGGTCAAACCGTCTGAGTAAATTTTATCCAGTTACAAATAAGCATTACCTCC
                                50
AGTGTATTGGTAGCTGTTAAGCCAAGGGCGGTAGCGTACCTGAAGAGATTAGGATCACA

TCATCAAATGGCAAGAAGTAAGCTT
```

Figure 5: DNA sequences of the *pkpsMII* nested promoter deletions (continued)

B.

>*pkpsM* 751 (751 bp)

Figure 5: DNA sequences of the *pkpsMII* nested promoter deletions (continued)

C.

>*pkpsM* 651 (651 bp)

```
                                                          600
GAATTCACTGAGGGATGGTGTTGGTTGTAACACTGGTTAAAATAAATACACAAAATAAT
                                                  550
CAAGATGTATATTTTAATCGACGAAAATAATTACCTTCGGGATTATTGATGCGACTTAA
                                          500
ATAACACCATTTAAATGTGATATAAATCACAAATATGACTGTAAAGAGGGGGCTGTAGA
                                  450
TATAAATAAGAAGTACACGAGTGAATTTTAAATAGGGAAATAGTTTCTCGGTGAACAAT
                          400
TTATTGGTAATCAATCGCGTGCGTTCTGGTTTGAAACAGATTGCAGGTATTGTTACGCA
350                                                       300
TAAAATCCATGGGGGTATTATAATCAAGTAGTTAACAGTAATAACAAAAAATAATTTCC
                                                  250
TGGGAAATTAACTCTACATTCTAAAACGGCCAGTTTTCTGAAATTACCAGAAGACCGAA
                                          200
TACATTCATGCCTGAAGAATGGATTAGAAGAAACGAGACGGAAATAGATCTATTTATCC
                                  150
CTGCGGAAATAATTTCTGCTGAAATTTTTTGCGGCTATTAAAAAGGTCAAACCGTCTGA
                          100
GTAAATTTTATCCAGTTACAAATAAGCATTACCTCCAGTGTATTGGTAGCTGTTAAGCC
                  50
AAGGGCGGTAGCGTACCTGAAGAGATTAGGATCACATCATCAAATGGCAAGAAGTAAGC

TT
```

Figure 5: DNA sequences of the *pkpsMII* nested promoter deletions (continued)

D.

>*pkpsM* 551 (551 bp)

```
                                                  500
GAATTCGATGCGACTTAAATAACACCATTTAAATGTGATATAAATCACAAATATGACTG
                                          450
TAAAGAGGGGGCTGTAGATATAAATAAGAAGTACACGAGTGAATTTTAAATAGGGAAAT
                                  400
AGTTTCTCGGTGAACAATTTATTGGTAATCAATCGCGTGCGTTCTGGTTTGAAACAGAT
                          350
TGCAGGTATTGTTACGCATAAAATCCATGGGGGTATTATAATCAAGTAGTTAACAGTAA
                  300
TAACAAAAAATAATTTCCTGGGAAATTAACTCTACATTCTAAAACGGCCAGTTTTCTGA
250                                                       200
AATTACCAGAAGACCGAATACATTCATGCCTGAAGAATGGATTAGAAGAAACGAGACGG
                                                  150
AAATAGATCTATTTATCCCTGCGGAAATAATTTCTGCTGAAATTTTTTGCGGCTATTAA
                                          100
AAAGGTCAAACCGTCTGAGTAAATTTTATCCAGTTACAAATAAGCATTACCTCCAGTGT
                                  50
ATTGGTAGCTGTTAAGCCAAGGGCGGTAGCGTACCTGAAGAGATTAGGATCACATCATC

AAATGGCAAGAAGTAAGCTT
```

E.

>*pkpsM* 451 (451 bp)

```
                                                  400
GAATTCAAATAGGGAAATAGTTTCTCGGTGAACAATTTATTGGTAATCAATCGCGTGCG
                                          350
TTCTGGTTTGAAACAGATTGCAGGTATTGTTACGCATAAAATCCATGGGGGTATTATAA
                                  300
TCAAGTAGTTAACAGTAATAACAAAAAATAATTTCCTGGGAAATTAACTCTACATTCTA
                          250
AAACGGCCAGTTTTCTGAAATTACCAGAAGACCGAATACATTCATGCCTGAAGAATGGA
                  200
TTAGAAGAAACGAGACGGAAATAGATCTATTTATCCCTGCGGAAATAATTTCTGCTGAA
150                                                       100
ATTTTTTGCGGCTATTAAAAAGGTCAAACCGTCTGAGTAAATTTTATCCAGTTACAAAT
                                                  50
AAGCATTACCTCCAGTGTATTGGTAGCTGTTAAGCCAAGGGCGGTAGCGTACCTGAAGA

GATTAGGATCACATCATCAAATGGCAAGAAGTAAGCTT
```

Figure 5: DNA sequences of the *pkpsMII* nested promoter deletions (continued)

F.

>*pkpsM* 351 (351 bp)

```
                                                300
GAATTCGGGGTATTATAATCAAGTAGTTAACAGTAATAACAAAAAATAATTTCCTGGGA
                                    250
AATTAACTCTACATTCTAAAACGGCCAGTTTTCTGAAATTACCAGAAGACCGAATACAT
                        200
TCATGCCTGAAGAATGGATTAGAAGAAACGAGACGGAAATAGATCTATTTATCCCTGCG
                150
GAAATAATTTCTGCTGAAATTTTTTGCGGCTATTAAAAAGGTCAAACCGTCTGAGTAAA
          100
TTTTATCCAGTTACAAATAAGCATTACCTCCAGTGTATTGGTAGCTGTTAAGCCAAGGG
50
CGGTAGCGTACCTGAAGAGATTAGGATCACATCATCAAATGGCAAGAAGTAAGCTT
```

Figure 5: DNA sequences of the *pkpsMII* nested promoter deletions

The panel shows the DNA sequences of the (**A**) 851 bp *kpsMII* promoter fragment, (**B**) 751 bp *kpsMII* promoter fragment, (**C**) 651 bp *kpsMII* promoter fragment, (**D**) 551 bp *kpsMII* promoter fragment, (**E**) 451 bp *kpsMII* promoter fragment, (**F**) 351 bp *kpsMII* promoter fragment. The CRP binding site indicated by CRP ChIP-seq is highlighted in yellow highlighted and bold bases. The DNA bases highlighted in grey represent the -10 element of the previously identified promoter (P1). The transcription starting site +1 of the P1 promoter is indicated with a black arrow at 750 bp. The *kpsMII* gene is shown in red bases. The sequence highlighted in dark green represents the restriction sites for EcoRI and HindIII. The numbering of the fragment starts from the HindIII site.

A.

>*pkpsM 440* (440 bp)

```
                                        400
GAATTCTTATTAATAGTTGCAATAAATCATTGAGTAACAATTGATAGGCCAAAACATAT
                                  350
AGGATAATTCTTGTGTGATCTGTATTTTGTGTAGCTTGGAAATTAGTAAAATTCCTGGA
                    300
GATAATCAGAAAGGGAATTTCAAATAAGCATAATAATTTCCAGTGAAACTATTCGTTGA
          250                                               200
TTTTAAGAAATGTCTGCTCAGGTATATCTACTGAGGGATGGTGTTGGTTGTAACACTGG
                                              150
TTAAAATAAATACACAAAATAATCAAGATGTATATTTTAATCGACGAAAATAATTACCT
                                    100
TCGGGATTATTGATGCGACTTAAATAACACCATTTAAATGTGATATAAATCACAAATAT
                        50
GACTGTAAAGAGGGGGCTGTAGATATAAATAAGAAGTACACGAGTGAATTTTAAATAGG

GAAATAGTTTCTCGGTGAACAAAGCTT
```

B.

>*pkpsM* p1KO (440 bp)

```
                                        400
GAATTCTTATTAATAGTTGCAATAAATCATTGAGTAACAATTGATAGGCCAAAACATAT
                                  *
                                  ↑ 350
AGGATAATTCTTGTGTGATCTGTGTTTTGTGTAGCTTGGAAATTAGTAAAATTCCTGGA
                    300
GATAATCAGAAAGGGAATTTCAAATAAGCATAATAATTTCCAGTGAAACTATTCGTTGA
          250                                               200
TTTTAAGAAATGTCTGCTCAGGTATATCTACTGAGGGATGGTGTTGGTTGTAACACTGG
                                              150
TTAAAATAAATACACAAAATAATCAAGATGTATATTTTAATCGACGAAAATAATTACCT
                                    100
TCGGGATTATTGATGCGACTTAAATAACACCATTTAAATGTGATATAAATCACAAATAT
                        50
GACTGTAAAGAGGGGGCTGTAGATATAAATAAGAAGTACACGAGTGAATTTTAAATAGG

GAAATAGTTTCTCGGTGAACAAAGCTT
```

Figure 6: DNA sequence of the *pkpsM* promoter fragment WT and with point mutations (continued)

C.

>*pkpsM* p2KO (440 bp)

```
                                   400
GAATTCTTATTAATAGTTGCAATAAATCATTGAGTAACAATTGATAGGCCAAAACATAT
                              350
AGGATAATTCTTGTGTGATCTGTATTTTGTGTAGCTTGGAAATTAGTAAAATTCCTGGA
                         300
GATAATCAGAAAGGGAATTTCAAATAAGCATAATAATTTCCAGTGAAACTATTCGTTGA
       250                                                 200
TTTTAAGAAATGTCTGCTCAGGTATATCTACTGAGGGATGGTGTTGGTTGTAACACTGG
                                                  150
TTAAAATAAATACACAAAATAATCAAGATGTATATTTTAATCGACGAAAATAATTACCT
                                   100    *
TCGGGATTATTGATGCGACTTAAATAACACCATTTAAATGTGATGTAAATCACAAATAT
                              50
GACTGTAAAGAGGGGGCTGTAGATATAAATAAGAAGTACACGAGTGAATTTTAAATAGG

GAAATAGTTTCTCGGTGAACAAAGCTT
```

D.

>*pkpsM* p3KO (440 bp)

```
                                   400
GAATTCTTATTAATAGTTGCAATAAATCATTGAGTAACAATTGATAGGCCAAAACATAT
                              *350
AGGATAATTCTTGTGTGATCTGTGTTTTGTGTAGCTTGGAAATTAGTAAAATTCCTGGA
                         300
GATAATCAGAAAGGGAATTTCAAATAAGCATAATAATTTCCAGTGAAACTATTCGTTGA
       250                                                 200
TTTTAAGAAATGTCTGCTCAGGTATATCTACTGAGGGATGGTGTTGGTTGTAACACTGG
                                                  150
TTAAAATAAATACACAAAATAATCAAGATGTATATTTTAATCGACGAAAATAATTACCT
                                   100    *
TCGGGATTATTGATGCGACTTAAATAACACCATTTAAATGTGATGTAAATCACAAATAT
                              50
GACTGTAAAGAGGGGGCTGTAGATATAAATAAGAAGTACACGAGTGAATTTTAAATAGG

GAAATAGTTTCTCGGTGAACAAAGCTT
```

Figure 6: DNA sequence of the *pkpsM* promoter fragment WT and the promoter with point mutations

A. Figure shows the DNA sequence, 440 bp, of the *pkpsM* promoter fragment (wild type), with the CRP binding site highlighted in yellow highlighted bases. The DNA bases highlighted in grey represent the -10 element of the previously identified promoters,

P1, and promoter P2. The sequences highlighted in dark green represents the restriction sites for EcoRI and HindIII. The numbering of the fragment starts from the HindIII site.

B. Figure shows the DNA sequence, 440 bp, of the derivative *pkpsM* promoter fragment (p1KO), with the CRP binding site highlighted in yellow. The DNA bases highlighted in grey represent the -10 element (P2). The arrow shows the point of mutation in P1 (A ➔ G) at position 352 bp. The sequencse highlighted in dark green represent the restriction sites for EcoRI and HindIII. The numbering of the fragment starts from the HindIII site.

C. Figure shows the DNA sequence, 440 bp, of the derivative *pkpsM* promoter fragment (p2KO), with the CRP binding site highlighted in yellow. The DNA bases highlighted in grey represent the -10 element of the previously identified promoter (P1). The arrow shows the point of mutation in P2 (A ➔ G) at position 95 bp. The sequences highlighted in dark green represent the restriction sites for EcoRI and HindIII. The numbering of the fragment starts from the HindIII site.

D. Figure shows the DNA sequence, 440 bp, of the derivative *pkpsM* promoter fragment (p3KO), with the CRP binding site highlighted in yellow. The arrow shows the point of mutations in P1 (A ➔ G) at position 352 bp and P2 (A ➔ G) at position 95 bp. The sequences highlighted in dark green represent the restriction sites EcoRI and HindIII. The numbering of the fragment starts from the HindIII site.

Appendix B

CRP targets identified by ChIP-seq analysis on the EAEC 042 chromosome

Peak Centre[1]	Score[2]	Annotated Gene[3]	*E. coli* K-12 Homologues[4]	Distance to TSS[5]	Matching Sequence[6]	*p*-value[7]	Regulated by CRP[8]	Gene Product[9]
4801045	1563	*proP*	*proP*	-203	**TGTGA**AGTTGA**TCACA**ATTT	1.26E-08	Y	Proline (betaine transporter)
4149109	17137	*mtlA*	*mtlA*	-260	**TGTGA**GTGATG**TCACA**TTTTT	1.26E-08	Y	Mannitol-specific PTS system EIICBA component
1748762	63	*ydeA*	*ydeA*	-75	**TGTGA**TCTGGA**TCGCG**TTTTC	2.31E-08	N	L-arabinose MFS transporter
1827122	631	*mlc*	*mlc*	-87	**TGTGA**TTAACA**GCACA**TTTTT	3.09E-08	Y	Protein Mlc (making large colonies protein)
3364989	113	*galP*	*galP*	-58	**TGTGA**TTTGCT**TCACA**TCTTT	4.09E-08	Y	Galactose-proton symporter (galactose transporter)
2034392	2844	*gapA*	*gapA*	-118	**TGTGA**TTTTCA**TCACG**ATTTA	4.09E-08	Y	Glyceraldehyde 3-phosphate dehydrogenase A
2205763	485	*EC042_2120*	*yedR*	-38	**TGTGA**TAAAGG**TCACA**TTTTT	5.38E-08	N	Putative membrane protein
1261443	508	*EC042_1182*	*bhsA*	-63	**TGTGA**TCCAGA**TCACA**TCTAT	7.03E-08	Y	Putative exported protein
147627	240	*gcd*	*gcd*	-56	**TGTGA**TCGTCA**TCACA**ATTCG	7.03E-08	Y	Quinoprotien glucose dehydrogenase
3231340	646	*EC042_3028*	In K-12 *(EO53_0363)*	-118	**TGTGA**CCTGGG**TCACG**AATTA	9.13E-08	-	Hypothetical protein
2496355	30	*EC042_2373*	*yohj*	-44	**TGTGA**TCGGTA**GCACG**TTTTA	1.51E-07	N	Putative membrane protein
4753197	338	*EC042_4429*	*yjcB*	-26	**TGTGA**TATAGT**TCACA**AAATT	1.93E-07	N	Putative membrane protein
3284051	40	*EC042_3081*	*ygeW*	-144	**TGTGG**GGTTGA**TCACA**AATTG	1.93E-07	N	Putative aspartate/ornithine carbamoyltransferase
461029	312	*EC042_0414*	S.G.	65	**TGTGA**TCTTGC**GCACA**TCATT	2.45E-07	-	Conserved hypothetical protein
263739	111	*EC042_0225*	S.G.	-67	**TGTGA**GCCGCA**TCACA**CTCTA	2.45E-07	-	Putative type VI secretion system protein
3459052	894	*kpsMII*	S.G.	-479	**TGTGA**TTTATA**TCACA**TTTAA	3.10E-07	-	Polysialic acid transport permease protein
608383	168	*EC042_0536*	S.G.	-112	**TGTGA**TGTAAA**GCGCA**TTTTT	3.10E-07	-	Putative adhesin (not virulence)
3747290	727	*EC042_3523*	*yhcN*	-41	**CGTGA**CCCATA**TCACA**AAATC	3.90E-07	N	Conserved hypothetical protein

3931572	607	*gntK*	*gntK*	-108	**TGTGA**GCTACT**TCAAA**TTTGT	4.89E-07	Y	Thermoresistant gluconokinase
3676811	377	*deaD*	*deaD*	-9	**TTTGA**GCCGGT**TCACA**CTTTT	4.89E-07	N	ATP-dependent RNA helicase (DEAD-box protein)
3160614	242	*EC042_2967*	*ygcW*	-44	**TGTGA**TTACGA**TCACA**TTAGC	4.89E-07	N	Putative short chain dehydrogenase
2051133	39	*EC042_1961*	*yeaQ*	-74	**CGTGA**TGCTCG**TCACA**AATCT	4.89E-07	N	Putative transglycosylase associated protein
1711834	361	*EC042_1630*	*aslA*	-274	**AGTGA**CATTCA**TCACA**TATTT	6.10E-07	N	Putative sulfatase
3751263	64	*aaeR*	*aaeR*	-18	**AGTGA**TTTAGA**TCACA**TAATA	7.57E-07	Y	LysR-family transcriptional regulator
2197691	115	*EC042_2113*	*yedP*	-102	**GGTGA**CGCGCG**TCACA**TTTCT	7.57E-07	N	Putative mannosyl-3-phosphoglycerate phosphatase
4062646	308	*dctA*	*dctA*	-119	**TGCGA**GCCAGC**TCAAA**CTTTT	9.37E-07	Y	C4-dicarboxylate transport protein
2844005	245	*hyfA*	*hyfA*	-177	**CGTGA**TCAAGA**TCACA**TTCTC	9.37E-07	Y	Hydrogenase-4 component A
5218617	4089	*deoC*	*deoC*	-130	**TGCGA**TCTGGT**TCAAA**TAATT	1.16E-06	Y	Deoxyribose-phosphate aldolase
4388273	416	*atpI*	*atpI*	-244	**CGTGC**TTCAGA**TCACA**TATTG	1.16E-06	N	ATP synthase protein I
3417179	116	*mchS4*	In K-12	1	**TGTGA**TTATTA**TCACA**ATTCA	1.16E-06	-	Conserved hypothetical protein
2877210	1487	*xseA*	*xseA*	1	**TGTGA**GCGAGA**TCAAA**TTCTA	1.16E-06	Y	Exodeoxyribonuclease VII large subunit
2506198	500	*mglB*	*mglB*	-246	**TGTGA**GTGATT**TCACA**GTATC	1.16E-06	Y	Galactose/glucose ABC transporter substrate-binding protein MglB
3858617	17	*EC042_3644*	In K-12	-75	**GGCGA**GGCGCT**TCACA**TTTTT	1.42E-06	-	Hypothetical protein
816974	62	*sdhC*	*sdhC*	-212	**TGTGA**TTTACA**TCACA**TAAAG	1.42E-06	Y	Succinate dehydrogenase cytochrome b-556 subunit
1881643	144	*EC042_1802 A*	In K-12 *(EO53_03635)*	39	**CGTGA**TCTTGA**TCACG**AATGA	1.42E-06	-	Conserved hypothetical protein
56171	77	*folA*	*folA*	-165	**TGTGA**TTTACG**TCACT**CTTTA	1.42E-06	N	Dihydrofolate reductase
3883094	9	*greB*	*greB*	90	**TGCGG**CCCAGG**TCACC**TTTTT	1.74E-06	N	Transcription elongation factor GreB
2993024	8657	*raiA*	*raiA*	-57	**TGAGA**TTTCCA**TCACA**CATTT	2.12E-06	Y	Ribosome-associated inhibitor A
2784492	7180	*ptsH*	*ptsH*	-318	**TGTGG**CCTGCT**TCAAA**CTTTC	2.12E-06	Y	Phosphocarrier protein Hpr
5020611	392	*cycA*	*cycA*	-179	**TGTGA**GCTGTT**TCGCG**TTATC	2.57E-06	N	D-serine, D-alanine, glycine transporter
3590224	20	*air*	*air*	-85	**TTTGA**TTTAGA**TCGCA**ATTTG	2.57E-06	Y	Aerotaxis receptor protein
3308819	37	*EC042_3098*	*ygfT*	-96	**TGCGA**CTGAGT**TCAAA**TTATT	2.57E-06	N	Formate-dependent uric acid utilization protein YgfT
3128509	52	*nlpD*	*nlpD*	678	**TGTGA**CCGTGG**TCGCA**GTTGG	2.57E-06	Y	Murein hydrolase activator NlpD

2525169	7	*EC042_2399*	*psuK*	-206	**TGTGC**ATCCCG**TCACA**ATTC	2.57E-06	N	Putative pseudouridine kinase
2499767	56	*EC042_2378*	*preT*	-87	**TGTGA**ATCCTT**TCACA**GTTTA	2.57E-06	Y	Putative oxidoreductase
2349836	25	*dacD/sbcB*	*dacD/sbcB*	- 74/ - 59	**TGTGA**CTACTA**TCTCA**TTTTT	2.57E-06	N	Penicillin-binding protein 6B/ exodeoxyribonuclease I
1381630	819	*adhE*	*adhE*	-217	**TTTGA**TTTGGA**TCACG**TAATC	2.57E-06	N	Aldehyde-alcohol dehydrogenase
4290090	31	*EC042_4032*	*yidG*	267	**CGCGA**GGGAGA**TCAAA**AATTT	3.12E-06	N	Putative membrane protein
1740383	90	*EC042_1653*	In K-12	-31	**CGTGA**TTACGA**TCACA**TTCTC	3.12E-06	-	Putative membrane protein
1344062	218	*EC042_1262*	*ychH*	-81	**TGCGG**GCGTGA**TCACA**ATTAC	3.12E-06	Y	Putative membrane protein
2627362	785	*glpT*	*glpT*	-110	**TGTGA**ATTACC**GCACA**CATTA	3.77E-06	Y	Glycerol-3-phosphate transporter
5094768	3314	*EC042_4748*	In K-12	-54	**TGTGA**TGATTG**TCGCA**TTTGA	4.53E-06	-	PTS system EIIA component
3638072	758	*tdcA*	*tdcA*	-63	**TGTGA**GTGGTC**GCACA**TATCC	4.53E-06	Y	Tdc operon transcriptional activator
3356318	6783	*EC042_3143*	S.G.	31	**CGTGT**TGTAGA**TCACA**AATAT	4.53E-06	-	Hypothetical protein
4596821	13	*cytR*	*cytR*	-123	**TGCGA**GGCGGA**TCGAA**AAATT	5.44E-06	Y	DNA-binding transcriptional regulator
2544035	94	*EC042_2414*	*yejG*	-134	**TGCGG**GCGTGG**TCACG**TTTGC	5.44E-06	N	YejG family protein
1336252	1427	*dhaR*	*dhaR*	-14	**GGTGT**TCCGGC**TCGCA**ATTTT	5.44E-06	N	PTS-dependent dihydroxyacetone kinase operon regulator (sigma-54 dependent transcriptional regulator)
4272121	295	*EC042_4013*	*nepI*	-7	**TGTGA**CGCATT**TAACG**TTTTT	6.52E-06	N	Purine ribonucleoside efflux pump NepI
3838726	12183	***ppiA***	*ppiA*	-121	**AGTGA**TTCTCA**TCACG**AAATA	6.52E-06	Y	Peptidyl-prolyl cis-trans isomerase
4456964	77	*cyaA*	*cyaA*	-148	**TGTTA**AATTGA**TCACG**TTTTA	9.27E-06	Y	Adenylate cyclase
2766013	32	*glk/ EC042_2608*	*glk/ yfeO*	- 92/ - 90	**TGTGA**CCCAGA**TCGAT**ATTTA	9.27E-06	N	Glucokinase/Ion channel protein
1865552	30	*uidA*	*uidA*	-162	**CGTGA**TATAGA**TCGCA**TTAAT	9.27E-06	Y	Glucuronidase ß
1314710	502	*dadA*	*dadA*	-92	**GGTGA**GCTGGC**TCACA**TCTCC	9.27E-06	Y	D-amino acid dehydrogenase small subunit
544222	102	*maa*	*maa*	-73	**TGTGA**TCTTTA**TCACA**CAGAT	9.27E-06	N	Maltose O-acetyltransferase
467174	392	*aroM*	*aroM*	-195	**AGGGA**TCTGCG**TCACA**TTTTT	9.27E-06	N	AroM family protein
5101081	53	*EC042_4755*	In K-12	-72	**TGTGA**CTGAGA**TCGCG**ATTAG	1.10E-05	-	Putative sugar kinase
5010825	22	*ulaA*	*ulaA*	-105	**TGCGG**GTCGCG**TCACA**TTTAA	1.10E-05	Y	PTS ascorbate transporter subunit IIC

3346695	76	*mscS*	*mscS*	-89	**TGTGA**TCTATT**TGGCA**AAATT	1.10E-05	N	Small-conductance mechanosensitive channel
943255	26	*EC042_0886*	*cecR*	-87	**TGTGA**CGCAGC**GCATA**AATTA	1.10E-05	N	TetR-family transcriptional regulator
4724645	552	*malE*	*malE*	-113	**TGTGA**TCTCTG**TTACA**GAATT	1.30E-05	Y	Maltose transport system substrate-binding protein MalE
4237814	145	*EC042_3992*	In K-12	-156	**CGTGC**TACCGG**TCACG**GTTTT	1.30E-05	-	Phage integrase family (tyrosine-type recombinase/integrase)
3771532	442	*dusB*	*dusB*	-278	**TGCGA**GCGATG**TCACA**AAAGG	1.30E-05	Y	tRNA-dihydrouridine synthase B
2769499	755	*nupC*	*nupC*	-60	**TGTGT**GTCAGA**TCTCG**TTTTC	1.30E-05	Y	Nucleoside permease NupC
4182076	16	*rfaK*	*rfaK*	646	**TGTGA**GCAACG**GCCAA**TTTTT	1.54E-05	N	Lipopolysaccharide N-acetylglucosaminyltransferase
3645715	857	*garD*	*garD*	-164	**TGCGC**GCTAAA**GCACA**GATTT	1.54E-05	N	D-galactarate dehydratase pseudo
3101613	429	*ascF*	*ascF*	-85	**GGTGA**CCGGTT**TCACA**AATAT	1.54E-05	Y	Arbutin-, cellobiose-, and salicin-specific PTS system EIIBC component
3985355	32	*EC042_3762*	*gntR*	-116	**ATCGA**CTCACG**TCACA**TTTTT	1.82E-05	Y	GntR-family transcriptional regulator
3723245	20	*gltB*	*gltB*	-180	**TGTGC**TTTAGC**GCAAA**ATTCT	1.82E-05	Y	Glutamate synthase [NADPH] large subunit
2497609	3901	*cdd*	*cdd*	-54	**TGAGA**TTCAGA**TCACA**TATAA	1.82E-05	Y	Cytidine deaminase
1250181	7061	*ptsG*	*ptsG*	-134	**TTTGA**CGGCTA**TCACG**TTTCA	1.82E-05	Y	PTS glucose transporter subunit IIBC
405459	29	*EC042_0361*	*yahK*	-153	**TTCGT**GGCAGA**TCACA**CTTTC	1.82E-05	N	NADPH-dependent aldehyde reductase YahK
3382409	2794	*nupG*	*nupG*	-44	**TGTTA**CCTGTG**GCAAA**TAATT	2.14E-05	Y	Nucleoside permease NupG
4287533	122	*ivbL*	*ivbL*	-85	**TGAGG**GGTTGA**TCACG**TTTTG	2.51E-05	Y	*ilvBN* operon leader peptide
2678707	161	*lrhA*	*lrhA*	-170	**CCTGA**AGTAGA**TCACA**GAATA	2.51E-05	N	NADH dehydrogenase operon transcriptional regulator
2272380	151	*EC042_2206*	*mtfA*	-173	**TGTGA**TTTTTG**TCACT**GACTT	2.51E-05	N	DgsA anti-repressor MtfA
4720171	269	*psiE*	*psiE*	-41	**TGTGA**CGGAGA**TCTAT**ATTTT	2.94E-05	Y	Putative phosphate starvation-inducible membrane protein
3190075	3516	*sdaC*	*sdaC*	-160	**AGTGA**TCTTGA**TCCCA**AATAC	2.94E-05	N	Serine transporter
1970534	77	*infC*	*infC*	1174	**CGTTA**ACTTCA**TCGCG**AATTT	2.94E-05	N	Translation initiation factor IF-3
409039	34	*EC042_0364*	In K-12	-22	**TGCGA**GAGAGA**TCACA**AAGTG	3.43E-05	-	LysE-family translocator
4266116	4	*EC042_4012*	S.G.	4601	**CGTGA**CGTTTG**TCGCG**GATGT	4.00E-05	-	Putative invasion '*air*'
3209376	59	*mltA*	*mltA*	-140	**TTTTA**CCCCGA**TCACG**GTTCT	4.00E-05	N	Murein transglycosylase A

2733499	23	*EC042_2583 /fadL*	*yfcZ/fadL*	-175	**AGTGA**CCGAAA**TCACA**CTTAA	4.00E-05	Y	Conserved hypothetical protein/ Long-chain fatty acid transport protein
1177277	639	*putA/putP*	*putA/putP*	-176/-225	**TCTGC**GGCAGT**TAACA**TTTTT	4.00E-05	Y	Bifunctional protein PutA/sodium proline symporter (proline permease) putP
150878	67	*EC042_0128*	*yadI*	-46	**TTTGA**CGGCTA**TCACC**CTTTA	4.00E-05	N	PTS sugar transporter subunit IIA
5030357	827	*EC042_4698*	*ytfJ*	-101	**TGCGG**TCAAGC**GCACA**AATCA	4.65E-05	N	YtfJ family protein
4236685	21	*EC042_3992*	In K-12	-156	**TGTGA**TCTTCC**GCCAA**ACTTC	4.65E-05	-	Tyrosine-type recombinase/integrase
2430107	2264	*EC042_2320*	*yegQ*	29	**TGTGA**TCGGTT**TCTAA**GAATT	5.40E-05	N	tRNA 5-hydroxyuridine modification protein YegQ
2095051	55	*pphA*	*pphA*	-14	**CGCGC**TAAAGA**TCACA**TAATC	5.40E-05	N	Serine/threonine protein phosphatase 1
1976942	41	*EC042_1890*	*yniA*	-55	**TGTGC**GTTAGC**TCGTG**TTTTT	5.40E-05	N	Fructosamine kinase family protein
4318893	16	*dnaA*	*dnaA*	-181	**TCTTC**TGTTTC**TCACA**GATTT	6.26E-05	N	Chromosomal replication initiator protein DnaA
2773567	1953	*EC042_2614*	*yfeC*	-75	**AGTTA**TTCATG**TCACG**GTTTC	6.26E-05	N	Putative DNA-binding transcriptional regulator
1623270	82	*EC042_1553*	*ydcH*	-72	**CACGA**TCCCGC**TCGCA**TTTTT	6.26E-05	N	YdcH family protein
1072066	118	*serC*	*serC*	-94	**TGCGA**TGTGTG**TCACT**GAATG	6.26E-05	Y	3-phosphoserine (phosphohydroxythreonine transaminase)
5027585	433	*cpdB*	*cpdB*	-66	**TGTGG**CATTCT**TCACT**GTTCT	7.24E-05	Y	2',3'-cyclic-nucleotide 2'-phosphodiesterase
2795558	1964	*ucpA*	*ucpA*	-74	**TGCGG**ATCAGC**TCACT**AATTC	7.24E-05	N	SDR family oxidoreductase UcpA
2531800	12889	*EC042_2405*	*mtlD*	-88	**TGTGA**GGGCGA**TCAAT**GAACT	7.24E-05	Y	Mannitol dehydrogenase family protein
2507357	39	*galS*	*galS*	-71	**TGTGA**CTCGAT**TCACG**AAGTC	7.24E-05	Y	HTH-type transcriptional regulator GalS
1190486	1700	*EC042_1101*	*ycdZ*	17	**TGAGA**TCGAGC**ACACA**TTTTA	7.24E-05	Y	DUF1097 domain-containing protein
4488945	19348	*udp*	*udp*	-68	**GGTGA**TGGGTA**TCACG**AAAAA	8.35E-05	Y	Uridine phosphorylase
4051417	311	*EC042_3818*	*yhjE*	-136	**TATGA**AATGCT**TCACA**TAATT	8.35E-05	N	MHS family MFS transporter
2068782	150	*sdaA*	*sdaA*	-129	**TGAGA**CAATCA**TCGCA**ATATT	8.35E-05	N	L-serine ammonia-lyase
1825856	114	*mlc*	*mlc*	1207	**TGTTG**CAGGGT**TAACA**TTTTT	9.62E-05	Y	Protein Mlc (making large colonies protein)
5121039	42	*EC042_4775*	In K-12	241	**ACTTA**GTCATG**TCACG**TTTTT	1.27E-04	-	Hypothetical protein
4530302	86	*glnA*	*glnA*	-106	**TGTGA**TCGCTT**TCACG**GAGCA	1.45E-04	Y	Glutamine synthetase
2054834	90	*EC042_1966*	*yeaV*	-219	**AGTGA**CGCGGT**TCAAA**TAAAC	1.45E-04	N	BCCT family transporter YeaV
228313	583	*gmhB*	*gmhB*	-661	**TGTGA**AGTGAT**TCACA**TCCGC	1.45E-04	N	D-*glycero*-β-D-*manno*-heptose-1,7-bisphosphate 7-phosphatase

4301258	22	*EC042_4041*	*yidE*	-63	**TTTGC**CTTATA**GCGCA**CATTA	1.90E-04	N	Putative transporter
2746629	6	*EC042_2592*	*insA*	221	**CGTTG**GCCTCA**ACACG**ATTTT	1.90E-04	N	Putative transposase (IS1)
719805	1585	*rnk*	*rnk*	-116	**TGTGA**CGCAAA**TCACT**TCAAA	1.90E-04	N	Nucleoside diphosphate kinase regulator
2622643	712	*EC042_2480 / EC042_2478*	*lysR/pseudogene*	- 72/ - 58	**TGCAA**AGTACA**TCACA**ATTTG	2.16E-04	N	LysR-family transcriptional regulator (partial)/ MFS transporter Pseudo
1302952	13	*minC*	*minC*	348	**TGTGA**CCGGCG**TTGTA**TTTTG	2.16E-04	N	Septum site determining protein MinC
1038450	134	*clpS*	*clpS*	-111	**TGTGA**CAGATG**TCGCT**GATGC	2.16E-04	N	ATP-dependent Clp protease adaptor protein
4466034	35	*EC042_4187*	S.G.	1	**CGAGG**TGTTGA**TCACG**AAACT	2.46E-04	-	Hypothetical protein
3353695	98	*EC042_3140*	*yggP*	487	**TGCGA**TGCTCA**TAAAC**ATATT	2.46E-04	N	Zinc-binding dehydrogenase
494392	1853	*tsx*	*tsx*	-110	**TATGT**TTCGTT**TCACA**GTTCT	2.46E-04	Y	Nucleoside-specific channel-forming protein Tsx
4217753	250	*EC042_3975*	S.G.	-20	**TGTTA**TCGTAT**ACACG**TTATT	2.80E-04	-	Hypothetical protein
4213344	4	*EC042_3970*	S.G.	301	**TGTGG**ATGAGA**GCAAA**GAACA	2.80E-04	-	Putative prophage protein
3996354	14	*EC042_3771*	*ybhG*	-211	**AGTTT**TATAGA**TCACA**CTTAT	2.80E-04	N	HlyD family secretion protein
2975867	1701	*EC042_2790*	*patZ*	710	**TCTGG**GTAGCA**TCACA**GCAGA	2.80E-04	Y	Protein lysine acetyltransferase
780664	122	*nagE/nagB*	*nagE/nagB*	-152/-159	**TGCGA**TACGAA**TTAAA**TTTTC	2.80E-04	Y	N-acetylglucosamine-specific PTS enzyme IIABC component/ Glucosamine-6-phosphate deaminase
4954732	3900	*aspA/fxsA*	*aspA/fxsA*	-132/-183	**AGTGA**TCCAGA**TTACG**GTAGA	3.17E-04	Y	Aspartate ammonia-lyase/ membrane protein FxsA
3846061	15	*frlA*	*frlA*	-107	**TGTGA**TCTTCC**TCCAC**ATTAC	3.17E-04	Y	Fructoselysine (psicoselysine) transporter
1868607	95	*malX/malI*	*malX/malI*	-70/-83	**TGTGA**TTTATG**CCTCA**CTATA	3.17E-04	Y	PTS maltose transporter subunit IIC/ mal regulon transcriptional regulator MalI
968344	134	*EC042_0909*	*ybiT*	1	**TGCGC**GGTTTG**TCATA**TCTCT	3.17E-04	Y	ABC transporter ATP-binding protein
48719	16242	*fixA*	*fixA*	-196	**TGTTT**TATAGA**TCACC**AATAT	3.17E-04	Y	Probable electron transfer flavoprotein subunit of oxidoreductase
3878947	818	*pckA*	*pckA*	-61	**CGCGA**CAAGGC**TCATA**GATTT	3.59E-04	Y	Phosphoenolpyruvate carboxykinase [ATP]
128586	5	*pdhR*	*pdhR*	-107	**TGTGC**ACAGTT**TCATG**ATTTC	3.59E-04	Y	Pyruvate dehydrogenase complex repressor
4160841	104	*lctP*	*lctP*	-38	**TCTGA**TGCTCA**TCTCC**TTGTC	4.06E-04	N	L-lactate permease
3966675	53	*EC042_3740*	In K-12	142	**CGTTT**GGCGCG**TCGCA**GCATT	4.06E-04	-	Putative acyltransferase
175121	61	*fhuC*	*fhuC*	553	**TGTGG**TGACTG**GCGCA**CTGTA	4.06E-04	N	Ferrichrome transport ATP-binding protein

5212600	384	*osmY*	*osmY*	-92	**AATGG**TCGCCA**TCACA**AAAGC	4.58E-04	Y	Osmotically inducible protein Y
5116358	827	*virK*	S.G.	774	**TGTGG**TGACTG**GTACA**GATTA	4.58E-04	-	Virulence protein
4167598	74	*grxC*	*grxC*	171	**ATTGA**CGCACA**GCACA**TTGGC	4.58E-04	N	Glutaredoxin 3
3615779	1543	*uxaC*	*uxaC*	-86	**ATTGA**TCTATC**TCACG**AAAAT	4.58E-04	Y	Glucuronateisomerase
773400	56	*EC042_0701*	In K-12	-481	**TCTGG**TGCGTC**TACCA**ATTTC	4.58E-04	-	Putative exported protein (pseudo)
5081393	155	*EC042_4739*	*ahr*	-121	**TGCGA**GCAAGC**TGGCG**CTTGC	5.16E-04	N	NADPH-dependent aldehyde reductase Ahr
3510478	251	*hyb0*	*hyb0*	-29	**TTTTG**CCTCAA**TCGCA**AAATC	5.16E-04	N	Hydrogenase-2 small chain
3335320	692	*serA*	*serA*	-95	**CGCGA**TGTTGA**GGAAA**TATCA	5.16E-04	Y	Phosphoglycerate dehydrogenase
3951363	1108	*rpoH*	*rpoH*	-94	**TGTGG**ATAAAA**TCACG**GTCTG	5.81E-04	Y	RNA polymerase sigma-32 factor RpoH
2158975	18	*ftnB*	*ftnB*	-98	**TGTGA**TCGTGG**AGTCA**ATTCT	5.81E-04	N	Ferritin-like protein 2
736066	860	*lipA*	*lipA*	-54	**TGTTG**TAATTA**TCAAC**TATTT	5.81E-04	N	Lipoyl synthase
4945596	33	*EC042_4606*	S.G.	787	**TCTGT**GCTCCA**TCAAG**CCAGT	6.52E-04	-	Hypothetical protein
4664179	66	*htrC*	*yjaZ*	120	**TCTTC**ACTGCT**TCACA**AAAGA	6.52E-04	N	Heat shock protein C
4101165	95	*cspA*	*cspA*	-39	**TATGG**CGTGCT**TTACA**GATTT	6.52E-04	N	RNA chaperone/antiterminator CspA
1444796	173	*stfR*	*stfR*	-106	**CCTGC**GTATCA**TAACA**ATATT	6.52E-04	N	Phage side tail fiber protein
927587	8	*moaA*	*moaA*	-11	**TGTGA**AGCCAT**GTACA**CCTTT	6.52E-04	N	Molybdenum cofactor biosynthesis protein A
4734493	69	*plsB/dgkA*	*plsB/dgkA*	-30/-119	**TGCTA**TCCTTG**CCGCG**CATTT	7.32E-04	N	Glycerol-3-phosphate acyltransferase PlsB/ diacylglycerol kinase
3293513	1002	*EC042_3089*	*ygfK*	-288	**AACGA**TTGCGT**TCAAA**TATTT	7.32E-04	N	Pyridine nucleotide-disulfide oxidoreductase
3152883	1050	*EC042_2961*	In K-12 *(ORF_o433)*	9	**TGTGT**CGGCGG**TCAAT**TTCCC	7.32E-04	-	Putative FAD-dependent oxidoreductase
2535535	3525	*spr*	*mepS*	-70	**CGTGT**CGCTTG**GCATT**TTTTT	7.32E-04	N	Lipoprotein
4926886	174	*set1A*	S.G.	39	**ACTGA**CGGTTT**TCCCA**GTCTT	8.20E-04	-	Enterotoxin 1
4817898	39	*dcuB*	*dcuB*	-29	**TGTGT**GAACCC**TCGCG**ATAAT	8.20E-04	N	Anaerobic C4-dicarboxylate transporter
3739372	164	*sspA*	*sspA*	-109	**TGTCG**GGTATT**GCTCA**TTTTT	8.20E-04	N	Stringent starvation protein A
1601165	3	*insB*	*insA1*	90	**TTTTA**CGCGTA**TGACA**GTCTC	8.20E-04	N	Transposase
33766	193	*rihC*	*rihC*	-35	**CGTGA**AGTCGA**TTAAG**TCATT	8.20E-04	N	Nonspecific ribonucleoside hydrolase (purine/pyrimidine ribonucleoside hydrolase)

20866	6	*insB*	*insB*	76	**TTTTA**CGCGTA**TGACA**GTCTC	8.20E-04	N	IS1 transposase B
5230955	45	*EC042_4888 /sltY*	*ettA/sltY*	-77/-112	**TGCGA**AGGTGA**TCGAC**CAACT	9.17E-04	N	Energy-dependent translational throttle protein EttA/ murein transglycosylase
4741151	625	*EC042_4418*	*pspG*	315	**TTTTA**TCCTTA**GAACA**AAATA	9.17E-04	N	Putative membrane protein
4398875	2967	*rbsD*	*rbsD*	-36	**GGTGC**CTTTTT**TCATT**TTTTT	9.17E-04	Y	Putative ribose transport (metabolism protein)
3517716	329	*EC042_3298*	*yghB*	-62	**TGCGT**CCGGGA**TCAAG**GCGTC	9.17E-04	N	Putative membrane protein
3196847	155	*fucP*	*fucP*	580	**TGTGT**TTATAC**GCACT**CAATT	9.17E-04	Y	L-fucose permease
2982786	81	*kgtP*	*kgtP*	-1097	**TGCTC**GCGCCG**TCACG**CTCGC	9.17E-04	N	Alpha-ketoglutarate permease
2945729	6	*EC042_2760*	*glrK*	-141	**GGTGC**CTGCCG**TCCAA**CTTCT	9.17E-04	N	Two-component system sensor kinase
1926713	17	*lpp*	*lpp*	-36	**TTTTA**TGTTGA**GAATA**TTTTT	9.17E-04	N	Major outer membrane lipoprotein
526093	29	*ppiD*	*ppiD*	-74	**CCCGT**TTCTTG**TCACA**ATAGG	9.17E-04	N	Peptidyl-prolyl cis-trans isomerase D
4944103	14	*EC042_4604*	S.G.	315	**TTTCG**CTGCCT**TCACA**AATTA	1.02E-03	-	Putative helicase (pseudogene)
4448062	161	*aslB*	*aslB*	-317	**CCTTC**GCATTG**TCACA**ACTTC	1.02E-03	N	Probable arylsulfatase-activating protein
4116976	107	*malS*	*malS*	-53	**TGAGA**GTTGAA**TCTCA**AATCA	1.02E-03	Y	Alpha-amylase
4089215	168	*EC042_3851*	*pseudo*	101	**AATGG**TTTCTG**TCACA**AATCC	1.02E-03	-	Conserved hypothetical protein
2576006	117	*napF*	*napF*	-76	**TTTGG**TCGCTC**TCAAT**TTTCA	1.02E-03	N	Ferredoxin-type protein
2559209	81	*EC042_2431*	S.G.	-223	**TGTTT**TGTAGA**TGTAA**TTTTA	1.02E-03	-	Putative prophage protein
4359930	46	*aadA1*	S.G.	61	**CGTTG**CTGGCC**GTACA**TTTGT	1.14E-03	-	Aminoglycoside adenylyltransferase
4243275	23	*EC042_3998*	S.G.	-45	**TGTGG**CGTTGA**TTGCG**ACACA	1.14E-03	-	Putative phage immunity repressor protein
3719403	3355	*elbB*	*elbB*	-72	**ACTGA**TTCATG**TAACA**AATCA	1.14E-03	N	Enhancing lycopene biosynthesis protein 2 (sigma cross-reacting protein 27A)
5093367	13	*EC042_4746*	S.G.	-696	**TGCGG**TTTTGC**GGCAA**GTTTT	1.27E-03	-	Conserved hypothetical protein
4747362	265	*aphA*	*aphA*	-33	**TGTGA**GATTTG**TTGCA**AAAAC	1.27E-03	N	Class B acid phosphatase
2677252	50	*nuoA*	*nuoA*	-296	**TGTGA**AGCAAT**GGAAA**AAATA	1.27E-03	N	NADH-quinone oxidoreductase subunit A
937104	77	*EC042_0881*	*ybhQ*	-32	**TCCGT**AAGACC**GCACC**TTTTT	1.27E-03	N	Putative membrane protein
855435	69	*aroG*	*aroG*	-126	**TGTAG**GAGAGA**TCTCG**TTTTT	1.27E-03	N	Phospho-2-dehydro-3-deoxyheptonate aldolase, Phe-sensitive
336180	16	*EC042_0305*	S.G.	876	**TCCTA**TTATCG**GCACC**ATTTA	1.27E-03	-	Conserved hypothetical protein

5118152	2	*EC042_4771*	In K-12	62	**TTTGA**GCAACA**GCCTG**TAATT	1.42E-03	-	Glycosyl transferase
4969712	85	*EC042_4632*	*yjeM*	-124	**TCCGC**TAAAGG**CCACA**ATTTA	1.42E-03	N	Putative permease
4522062	267	*engB*	*engB*	866	**CGCGA**CTAACT**TTACT**CTTTT	1.42E-03	N	Probable GTP-binding protein
4006081	208	*EC042_3779*	*dtpB*	-106	**TCTGA**CAGATG**TAAAC**TTTTT	1.42E-03	N	Putative oligopeptide transporter
3948285	20	*EC042_3720*	*panM*	-192	**TCTGC**CCTTTG**TTCCG**TTATT	1.42E-03	N	Putative acetyltransferase
2073958	3353	*manX*	*manX*	-142	**ACGGA**TCTTCA**TCACA**TAAAA	1.42E-03	Y	Mannose-specific PTS system EIIAB component
27718	12	*rpsT*	*rpsT*	-95	**GGCGC**TTATTT**GCACA**AATCC	1.42E-03	N	30S ribosomal protein S20
3892798	55	*gntT*	*gntT*	-136	**TGTTA**CCCGTA**TCATT**CACGT	1.57E-03	Y	High-affinity gluconate transporter
3204810	42	*gcvA*	*gcvA*	-211	**TGCGA**TTCAGA**CCATG**GTAGC	1.57E-03	N	Glycine cleavage system transcriptional activator
2315229	25	*EC042_2225*	*yeeO*	-50	**TCTGA**CTGGAC**TCGAA**CCAGT	1.57E-03	N	Putative membrane protein
2313387	14	*EC042_2224*	*yeeN*	1195	**TCTGA**CTGGAC**TCGAA**CCAGT	1.57E-03	N	Conserved hypothetical protein
2273560	22	*EC042_2207*	In K-12	-363	**TCTGA**CTGGAC**TCGAA**CCAGT	1.57E-03	-	Integrase
851102	39	*EC042_0769*	S.G.	-381	**TCTGA**GTTTGA**TGTTA**AAATT	1.57E-03	-	Hypothetical protein
5089682	1	*EC042_4744*	S.G.	-1219	**GGTGA**ACTCCA**TCTTG**GAATT	1.75E-03	-	Conserved hypothetical protein noise background
4956885	12	*groS*	*groS*	-93	**TGTGC**TGATCA**GAATT**TTTTT	1.75E-03	N	10 kDa chaperonin
4697783	24	*metH*	*metH*	-72	**TGTTG**CGTGGT**GGTCG**CTTTT	1.75E-03	N	Methionine synthase (5-methyltetrahydrofolate-homocysteine methyltransferase)
3608027	45	*EC042_3380*	*ygjR*	-17	**CGTGC**GCAGAT**ACGCA**TTATT	1.75E-03	N	Putative oxidoreductase
3012994	857	*EC042_2821*	In K-12	-194	**TCCGC**TCTGGC**TCATA**CTAAA	1.75E-03	-	Conserved hypothetical protein
263036	97	*EC042_0224*	S.G.	108	**ATTGC**CGGTGA**TCAAT**AAATT	1.75E-03	-	Putative type VI secretion system protein
3872334	283	*nudE*	*nudE*	-89	**TGTGA**ATTGCA**TGTTA**TTTAC	1.93E-03	N	ADP compounds hydrolase
965159	19	*EC042_0905*	*opgE*	-166	**TGCAA**TCCCCT**TCGCA**AAAGA	1.93E-03	N	Putative membrane protein
784980	12	*EC042_0709*	In K-12	-72	**TTTTT**CGCTGT**TCACC**TTTGG	1.93E-03	Y	Putative exported protein
540752	28	*EC042_0493*	*ybaA*	-171	**TGTTG**GTTCTC**TCGCA**ACGCT	1.93E-03	N	Conserved hypothetical protein
125271	40	*ampD*	*ampD*	-35	**CGTTA**GCTATC**TGGAG**TTTTA	1.93E-03	Y	N-acetyl-anhydromuramyl-L-alanine-amidase
4219192	26	*EC042_3978*	*yicG*	-202	**TCTTT**TCCACC**CCTCA**ATTTT	2.14E-03	N	Putative membrane protein
3687170	262	*argG*	*argG*	-303	**CCTGA**AACGTG**GCAAA**TTCTA	2.14E-03	Y	Argininosuccinate synthetase

3565809	4	*EC042_3343*	In k-12	1874	**CCTTG**CTGACT**TCCCG**ATTGT	2.14E-03	-	Putative membrane protein
3545591	258	*icc*	*icc*	-135	**TGTGT**TCAAGC**CAGCA**GATTT	2.14E-03	N	Repressor protein of division inhibition gene
3337434	10	*sbm*	*scpA*	-172	**AATGC**CTGATA**GCACA**TATCA	2.14E-03	N	Methylmalonyl-CoA mutase
2941936	195	*glnB*	*glnB*	-38	**TGTGT**TACGTT**TAGCA**GATCA	2.14E-03	N	Nitrogen regulatory protein p-II
2255919	34	*EC042_2180*	*pseudo*	612	**TCTGG**TTTGCG**GGGAG**TTTGT	2.14E-03	-	Putative prophage protein
459356	1136	*EC042_0411*	*yaiZ*	112	**TGTTG**TCTACG**CCACA**GACGG	2.14E-03	N	Putative membrane protein
4362407	3	*EC042_4096*	In K-12	275	**CGCGG**TGCGAC**GGACA**CATGC	2.37E-03	-	Putative acetyltransferase
4202532	33	*EC042_3956*	S.G.	435	**TATGA**CGCGCC**GAGCG**TCTTT	2.37E-03	-	Putative prophage protein
4177213	9	*htrL*	*htrL*	-152	**TGCTA**ATTGCA**ACAAA**CTAGT	2.37E-03	N	Putative lipopolysaccharide biosynthesis protein
2079650	271	*EC042_1989*	*pseudo*	74	**GTTGT**ATGTGA**TAACA**GATTT	2.37E-03	-	Conserved hypothetical protein (pseudogene)
1418735	75	*EC042_1347*	In K-12	-637	**ACAGG**CTTTCA**TCACA**TCTGA	2.37E-03	-	Putative phage protein
167375	194	*sfsA*	*sfsA*	103	**TGTGC**AACGCC**TGGCG**ATACC	2.37E-03	Y	Sugar fermentation stimulation protein
4590079	3873	*glpF*	*glpF*	271	**TGTTT**CGACAA**GCGCA**AAGTT	2.61E-03	Y	Glycerol uptake facilitator protein
3832206	568	*EC042_3619*	In K-12 *EO53_00955*	46	**ATTGA**TGTACT**GCATG**TATGC	2.61E-03	-	Hypothetical protein
1915871	36	*EC042_1835*	*ydhR*	-154	**AGTTG**GTTAGA**GCACC**ACCTT	2.61E-03	N	Conserved hypothetical protein
40727	46	*caiF*	*caiF*	-110	**CAGGA**TTTAGC**TCACA**CTTAT	2.61E-03	Y	Transcriptional activator
172	301	*thrA*	*thrA*	-164	**TCTGT**GCGCTA**TGCCT**ATATT	2.61E-03	N	Bifunctional aspartokinase I/homoserine dehydrogenase I
4353137	229	*cat*	In K-12	-58	**TTTTT**GAGTTA**TCGAG**ATTTT	2.88E-03	-	Chloramphenicol acetyltransferase
3932650	49	*gntR*	*gntR*	-37	**TGTTT**TCCCCG**GGAAG**GTTCT	2.88E-03	Y	Gluconate utilization operon repressor
2917822	42	*EC042_2736*	*trmJ*	-8	**CGTCT**CCACCA**GCACA**ATTCG	2.88E-03	N	Putative RNA methyltransferase
5141146	4	*EC042_4800*	In K-12	-156	**TGTTC**TTTCAT**TCATA**ACCTT	3.17E-03	-	Conserved hypothetical protein
3548332	26	*EC042_3327*	*ygiB*	-35	**TTTTG**TCCGTT**TCATT**TTTGT	3.17E-03	N	Putative lipoprotein
2480548	911	*EC042_2359*	*btsS*	-30	**TTTGC**GGGTAT**GTACG**ATTTT	3.17E-03	N	Two-component system sensor kinase
2416645	544	*mdtA*	*mdtA*	-382	**TCTTA**TCGTGA**AGGCA**TACTT	3.17E-03	N	Multidrug resistance protein
4978451	41	*orn*	*orn*	1136	**CGTTT**CCCGCT**CCAAA**TTCTT	3.49E-03	N	Oligoribonuclease

4682235	2044	*purH*	*purH*	-309	**TGTCG**ATGGAG**GCGCA**TTATA	3.49E-03	N	Bifunctional purine biosynthesis protein
4451798	32	*aslA*	*aslA*	-437	**TGCTC**ATTCCA**TCTCT**TATGT	3.49E-03	N	Arylsulfatase
3640881	17	*garK*	*garK*	1472	**CTTGA**GCCAGT**GAGCG**ATTGC	3.49E-03	N	Glycerate kinase 2
3421201	615	*mchA*	In K-12	-130	**AATGA**GTTACA**GGACA**GTTAT	3.49E-03	-	Microcin activation protein
3350600	5944	*epd*	*epd*	-144	**TCTTC**GGCTGG**ACAAA**CATTC	3.49E-03	Y	D-erythrose 4-phosphate dehydrogenase
5240772	4	*arcA*	*arcA*	-141	**TGCGT**TCTTGA**TGCAC**TTTCC	3.83E-03	N	Aerobic respiration control protein
4560899	4155	*fdhD*	*fdhD*	-18	**TGCTG**TGTTTT**TTTCA**CTTTC	3.83E-03	N	Putative formate dehydrogenase accessory protein
3690953	20	*EC042_3465*	S.G.	-108	**CCCGT**CCTCGG**TACCA**AATTC	3.83E-03	-	Hypothetical protein
3327528	16	*gcvT*	*gcvT*	-96	**CCCGG**TCCCCA**ACGCA**ATCGT	3.83E-03	Y	Aminomethyltransferase (glycine cleavage system protein
2782302	30	*zipA*	*zipA*	-48	**TGAGG**TAATCG**GCAAA**TACTC	3.83E-03	N	Cell division protein
1742072	146	*uxaB*	*uxaB*	-80	**TCAGG**AGTTGA**TCAAA**AAAGG	3.83E-03	Y	Altronate oxidoreductase
4968382	25	*frdA*	*frdA*	-88	**TGCGA**ACGCTA**TTCCA**CTGCT	4.20E-03	N	Fumarate reductase flavoprotein subunit
3187106	9	*syd*	*syd*	157	**CGTTT**ACCGGG**GAACA**AAATG	4.20E-03	N	Putative SecY-interacting protein
1149859	8	*cspG*	*cspG*	-35	**TTTGC**CGAAAG**GCCCA**AAATG	4.20E-03	N	Old shock-like protein
3276979	24	*EC042_3075*	S.G.	-248	**CCTTC**GCCCGC**TCCAA**TTTAT	4.61E-03	-	Conserved hypothetical protein
2151336	4034	*flhD*	*flhD*	-216	**CTTTA**CACTTA**TCTAA**GATTT	4.61E-03	Y	Transcriptional activator FlhD
2132403	20	*cutC*	*cutC*	519	**CGTGA**TGCTCC**AATCA**TTATG	4.61E-03	N	Copper homeostasis protein
1475788	284	*EC042_1404*	*lapA*	-67	**TGCGG**TAAGTT**GACCA**TAATT	4.61E-03	N	Putative membrane protein
1139173	29	*EC042_1056*	*yccA*	-274	**TCCGG**AATGAC**GCGCA**CTATA	4.61E-03	N	Putative membrane protein
816974	62	*sdhC*	*sdhC*	-212	**GTATA**GGCCGT**TCACA**AAATC	4.61E-03	Y	Succinate dehydrogenase cytochrome b-556 subunit
379869	89	*EC042_0340*	In K-12	-136	**CCTCA**TTTGTA**TGACA**AATGA	4.61E-03	-	Conserved hypothetical protein
4761274	7	*EC042_4436*	*pseudo*	-183	**TGAGC**AAGGCC**GCGCG**CATTA	5.04E-03	-	Putative type III effector protein (pentapeptide repeat protein) (pseudogene)
4674584	42	*hupA*	*hupA*	-118	**TTTGT**CGTACC**TGGAG**TCTTC	5.04E-03	Y	DNA-binding protein HU-alpha
4354065	11	*EC042_4087 A*	In K-12	367	**GGAGA**GCCTGA**GCAAA**CTGGC	5.04E-03	-	Putative plasmid-related protein
3899333	300	*malT*	*malT*	-72	**AATGC**AAGCGA**TGACG**TTTTT	5.04E-03	Y	Regulatory protein

2551890	24	*EC042_2422*	*intA*	-230	**TCCGA**CCCCCG**ACACC**CCATG	5.04E-03	N	Integrase
732988	1014	*cspE*	*cspE*	-54	**CGCTT**TGTCCA**GTGCA**ACTTT	5.04E-03	Y	Cold shock-like protein
5007735	171	*EC042_4669*	*bsmA*	-9	**TGGGG**GAATGT**TCCCG**AATTG	5.51E-03	N	Putative lipoprotein
4999291	10	*EC042_4657*	In K-12	-7	**TTTTT**GCATAA**TCATA**TTCTT	5.51E-03	-	Putative D-galactarate dehydratase/altronate hydrolase
3363224	24	*EC042_3148*	*yqgD*	191	**CGCGT**TTTATA**CCCCA**CAATG	5.51E-03	N	Putative membrane protein
5015959	41	*rpsF*	*rpsF*	-114	**CGTTA**TAGTAC**TGACA**TACCC	6.02E-03	Y	30S ribosomal subunit protein S6
4374307	45	*tetR*	S.G.	108	**TCTGA**CGACAC**GCAAA**CTGGC	6.02E-03	-	Tetracycline repressor
4260185	5	*EC042_4011*	S.G.	2663	**GGTTG**GAGCGA**TGCCA**ATTAT	6.02E-03	-	Transcriptional regulator
4192231	13	*rpmB*	*rpmB*	-151	**TCGGG**ACTTGA**GCACA**TCGCT	6.02E-03	N	50S ribosomal subunit protein L28
3186304	27	*EC042_2990*	*yqcC*	-208	**TGTGA**GCAGGA**AGCAA**TAGTT	6.02E-03	N	Conserved hypothetical protein
1866407	1300	*uidR*	*uidR*	-34	**TCTGC**ATGTTA**TCCAT**CATTA	6.02E-03	N	Uid operon repressor (TetR-family transcriptional regulator)
1127654	12	*EC042_1043*	In K-12 *(EO53_13390)*	-35	**GCTTA**TGCCGA**GCACC**CCTGG	6.02E-03	-	Conserved hypothetical protein
841420	39	*cydA*	*cydA*	-133	**GGTTT**ATTCAT**TAACA**TATTT	6.02E-03	N	Cytochrome d ubiquinol oxidase subunit 1
407967	9	*EC042_0363*	S.G.	153	**TATGA**AAAAAA**TCATA**ATTCC	6.02E-03	-	Hypothetical protein
5129030	57	*fecI*	*fecI*	-146	**TGTGT**CGGTCA**GAATG**ACTCA	6.57E-03	N	RNA polymerase sigma factor
4852738	11	*EC042_4524*	S.G.	-137	**ATCGA**TATAAT**GCAAA**TAATC	6.57E-03	-	Putative type VI secretion protein
4431631	17	*rho*	*rho*	-206	**ATCTG**TTCACT**TCGCA**TTTAA	6.57E-03	N	Transcription termination factor
3332837	78	*EC042_3122*	*fau*	-115	**ACCGC**GGAGCG**CCACA**TTCTT	6.57E-03	N	Conserved hypothetical protein
3080457	78	*EC042_2887*	*yqaB*	-554	**GCTGA**TATTGA**TACCG**CTACT	6.57E-03	N	Putative phosphatase
2529223	4498	*fruB*	*fruB*	-102	**GCTGA**AACGTT**TCAAG**AAAGC	6.57E-03	N	Multiphosphoryl transfer protein
1235648	3	*rluC*	*rluC*	-277	**CGCTA**ACTGCC**TGAAA**GATCA	6.57E-03	N	Ribosomal large subunit pseudouridine synthase C
5133814	2	*EC042_4789*	In K-12	793	**CCCGG**TCTTCG**TCTTC**TTATT	7.16E-03	-	Transposase
4752086	197	*ssb*	*ssb*	-56	**CGGAA**CCGAGG**TCACA**ACATA	7.16E-03	N	Single-stranded binding protein
4650149	75	*tufA*	*tufA*	-188	**GTTGG**TAGAGC**GCACC**CTTGG	7.16E-03	N	Elongation factor Tu
2833992	209	*EC042_2670*	S.G.	179	**GGAGG**TAGTCA**TCGCA**GCAAT	7.16E-03	-	Conserved hypothetical protein

5192683	765	*EC042_4851*	*btsT*	-130	**AGTTA**TTTACC**TTACT**TTACG	7.79E-03	Y	Putative carbon starvation protein
2842820	3	*dapA*	*dapA*	-57	**TGTTG**TATGCA**TGTTT**TTTTT	7.79E-03	N	Dihydrodipicolinate synthase
2310574	176	*amn*	*amn*	179	**AGTGA**TGGATA**GACAA**AAAGA	7.79E-03	N	AMP nucleosidase
1068899	7	*focA*	*focA*	78	**CCTGC**CTCTTC**GGCCA**CTTTG	7.79E-03	Y	Probable formate transporter 1
4325818	71	*tnaA*	*tnaA*	-326	**TTCGA**GGATAA**GTGCA**TTATG	8.47E-03	Y	Tryptophanase
3011421	48	*EC042_2820*	In K-12	-450	**GGCGG**TTGGCC**TCGTA**AAAAG	8.47E-03	-	Integrase
4412462	285	*hdfR*	*hdfR*	1067	**CGCGA**CCCCCT**GCGTG**ACAGG	9.20E-03	N	LysR-family transcriptional regulator (H-NS-dependent flhD regulator)
3419417	7	*mchS1*	In K-12	887	**GCTGG**TCGTTC**TAATA**CTATA	9.20E-03	-	Putative microcin esterase
242043	48	*EC042_0209*	In K-12	-361	**CGCGA**CCCCCT**GCGTG**ACAGG	9.20E-03	-	Putative lipoprotein
4812398	90	*melR*	*melR*	-19	**GCTGC**TGCACA**TAAAC**GTATC	9.98E-03	Y	Melibiose operon regulatory protein
3485206	29	*glcC*	*glcC*	-65	**CTTGT**CTTGGT**TAACT**TAATG	9.98E-03	Y	Glc operon transcriptional activator
3377662	3657	*ansB*	*ansB*	-169	**TTTTA**TGCCGT**TTAAT**TCTTC	9.98E-03	Y	L-asparaginase 2
2776365	79	*EC042_2617*	*fixA-like family protein*	-245	**AGCTG**AGCTAA**TCACC**CACTT	9.98E-03	N	Conserved hypothetical protein
4511085	67	*mob*	*mob*	2529	**TGATA**CCTCCA**GCAAC**CCTCA	1.08E-02	N	Molybdopterin-guanine dinucleotide biosynthesis protein B
516051	6	*cyoA*	*cyoA*	-189	**TTTGT**TATAAC**GCCCT**TTTGC	1.08E-02	Y	Cytochrome o ubiquinol oxidase subunit 2
3512048	18	*EC042_3292*	*gpr*	-37	**TCCTT**TCCACC**GTTCA**ATTTC	1.17E-02	N	Putative aldo/keto reductase
2685148	22	*EC042_2536*	*yfbV*	-134	**TTTTC**AGCCAA**TGGCA**TAATG	1.17E-02	N	Putative membrane protein
4646221	64	*murB*	*murB*	-229	**TACGG**CCGCCA**GGCAA**ATTCT	1.27E-02	N	UDP-N-acetylenolpyruvoylglucosamine reductase
3784875	200	*EC042_3554*	*yhdZ*	915	**ACTGC**CGCCAG**ACAAA**TTCTT	1.27E-02	N	ABC transporter, ATP-binding protein
4005281	31	*uspA*	*uspA*	-156	**TGTTA**AGACCA**TCCTT**ACCTT	1.37E-02	N	Universal stress protein A
3407089	69	*EC042_3190*	S.G.	33	**ACTCA**TTCAGT**TCATG**TTTCA	1.37E-02	-	Conserved hypothetical protein
4946926	8	*insB*	*insA*	43	**TTTGC**CGTTAC**GCACC**ACCCC	1.48E-02	N	Transposase
461937	112	*phoA*	*phoA*	230	**TGGGA**TGGGGG**ACTCG**GAAAT	1.48E-02	N	Alkaline phosphatase
5207449	48	*EC042_4865*	*yjjZ*	376	**TGCGA**AGGTGG**CGGAA**TTGGT	1.71E-02	N	Putative membrane protein
5167710	103	*EC042_4823*	*pseudo*	6	**TATGT**TCTGTA**ACAAG**TAAAA	1.84E-02	-	Conserved hypothetical protein

4046225	299	*treF*	*treF*	-160	**TTTCA**TTTTCA**GAACA**TTCAA	1.84E-02	N	Cytoplasmic trehalase
3740611	156	*rplM*	*rplM*	-165	**AGTTT**TTTTCC**CAAAA**CTTTT	1.98E-02	Y	50S ribosomal subunit protein L13
2214844	95	*EC042_2130*	S.G.	769	**GGGTT**AGATCG**TCACT**GTTTT	1.98E-02	-	Putative prophage protein
2164036	99	*EC042_2071*	*yecA*	-280	**TTTCT**TCCCGA**GCCCG**GATGG	1.98E-02	N	Conserved hypothetical protein
3562340	187	*EC042_3340*	In K-12 *ORF_o116*	-76	**GGTGG**GGAATT**TCGCC**CCGCC	2.45E-02	-	Conserved hypothetical protein
2647958	252	*pmrD*	*pmrD*	-46	**GCTAA**GAGTTT**TCACA**TCAAT	2.80E-02	N	Polymyxin B resistance protein

1 Peak centre indicated by MACS2 and annotated to EAEC 042 chromosome. Locations with more than one CRP site are highlighted in red.

2 Score represents the number of reads mapped to each position within the EAEC 042 genome.

3 Annotated gene to an identified CRP site.

4 EAEC 042 homologues to *E. coli* K-12 genes. Genes underlined represent matching genes targeted by CRP and previously indicated by Grainger *et al.* 2005. S.G denotes specific genes to EAEC 042. In K-12 indicates genes that are present in *E. coli* K-12 but with unknown functions.

5 Distance between an identified CRP site and the TSS (Translation Start Site) of annotated gene.

6 Matching sequence obtained from MEME SUITE (Bailey *et al.*, 2009).

7 *p*-value for CRP motif matches from MEME SUITE.

8 Reported to be regulated by CRP, as listed for *E. coli* K-12 in RegulonDB (Santos-Zavaleta *et al.*, 2019).

9 Product of encoding gene (Chaudhuri *et al.*, 2010).

Appendix C

The influence of CRP transcription factor on EAEC growth

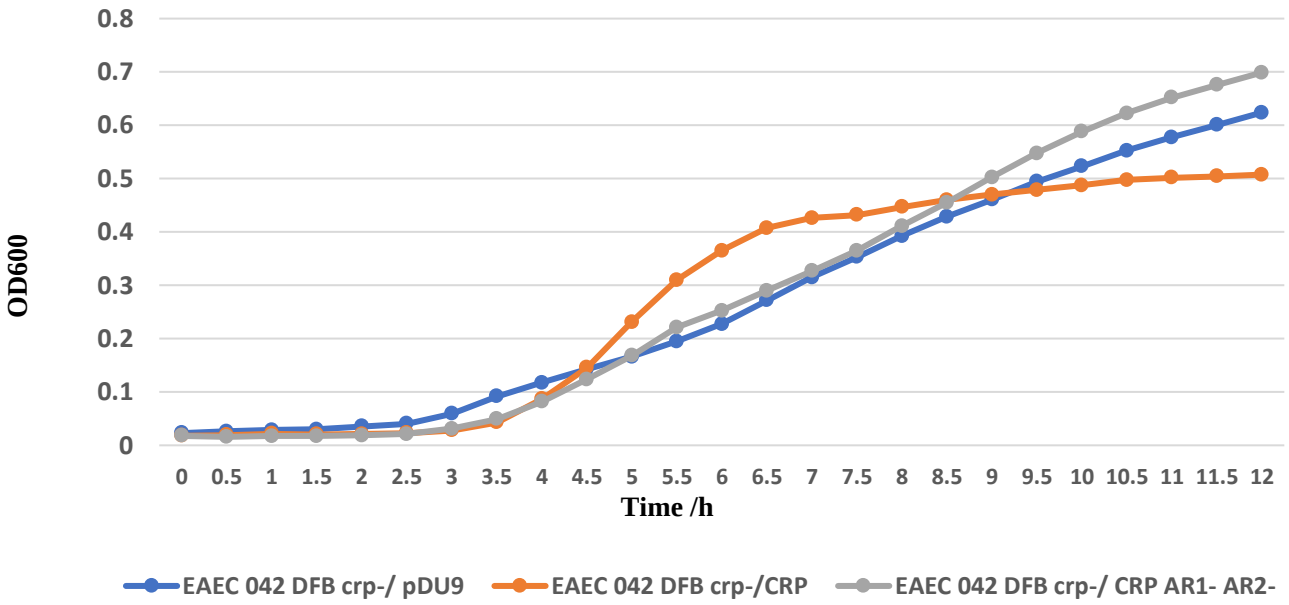


Figure 1: Growth kinetics of EAEC 042DFB Δ*crp*.

The line chart illustrates the growth curve of EAEC 042DFB Δ*crp* carrying pDU9 (blue line), EAEC 042DFB Δ*crp* carrying pDCRP (red line) and EAEC 042DFB Δ*crp* carrying pDCRP AR1^{-}AR2^{-} (green line). Bacterial cultures of each strain with desire plasmid were incubated for 12 hours at 37℃ with culture absorbance every 30 min at OD_{600} nm using Fluostar plate reader.

Appendix D

https://doi.org/10.1080/21505594.2022.2111754

Virulence

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/kvir20

Novel organisation and regulation of the *pic* promoter from enteroaggregative and uropathogenic *Escherichia coli*

Munirah M. Alhammadi, Rita E. Godfrey, Joseph O. Ingram, Gurdamanjit Singh, Camilla L. Bathurst, Stephen JW. Busby & Douglas F. Browning

To cite this article: Munirah M. Alhammadi, Rita E. Godfrey, Joseph O. Ingram, Gurdamanjit Singh, Camilla L. Bathurst, Stephen JW. Busby & Douglas F. Browning (2022) Novel organisation and regulation of the *pic* promoter from enteroaggregative and uropathogenic *Escherichia coli*, Virulence, 13:1, 1393-1406, DOI: 10.1080/21505594.2022.2111754

To link to this article: https://doi.org/10.1080/21505594.2022.2111754

VIRULENCE
2022, VOL. 13, NO. 1, 1393–1406
https://doi.org/10.1080/21505594.2022.2111754

Taylor & Francis
Taylor & Francis Group

RESEARCH PAPER

OPEN ACCESS Check for updates

Novel organisation and regulation of the *pic* promoter from enteroaggregative and uropathogenic *Escherichia coli*

Munirah M. Alhammadi[a,b*], Rita E. Godfrey[a*], Joseph O. Ingram[a], Gurdamanjit Singh[a], Camilla L. Bathurst[a], Stephen J.W. Busby[a], and Douglas F. Browning [a,c]

[a]Institute of Microbiology and Infection, School of Biosciences, University of Birmingham, Birmingham, UK; [b]Biology Department, Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia; [c]College of Health & Life Sciences, Aston University, Birmingham, UK

ABSTRACT

The serine protease autotransporters of the *Enterobacteriaceae* (SPATEs) are a large family of virulence factors commonly found in enteric bacteria. These secreted virulence factors have diverse functions during bacterial infection, including adhesion, aggregation and cell toxicity. One such SPATE, the Pic mucinase (protein involved in colonisation) cleaves mucin, allowing enteric bacterial cells to utilise mucin as a carbon source and to penetrate the gut mucus lining, thereby increasing mucosal colonisation. The *pic* gene is widely distributed within the *Enterobacteriaceae*, being found in human pathogens, such as enteroaggregative *Escherichia coli* (EAEC), uropathogenic *E. coli* (UPEC) and *Shigella flexneri* 2a. As the *pic* promoter regions from EAEC strain 042 and UPEC strain CFT073 differ, we have investigated the regulation of each promoter. Here, using *in vivo* and *in vitro* techniques, we show that both promoters are activated by the global transcription factor, CRP (cyclic AMP receptor protein), but the architectures of the EAEC and the UPEC *pic* promoter differ. Expression from both *pic* promoters is repressed by the nucleoid-associated factor, Fis, and maximal promoter activity occurs when cells are grown in minimal medium. As CRP activates transcription in conditions of nutrient depletion, whilst Fis levels are maximal in nutrient-rich environments, the regulation of the EAEC and UPEC *pic* promoters is consistent with Pic's nutritional role in scavenging mucin as a suitable carbon source during colonisation and infection.

ARTICLE HISTORY
Received 16 May 2022
Revised 27 July 2022
Accepted 6 August 2022

KEYWORDS
Bacterial gene regulation; CRP; enteroaggregative *Escherichia Coli* (EAEC); Fis; mucin; Pic; uropathogenic E. *coli* (UPEC); virulence

Introduction

Many bacteria use the regulation of transcript initiation as a key strategy to control the expression of their genes, and it is generally accepted that the evolution of this regulation is an essential contributor to bacterial fitness, as they acquire the ability to colonise different habitats. Genomic studies have shown that bacterial pathogens have acquired many genes that encode specific virulence determinants, which facilitate colonisation of their hosts [1,2]. The expression of these genes is tightly regulated at the level of transcription, involving the interplay of transcription factors. Many common pathogens possess a virulence-associated regulon whose transcription is controlled by a "dedicated" transcription factor [1,2]. Such factors appear to have been recruited and adapted specifically to control the expression of virulence genes. This appears to be the case in strains of enteroaggregative *Escherichia coli* (EAEC), an important human pathogen that causes diarrhoea in adults and children and is endemic both in industrialized and developing countries [3–6]. Here, the AggR transcription factor activates the expression of dozens of virulence determinants, *e.g.* the attachment adherence fimbriae (AAF) required for colonization, the anti-aggregative protein dispersin (Aap) and its dedicated secretion system [7–11], and, thus, AggR is the master controller of virulence in these strains [10–13]. However, expression of some EAEC toxins and secreted proteins, such as the plasmid-encoded toxin (Pet), enteroaggregative heat-stable toxin (EAST-1), *Shigella* enterotoxin 1 (ShET1) and the Pic mucinase (protein involved in colonisation), are not controlled by AggR [10,13,14].

The Pet toxin, a cytotoxic autotransporter protein, which is secreted during infection by EAEC strain 042 [15,16], belongs to a large family of secreted virulence factors (the serine protease autotransporters of the *Enterobacteriaceae* (SPATEs)), commonly found in enteric bacteria [17]. Expression of Pet is positively controlled by the cyclic AMP receptor protein (CRP)

CONTACT Stephen J.W. Busby s.j.w.busby@bham.ac.uk; Douglas F. Browning d.browning@aston.ac.uk
*These authors contributed equally to this work.
Supplemental data for this article can be accessed online at https://doi.org/10.1080/21505594.2022.2111754

and the nucleoid-associated protein Fis (factor for inversion stimulation) [18,19]. CRP is a global transcription factor whose activity is triggered by 5`-3` cyclic AMP (cAMP), an effector that signals depletion of certain nutrients and stress [20,21]. CRP functions as a dimer, with a helix-turn-helix motif in each subunit binding to adjacent major grooves at its DNA target [22,23]. The consensus base sequence for binding is 5`-TGTGANNNNNNTCACA-3` and, upon binding, at many targets, CRP interacts directly with RNA polymerase holoenzyme (RNAP), resulting in CRP-dependent activation of transcription by RNAP recruitment. CRP can make two distinct types of activatory interaction with RNAP [22,23]. In Class I activation, a surfaceexposed determinant on one subunit of the CRP dimer, Activating Region 1 (AR1), contacts a complementary determinant in the C-terminal domain of the RNAP alpha (α) subunit. In this instance, the DNA site for CRP is often located 49 base pairs upstream from the promoter −10 element (5`TATAAT-3`) [22–24]. In Class II activation, CRP usually binds 29 base pairs upstream of the −10 element and, additionally, contacts the *N*-terminal domain of the RNAP α subunit, with surfaceexposed determinant Activating Region 2 (AR2) [22–24]. In all *E. coli* strains studied so far, transcript initiation at hundreds of promoters is activated by CRP, functioning either by the Class I or Class II mechanism.

Like Pet, the Pic mucinase is a SPATE, and was originally identified in EAEC strain 042 and *Shigella flexneri* 2a, but, later, discovered in Uropathogenic *E. coli* (UPEC), and found to be distributed widely within the *Enterobacteriaceae* [17,25,26]. The Pic mucinase cleaves mucin, allowing EAEC both to use this substrate as a carbon source, and, to penetrate the mucus lining in the gut, increasing colonisation, and stimulating mucin secretion [25–30]. In addition, this important SPATE cleaves a wide range of substrates, including complement proteins, human leukocyte surface glycoproteins, and coagulation factor V, implicating Pic in bacterial serum resistance, haemagglutination and immunomodulation [25,26,31–33]. Here, we report the cloning and characterisation of the *pic* gene regulatory region from EAEC strain 042 and report that *pic* transcription is activated by CRP, using a Class I mechanism. We also investigated the *pic* promoter from UPEC strain CFT073, showing that it is also activated by CRP, but despite striking base sequence similarities with the EAEC *pic* promoter, activation is predominantly via a Class II mechanism. We show that both promoters have unusual features, and that CRP-dependent activation is counteracted by the global transcription factor Fis. We argue that what we observe is the result of evolutionary processes whereby transcription units are recruited into the regulons of different transcription factors.

Materials and methods

Bacterial strains, plasmids and DNA fragments

Table S1 describes the different *E. coli* strains that we employed in this study, together with details of plasmids and DNA fragments carrying promoters. Table S2 gives details of synthetic oligodeoxynucleotide primers that we used to amplify or to alter promoter-carrying DNA fragments Throughout, standard molecular biology procedures were used [34]. Each cloned promoter DNA fragment carries flanking EcoRI and HindIII sites to facilitate cloning and DNA sequences are numbered from the transcript start (designated as +1) as defined by the 5` RACE and potassium permanganate footprinting experiments. Upstream and downstream promoter locations are designated '-' and "+," respectively. Base pair substitutions in both the UPEC and EAEC *pic* promoters are designated pNX, where N is the substitution's position relative to the transcript start and X is the substitution in the non-template strand. As a source of DNA for EMSA (electromobility shift assays) and footprinting procedures, promoter fragments were cloned into pSR [35]. To determine promoter activity, each promoter fragment was cloned into the *lac* expression vectors pRW224 and pRW225 to generate either a *lacZ* transcriptional or translational fusion, respectively, [36]. Derivatives of pSR were selected for using media supplemented with 100 μg ml^{-1} ampicillin, whilst for pRW224 and pRW225 constructs, 15 μg ml^{-1} tetracycline was used.

DNA construction

The EAEC *pic*p042 and UPEC *pic*p073 promoter fragments were amplified by PCR using primers in Table S2, with EAEC 042 and UPEC CFT073 genomic DNA, respectively, as template. Flanking EcoRI and HindIII sites facilitated cloning into different plasmids. To introduce the p37A and p34T mutations in the UPEC *pic* promoter, the *pic*p073 fragment was amplified by PCR, using either primer picCTFp37A or picCTFp34T and primer lacZRev, with pRW224/*pic*p073 as template. Products were cloned into pRW224 and verified by DNA sequencing.

Point mutations were introduced into the putative −10 elements of *pic*p042 and *pic*p073 using megaprimer PCR [37]. During the first round of PCR, megaprimers were generated using primers *pic*p042-UKO, *pic*p073-

UKO, *pic*p042-DKO or *pic*p073-DKO with primer lacZRev and plasmids pRW224/*pic*p042 or pRW224/*pic*p073 as template. In the second round of PCR, purified megaprimer was used with primer D10520 and various pRW224/*pic*p042 or pRW224/*pic*p073 constructs as template. All DNA fragments were restricted with EcoRI and HindIII and cloned into pRW224.

The *pic*p042 and *pic*p073 promoter fragments were cloned into pRW225 to generate *lacZ* translational fusions. Fragment DNA was PCR amplified using either primers pic042–225 or pic073–225 with D10520 and pRW224/*pic*p042 or pRW224/*pic*p073 as template. PCR products were cut with EcoRI and HindIII and cloned into pRW225. All constructs were confirmed using Sanger DNA sequencing.

Error-Prone PCR of the EAEC pic*p042* promoter fragment

To introduce random substitutions throughout the EAEC *pic*p042 promoter fragment, error-prone PCR was used. Primers D10520 and lacZRev were used with pRW224/*pic*p042 as template with standard *Taq* DNA polymerase. PCR products were restricted with *Eco*RI and *Hin*dIII and cloned into pRW224. *E. coli* K-12 BW25113 cells were transformed with the resultant plasmid DNA and plated onto MacConkey agar plates, containing tetracycline. Colonies, in which the intensity of the Lac+ phenotype had decreased, were selected. Plasmid DNA was purified, and the sequence of *pic* promoter fragments obtained using Sanger Sequencing.

5` RACE (rapid amplification of cDNA ends)

Overnight cultures of BW25113 cells, containing either pRW224/*pic*p042 or pRW224/*pic*p073 were grown in LB medium at 37 °C with shaking until an $OD_{600} = 1$. Total RNA was isolated using an Isolate II RNA Mini Kit (Meridian Bioscience), as specified by the manufacturers. Specific mRNA was then converted to cDNA using primer SP1 and a 5`Race kit 2nd Generation (Roche), according to the manufacturer's instructions. Strand specific cDNA was A-tailed on the 3` end and then amplified by PCR using primers dT-anchor and lacZRev. PCR products were then purified and cloned into pJET1.2, using the Clone JET PCR Cloning Kit (Thermo Scientific), as specified by the manufacturers. Constructs were sequenced using Sanger Sequencing.

Proteins

Purified RNAP was obtained from New England Biolabs (Ipswich, MA, USA). Fis protein was purified as in Osuna *et al.* [38]. CRP over-expression and purification was carried out as described in Ghosaini *et al* [39].

Electromobility shift assays (EMSA)

EMSA assays using purified Fis were carried out as detailed in Godfrey *et al.* [40]. Purified *pic* promoter fragments were prepared from pSR based vectors and end labelled with [γ-^{32}P]-ATP. In each reaction, 0.5 ng of each ^{32}P labelled fragment was incubated with different amounts of Fis protein. The composition of the EMSA reaction buffer was 20 mM HEPES (pH 8.0), 5 mM $MgCl_2$, 50 mM potassium glutamate, 1 mM DTT, 25 μg ml^{-1} herring-sperm DNA, 0.5 mg ml^{-1} BSA and 5% (v/v) glycerol (10 μl final volume). Samples were incubated at 37°C for 20 min, and then immediately separated on a 6% polyacrylamide gel (12 V cm^{-1}), containing 0.25 x TBE and 2% glycerol. Dried gels were analysed with a Bio-Rad Molecular Imager FX and Quantity One software (Bio-Rad).

Potassium permanganate footprinting experiments

Potassium permanganate footprinting experiments were carried out on ^{32}P-end-labelled *pic* promoter fragments, as in Browning *et al.* [41]. Reactions contained 1.35 nM template DNA in 20 mM HEPES (pH 8.0), 50 mM potassium glutamate, 5 mM $MgCl_2$, 1 mM DTT, 0.5 mg ml^{-1} BSA and 200 μM cAMP (final volume 20 μl). 50 nM *E. coli* RNA polymerase (NEB) was added to reactions, where indicated. Samples were analysed using denaturing gel electrophoresis and calibrated with Maxam-Gilbert "G+A" sequencing reactions. Gels were visualised using Quantity One software (Bio-Rad) and a Bio-Rad Molecular Imager FX.

Assays of pic promoter activity

To analyse expression from the *pic* promoter, DNA fragments were cloned into the pRW224 and pRW225 *lac* expression vectors, and each construct was transferred into various *E. coli* K-12 Δ*lac* cells (Table S1). β-galactosidase activity was determined using a Miller [42] protocol as in [43]. Cells were cultured in LB medium with shaking at 37°C until they reached mid-logarithmic phase ($OD_{650} = 0.40.6$). In all cases, βgalactosidase activities are stated as nmol of ONPG

hydrolysed min^{-1} mg^{-1} dry cell mass and activities are the average of at least three independent biological replicates and average values are reported.

Isolation of the Pic protein from the extracellular medium

To examine the secretion of the Pic mucinase into the extracellular medium, *E. coli* K12 BW25113 cells and its isogenic Δ*crp* and Δ*fis* derivatives were transformed with plasmid pPic, which carries the EAEC *pic* gene and promoter region [27]. Cultures were grown overnight with shaking at 37°C in 50 ml of LB medium, supplemented with tetracycline. The OD_{600} of each culture was normalized to allow secreted protein levels to be compared, and cells were pelleted by centrifugation at 10,000 x g for 10 minutes at 4°C. Supernatants, which contain the secreted Pic protein, were filtered through 0.22 µm filters (Millipore) and proteins were precipitated by adding 10% (w/v) trichloroacetic acid (Fisher Scientific), as in our previous work [44]. Secreted Pic protein were analysed using SDS-PAGE and stained with Coomassie blue.

Modelling of mRNA secondary structure

Secondary structure prediction of the EAEC and UPEC *pic* mRNA transcripts was carried out using Mfold [45] .

Results and discussion

The Pic promoter from EAEC strain 042 is activated by CRP

To investigate the regulation of Pic mucinase expression, we inspected the DNA base sequence immediately upstream of the translation start of the annotated *pic* gene on the EAEC strain 042 chromosome. Using PCR with synthetic oligodeoxynucleotide primers, we amplified the *pic*p042 fragment, which carries the 167 base pair sequence, immediately upstream from the Shine-Dalgarno sequence of the *pic* gene, flanked by an upstream EcoRI and a downstream HindIII site (Figure 1(a,b)). This was then cloned into pRW224, a well-characterised low copy number *lac* expression vector such that any promoter in the fragment will drive *lac* gene expression [36] . Colonies of *Δlac* *E. coli* K-12 strain M182 carrying pRW224 with the *pic*p042 fragment scored as red Lac+ on MacConkey lactose indicator plates, whereas *Δlac* strain M182 carrying pRW224 with the starting short linker fragment scored as white Lac-.

In a preliminary experiment, we checked that longer fragments gave similar expression, and searched for base changes in the *pic*p042 fragment that reduced expression. To do this, we amplified the *pic*p042 fragment using error-prone PCR, cloned the resulting product into pRW224 and transformed plasmids into M182 *Δlac* cells, plating transformants on MacConkey lactose plates. The base sequence of the *pic*p042 fragment was then determined from colonies that appeared pale pink rather than red Lac+. The first mutations that we found by this method were located 50 bp from the EcoRI end of the fragment (Figure 1b), and we noted that this region carries the sequence element 5`-GGCGGTTCAGTTCACA-3`, which is a 7/10 match to the well-established consensus DNA target for the cyclic AMP receptor protein (CRP) [22,23]. Hence, we concluded that the *pic*p042 fragment carries the EAEC 042 *pic* promoter, and used the pRW224 *lac* gene expression vector to measure its activity and investigate its dependence on CRP. Figure 1(c) illustrates levels of βgalactosidase activity, measured in the M182 *Δlac* host, and its *Δcrp* derivative, carrying pRW224 with the starting *pic*p042 fragment, as well as derivatives carrying the p58A or p56C mutations that alter key bases in the suggested DNA site for CRP (Figure 1b). The data show that the *pic*p042 fragment carries one or more CRP-dependent promoters.

Previous research has shown that, for most CRP-activated promoters in *E. coli*, the promoter 10 hexamer element is located either 29 or 49 base pairs downstream from the centre of the DNA site for CRP [23]. Inspection of the *pic*p042 sequence shows good matches to the consensus 10 hexamer sequence (5`-TATAAT-3`) at both locations: 5`-TATATT-3` and 5`-TAACGT-3`, located 29 and 49 base pairs downstream, respectively (Figure 1b). Hence, we experimentally determined the transcript start point by sequencing the 5` end of RNA extracted from growing cells (Figure 1d) and used an *in vitro* potassium permanganate unwinding assay to measure the location of CRP-induced unwinding in the *pic*p042 fragment (Figure 1e). Figure 1d illustrates the result of primer extension analysis of the 5` end of the RNA transcript that initiates in the *pic*p042 fragment, after amplification using the 5`RACE method (rapid amplification of cDNA ends). The experiment shows that transcription in the *pic*p042 fragment starts 61 base pairs downstream from the centre of the DNA site for CRP and suggests that the functional 10 element is 5`-TAACGT -3`. This was confirmed by the *in vitro* experiment illustrated in Figure 1e, where potassium permanganate was used to detect local CRP-dependent unwinding. In brief, the purified fragment was 5`-end labelled with

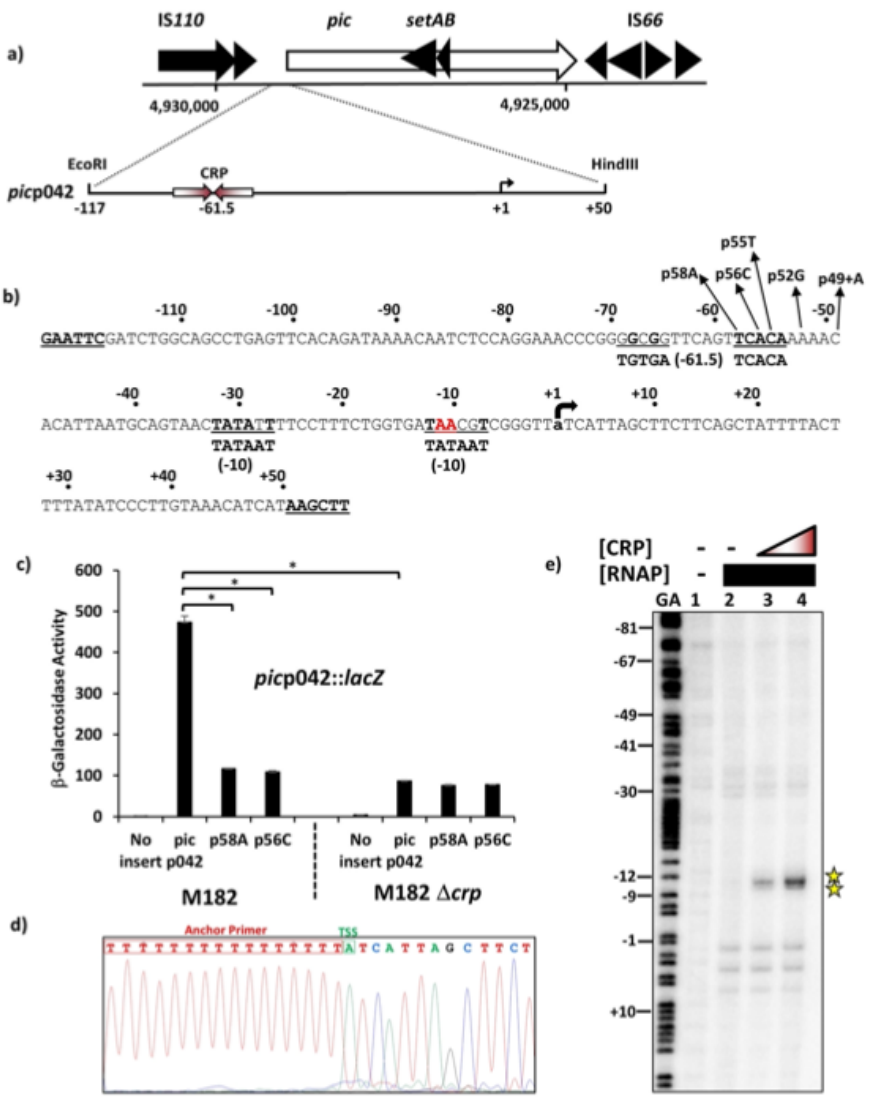


Figure 1. CRP-Dependent activation of the EAEC 042 *pic* promoter. (a) the panel shows the organisation of the *pic* locus on the EAEC 042 chromosome, detailing the flanking IS*110* and IS*66* family elements. The exploded view shows a schematic representation of the *pic*p042 promoter fragment. The CRP binding site is shown as inverted arrows and the start site of transcription (+1) is indicated by a bent arrow. The flanking EcoRI and HindIII sites used to clone the fragment are indicated. (b) the base sequence of the *pic*p042 promoter fragment from EAEC 042. The CRP-binding site and two proposed −10 promoter elements are underlined and matches to their respective consensus sequences are in bold [23,24]. The transcription start site (+1) is lower case and indicated by a bent arrow and substitutions, which change the CRP-binding site and adjacent sequences, are shown. The bases that were identified by potassium permanganate footprinting as being single stranded in the open complex are highlighted in red. Terminal EcoRI and HindIII sites are bold and underlined. (c) the panel details β-galactosidase activities determined in the Δ*lac E. coli* K-12 strain M182 and its *δcrp* derivative. Cells carried the *lac* expression vector pRW224 into which EAEC 042 *pic*p042 promoter fragments were cloned. The p58A and p56C substitutions disrupt the CRP binding site at the *pic* promoter, see (b). Cells were cultured in LB medium and βgalactosidase activities are stated as nmol of ONPG hydrolysed min^{-1} mg^{-1} dry cell mass, activities are the average of at least three independent biological replicates. Standard deviations are shown and * indicates $P < 0.01$ using a Student's *t*-test. (d) the panel shows the sequence trace from a 5` RACE experiment, which determined the EAEC 042 *pic* promoter transcription start site (TSS: green box). (e) End-labelled *pic*p042 AatII-HindIII fragment was incubated with RNA polymerase and CRP and subjected to potassium permanganate footprinting. The concentration of CRP was as follows: lanes 1 and 2, no protein; lane 3, 400 nM; lane 4, 800 nM. Each reaction contained 50 nM RNA polymerase and 200 µm cAMP. Maxam-Gilbert 'G+A' sequencing reactions have been run to calibrate the gel. The location of cleavage sites within the EAEC 042 *pic* promoter are highlighted by stars (and in red in panel (b)).

^{32}P at the HindIII end and incubated with purified *E. coli* RNA polymerase holoenzyme, either with or without cAMP and CRP. Permanganate modifies single-stranded thymine residues and results in strand cleavage that is detected by polyacrylamide gel analysis [46]. The data in Figure 1e show that CRP drives unwinding of bases in the 5`-TAACGT-3` hexamer, located 49 base pairs downstream of the centre of the DNA site for CRP. These results argue that transcription of the EAEC strain 042 *pic* gene is primarily driven by a Class I CRP-dependent promoter.

The UPEC strain CFT073 picU *promoter is also activated by CRP*

The paradigm UPEC strain CFT073 also secretes the Pic mucinase (PicU), which is expressed during colonisation of the bladder and plays a role during systemic infection [26,47,48]. Inspection of the base sequence upstream of the UPEC *picU* translation start revealed many similarities to the corresponding *pic* gene of EAEC 042. Strikingly, this region carries the sequence element 5`-TGTAACAGACATCACA-3`, which is a 9/10 match to the consensus DNA target for CRP (Figure 2a) [23]. Thus, again using PCR and synthetic oligodeoxynucleotide primers, we amplified the EcoRI-HindIII *pic*p073 fragment, carrying UPEC CFT073 sequence immediately upstream of the *picU* Shine-Dalgarno sequence, and cloned it into pRW224. As expected, colonies of M182 carrying pRW224 containing the *pic*p073 fragment scored as red Lac+ on MacConkey lactose plates. Figure 2(b) illustrates levels of βgalactosidase activity, measured in the M182 *Δlac* host, and its *Δcrp* derivative, carrying pRW224 with the starting *pic*p073 fragment, whilst Figure 2(c) shows the effect of derivatives carrying either the p34T or the p37A mutations that alter key bases in the suggested DNA site for CRP. The data show that the *pic*p073 fragment carries one or more CRP-dependent promoters. Remarkably, the *pic*p073 sequence, like *pic*p042, also contains hexamer elements with a good match to the consensus 10 hexamer (5`TATAAT3`) at both 29 base pairs and 49 base pairs downstream from the centre of the DNA site for CRP (respectively, 5`TATATT-3` and 5`-TAACGT-3`). Hence, again, we experimentally determined the transcript start point in growing cells (Figure 2d) and used *in vitro* potassium permanganate footprinting to measure the location of CRP-induced unwinding in the *pic*p073 fragment (Figure 2e). The data in Figure 2(d) show that transcription in the *pic*p073 fragment starts 40 base pairs downstream from the centre of the DNA site for CRP, suggesting that the functional 10 element must be 5`-TATATT-3`. This is confirmed by the *in vitro* potassium permanganate footprinting experiment illustrated in Figure 2(e), with CRP directing the unwinding of bases in the 5`-TATATT-3` hexamer, located 29 base pairs downstream of the centre of the DNA site for CRP. These results argue that transcription of *picU* in UPEC CFT073 is primarily driven by a Class II CRP-dependent promoter.

Similarities and differences between the EAEC 042 and UPEC CFT073 pic *promoters*

The activation of transcript initiation by CRP at target promoters is dependent on specific CRP-RNAP interactions involving AR1 at Class I promoters, and both AR1 and AR2 at Class II promoters [22,23]. To measure the effects of disrupting the different activating regions on *pic* promoter activity, we transformed the *lac* expression vector, pRW224, carrying either the *pic*p042 or *pic*p073 fragment into M182 *Δlac Δcrp* cells and then introduced a second plasmid, pD, carrying wild type CRP, or CRP carrying defective AR1, defective AR2 or defective AR1 and AR2. The data, illustrated in Figure 3a, indicate that, with the EAEC *pic*p042 fragment, disruption of AR1 prevents CRP-dependent activation, whilst disruption of AR2 increases activation. In contrast, with the UPEC *pic*p073 fragment, alteration of either AR1 or AR2 leads to a reduction in CRP-dependent activation.

In a second set of comparison experiments, we exploited previous findings that the activity of bacterial promoter −10 elements is reduced to near zero by alteration of the highly conserved A at the second position of the hexamer. Hence starting with either the *pic*p042 or *pic*p073 fragments cloned in pRW224, we inactivated either the upstream -10 element, or the downstream −10 element, or both (Figure 3b). Data in Figure 3(c) show that, with the *pic*p042 fragment, a large reduction in promoter activity is caused by inactivation of the downstream -10 element. In contrast, with the *pic*p073 fragment, the biggest reduction is observed when the upstream -10 element is mutated, although changing the downstream element does cause a significant reduction.

Taken together, our data suggest that CRP at the EAEC strain 042 *pic* promoter almost exclusively uses a Class I activation mechanism, but, at the UPEC CFT073 *picU* promoter, Class II activation predominates, though Class I activation is possible. Since, the consequence of initiation at one locus or another is to alter the 5`-end transcript sequence, we extended the HindIII end of each fragment to include the corresponding *pic* gene translation start, and cloned the

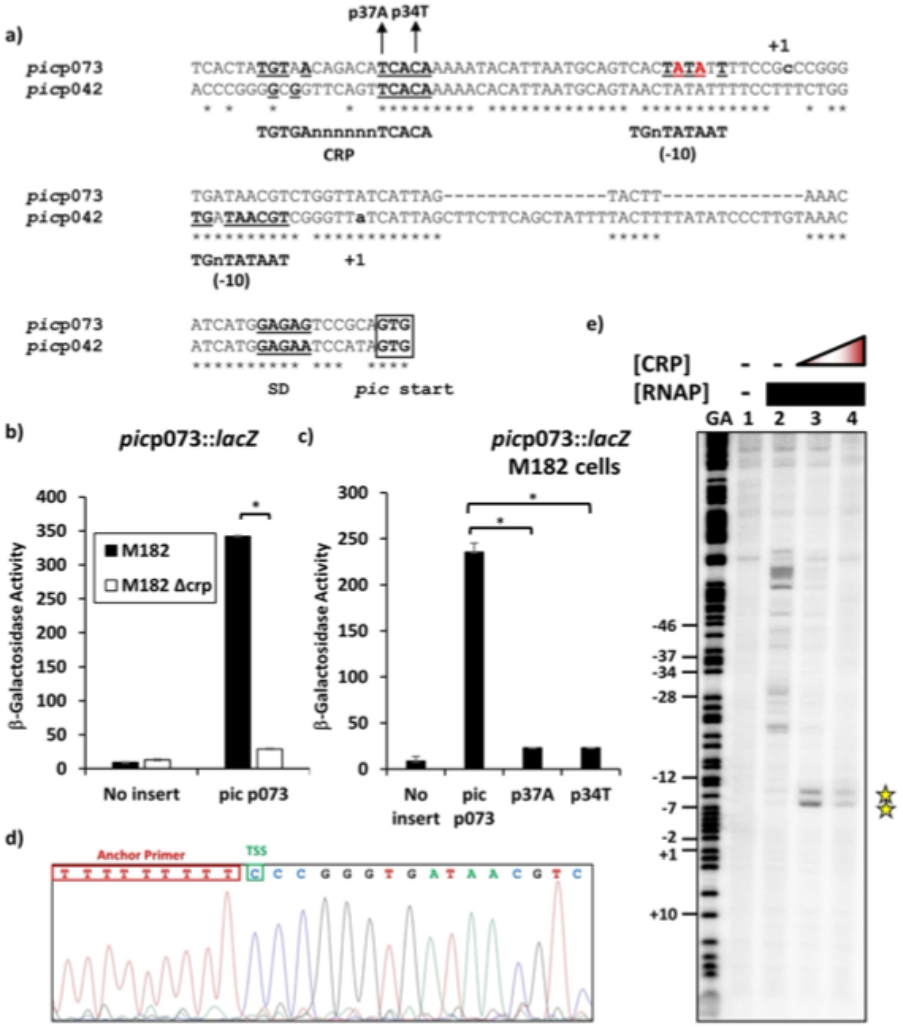


Figure 2. CRP-Dependent activation of the UPEC CFT073 *picU* promoter. (a) the panel shows an alignment of the UPEC *pic*p073 promoter sequence with the sequence from EAEC *pic*p042. The CRP-binding site and −10 promoter elements are underlined, with matches to their respective consensus sequences in bold [23,24]. For both fragments, transcription start sites (+1) are lower case bold, translation initiation codons (GTG) are boxed and the Shine–Dalgarno sequences (SD) are bold and underlined. The position of the p37A and p34T substitutions, which disrupt the CRP-binding site in the UPEC *pic*p073 promoter fragment, are shown and the bases, identified by potassium permanganate footprinting as being single stranded in the open complex, are in red. (b) the panel details β-galactosidase activities determined in the Δ*lac E. coli* K-12 strain M182 and its *δcrp* derivative. Cells carried the *lac* expression vector pRW224 into which the UPEC *pic*p073 promoter fragment was cloned. (c) the panel desplays β-galactosidase activities determined in the strain M182, with cells carrying various UPEC *pic*p073 promoter fragments, cloned into pRW224. The p37A and p34T substitutions disrupt the CRP binding site within the *pic*073 promoter fragment, see (a). In both (b) and (c), cells were cultured in LB medium and β-galactosidase activities are the average of at least three independent biological replicates. Standard deviations are shown and * indicates $P < 0.01$ using a Student's *t*-test. (d) the panel shows the sequence trace from a 5` RACE experiment, which determined the UPEC *picU* promoter transcription start site (TSS: green box). (e) End-labelled *pic*p073 AatII-HindIII fragment was incubated with RNA polymerase and CRP and subjected to potassium permanganate footprinting. The concentration of CRP was as follows: lanes 1 and 2, no protein; lane 3, 400 nM; lane 4, 800 nM. Reactions contained 50 nM RNA polymerase and 200 µm cAMP and Maxam-Gilbert 'G+A' sequencing reactions have been included. The location of potassium permanganate cleavage sites are shown starred.

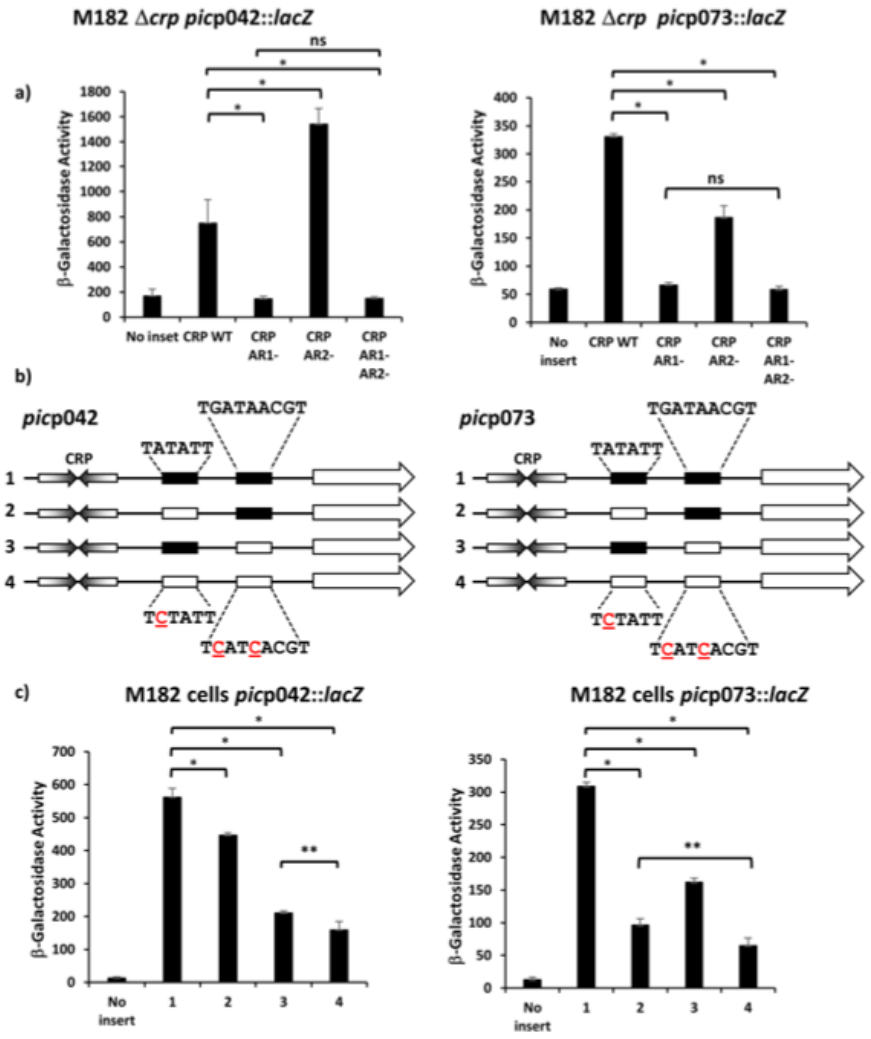


Figure 3. The EAEC and UPEC *pic* promoters use different −10 elements. (a) the panel details β-galactosidase activities from *E. coli* strain M182 *Δcrp*, carrying the *lac* expression vector pRW224 into which the EAEC *pic*p042 or UPEC *pic*p073 promoter fragments were cloned. Cells also carried either plasmid pDCRP or derivatives of pDCRP, encoding substitutions in CRP activating regions AR1 (HL159) and/or AR2 (KE101). (b) the panel shows schematic representations of the wild type EAEC *pic*p042 and UPEC *pic*p073 promoter fragments and fragments carrying substitutions in putative upstream and downstream −10 promoter elements. CRP binding sites are shown as inverted arrows and wild-type and mutant −10 elements are shown as filled or empty boxes, respectively. (c) the panel displays β-galactosidase activities in *E. coli* strain M182, carrying pRW224 into which the EAEC *pic*p042 or UPEC *pic*p073 promoter fragments depicted in (b) were cloned. Note that the numbering of promoter inserts is the same in both (b) and (c). For all experiments, cells were grown in LB, βgalactosidase activities are stated as nmol of ONPG hydrolysed min^{-1} mg^{-1} dry cell mass and are the average of at least three independent biological replicates. Standard deviations are shown and * indicates $P < 0.01$, ** indicates $P < 0.05$ and ns (not significant) $P > 0.05$, using a Student's *t*-test.

resulting fragments into pRW225, a derivative of pRW224 that permits translation fusions, thereby fusing the *N*-terminal end of the *pic* open reading frame to the *lacZ* gene. Hence, for each fragment, we could measure activity due to transcription only (using the pRW224 derivatives), or due to transcription and translation (using the pRW225 derivatives), and differences in the ratio of the two values will reflect differences in translation, due to altered initiation or changes in transcript stability. With promoter fragments derived from *pic*p042, this ratio, within experimental error, remains constant, irrespective of the mutations in either −10 element (Figure 4a). In sharp contrast, with the fragments derived from *pic*p073, relative expression of the

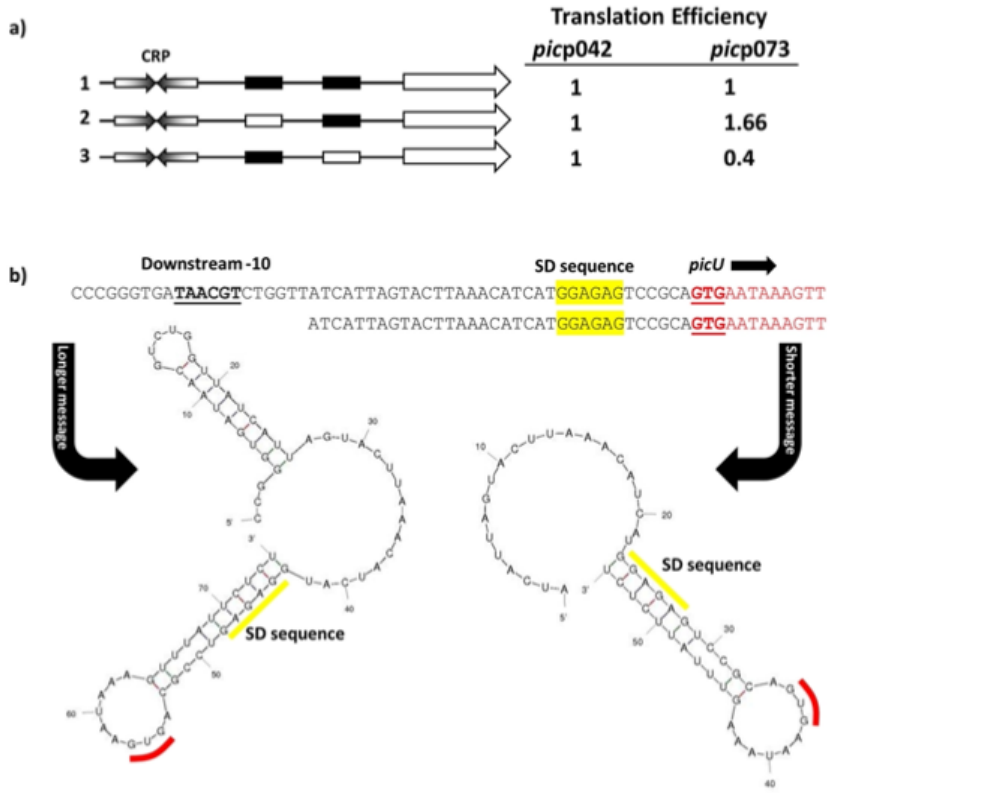


Figure 4. Translation efficiency differs for the EAEC *pic*p042 and UPEC *pic*p073 promoter fragments. (a) the panel shows schematic representations of the wild-type EAEC *pic*p042 and UPEC *pic*p073 promoter fragments and derivatives with substitutions in the upstream or downstream −10 promoter elements. CRP binding sites are shown as inverted arrows and wild-type and mutant −10 elements are shown as filled or empty white boxes, respectively. Fragments were cloned into pRW224 and pRW225 to generate lacZ transcription fusions and translation fusions, respectively. β-galactosidase activities were measured in *E. coli* M182 carrying either pRW224 or pRW225 into which the various EAEC *pic*p042 and UPEC *pic*p073 promoter fragments were cloned. Translation Efficiency is shown as the ratio of the values observed for pRW224 (transcription only) and pRW225 (transcription and translation) and is set as 1 for each starting wild type promoter. Cells were cultured in LB medium, βgalactosidase activities are the average of at least three independent values. (b) the panel shows the predicted secondary structure of the 5` ends of the mRNA when initiated from either the upstream or downstream -10 element in the UPEC *pic*p073 fragment. The location of the downstream −10 element, the Shine-Dalgarno sequence (SD) and translation initiation codon (GUG) are shown.

Pic-LacZ fusion protein is higher when Class II activation is prevented by mutation of the upstream -10 element, but is lower when Class II activation is optimised by mutation of the downstream −10 element, which results longer transcripts (Figure 4a). Inspection of secondary structures at the 5` end of transcripts that initiate in *pic*p073 (Figure 4b) suggests a simple reason for this, with the longer transcript, resulting from Class II activation, carrying an inhibitory 5` stem-loop structure that is absent from shorter transcripts. Note that, because promoter activity in the *pic*p042 fragment mostly uses the downstream -10 hexamer element, and because the EAEC strain 042 sequence contains two short insertions just upstream of the *pic* gene Shine-Dalgarno sequence (see Figure 2a), the lengths of the 5` untranslated sequences for the EAEC strain 042 and UPEC CFT073 genes are similar, but the EAEC *pic* transcript lacks the inhibitory structure.

Repression of *pic* promoter activity by Fis

To our knowledge, the only other EAEC virulence factor regulated by CRP is the *pet* gene, which encodes an autotransporter-toxin. Expression of the EAEC strain 042 *pet* gene, and other related toxins in other strains, is co-activated by CRP and Fis, another global DNA-binding protein that functions to modulate transcript initiation at specific promoters [18,19]. Thus, *pet* gene expression is greatly reduced in *fis* mutant strains. To

determine whether Fis plays any role at the *pic* gene promoter, we measured βgalactosidase activity in *fis*$^+$ and *Δfis* cells containing pRW224, carrying either the *pic*p042 or *pic*p073 promoter fragment. Data illustrated in Figure 5(a) show that, with both fragments, promoter expression is higher in the *Δfis* background, suggesting that Fis functions as a repressor. To examine this further, *E. coli* K-12 BW25113 cells and *Δfis* and *Δcrp* isogenic derivatives were transformed with plasmid pPic, which carries the complete EAEC 042 *pic* transcription unit [27], and Pic protein secreted into the external medium was analysed by SDS PAGE (Figure 5b). The results confirm that CRP and Fis work positively and negatively, respectively, at the *pic* promoter. Furthermore, EMSAs, performed with the *pic*p042 fragment show that Fis binds to multiple targets, consistent with a role as a repressor, with similar results observed for the *pic*p073 fragment (Figure 5c).

Conclusions

Bacterial transcription factors evolved to control the transcription of genes, in many cases, in response to external signals that are mediated by effectors [49]. Some transcription factors regulate the expression of a small number of genes, whilst others regulate hundreds [50]. Clearly, it is advantageous for bacteria to adapt to their surroundings, and this is especially true for the expression of virulence determinants that facilitate infection of mammalian hosts. Hence, in many bacterial pathogens, the expression of virulence factors, the proteins that are needed to establish and maintain successful infection, is regulated by one or more "dedicated" transcription factors [1,2]. These are often encoded by genes that are associated with plasmids or "pathogenicity islands" of genes encoding the virulence factors [1,2]. However, it is apparent that the expression of some genes that are essential for infection are

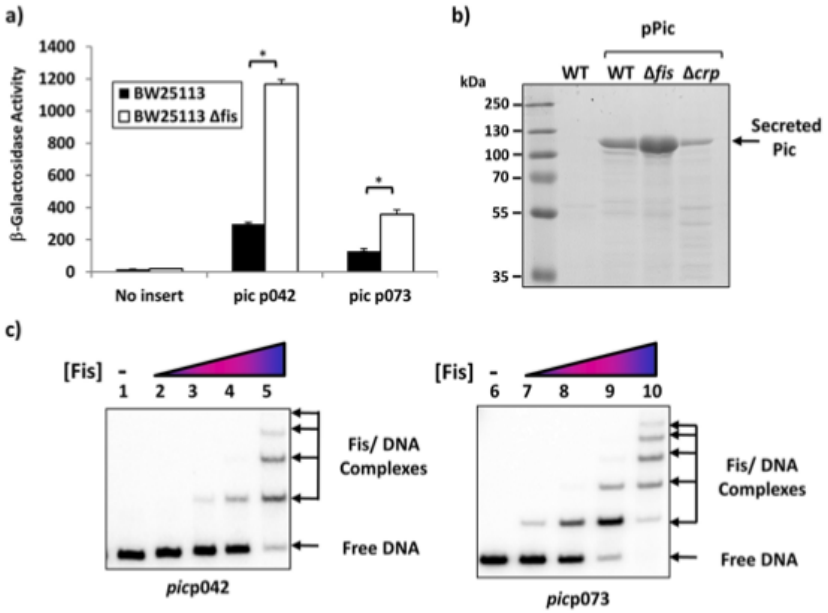


Figure 5. Fis represses the EAEC and UPEC *pic* promoters. (a) the panel illustrates β-galactosidase activities measured in the Δ*lac E. coli* K-12 strain BW25113 and its *δfis* derivative. Cells carried the *lac* expression vector pRW224 into which the EAEC 042 *pic*p042 or UPEC CFT073 *pic*p073 promoter fragments were cloned. Cells were grown in LB medium and βgalactosidase activities are the average of at least three independent biological replicates. Standard deviations are shown and * indicates $P < 0.01$ using a student's *t*-test. (b) the panel shows an SDS-PAGE gel run to quantify the amount of Pic protein secreted into the external medium by wild-type strain BW25113 (WT) and its *Δfis* and *Δcrp* derivatives. Where indicated, cells carried plasmid pPic, in which the EAEC 042 *pic* promoter and gene are cloned into plasmid pACYC184 [27]. The arrows indicate the location of secreted Pic protein (116 kDa). PageRuler prestained protein ladder (Fisher Scientific) were run to calibrated the gel. (c) Gel retardation assays of the EAEC *pic*042 and UPEC *pic*073 promoter fragments with purified Fis protein. End-labelled *pic* promoter fragments were incubated with purified Fis protein: lanes 1–5, *pic*p042 *Eco*RI-*Hin*dIII fragment; lanes, 6–10, *pic*p073 *Eco*RI-*Hin*dIII fragment. The amount of Fis protein in each reaction was: lanes 1 and 6, no protein; lanes 2 and 7, 200 nM; lanes 3 and 8, 400 nM; lanes 4 and 9, 800 nM; lanes 5 and 10, 1.35 μm. The location of Fis/DNA complexes and unbound DNA is indicated.

regulated by "regular" transcription factors, whose activity is not restricted to infection situations [18,19,51]. This appears to be the case for expression of the *pic* mucinase gene in both EAEC 042 and UPEC CFT073.

The Pic mucinase is distributed widely in enteric bacteria [25,47,52]. The EAEC 042 *pic* allele and corresponding promoter region is found in *S. flexneri* and many EAEC strains, including the Shiga-toxin-producing EAEC O104:H4 strain C227–11, which carries two chromosomal copies of this variant (Figure S1) [25,53]. The, UPEC *picU* gene and promoter is found in many UPEC strains (*e.g., E. coli* strain ABU83972 and *E. coli* clone D i2) but also faecal isolates, including the well-known probiotic *E. coli* strain Nissle 1917 (Figure S1) [54–56]. As faecal *E. coli* are considered to be the major source of bacteria causing urinary tract infections, this is perhaps not surprising. However, it is of note that Nissle 1917 has many virulence determinants that are similar to UPEC isolates (Table S3). Many studies that have examined the distribution of Pic have not distinguished between the different Pic variants carried by bacteria (*i.e.*, *pic* verses *picU*) and so it is unclear if *picU* is more likely to be carried by UPEC isolates [25,47,52].

Previously, Behrens *et al.* [57] investigated the EAEC 042 *pic* promoter and concluded that a single strong promoter was responsible for the majority of *pic* expression. Here we have reported the location of this promoter and the transcription factors involved in Pic regulation. The EAEC and UPEC *pic* promoters are unusual as they have evolved two precisely placed 10 hexamer elements, and, in the UPEC *pic*p073 fragment, both appear to be functional. Thus, the UPEC strain CFT073 *picU* promoter is the first example of what we dub as an ambiguous Class I/II CRP-dependent promoter. The advantage of this is still unclear, but data in Figure 4 show that *pic* gene translation levels differ according to which 5` transcript end is selected. Similarly, the factors that bias CRP-dependent activation in the *pic*p042 fragment towards the Class I option are unclear, but this study underscores that bacterial promoters can be active, whilst not conforming to simple textbook models. Interestingly, *Citrobacter rodentium* also carries the PicC mucinase, which has 79.15% and 78.57% identity to those from EAEC and UPEC, respectively (Figure S2) [58]. Comparison of the *picC* promoter with that from EAEC and UPEC suggests that it is likely CRP-dependent but a hybrid version of these two *E. coli* promoters, carrying the upstream of the UPEC *picU* promoter and the downstream region of the EAEC version (Figure S1). We surmise that the *C. rodentium pic* promoter functions by a Class II mechanism only, as the downstream −10 element is poor. Note that CRP in *C. rodentium*, and the *Enterobacteriaceae* in general, is highly conserved (Figure S3). Thus, it seems that evolution has produced three flavours of the *pic* promoter, CRP-dependent Class I only (EAEC), CRP-dependent Class II only (*C. rodentium picC*) and CRP-dependent Class I/II ambiguous (UPEC *picU*).

E.coli genomes encode over 300 transcription factors that support a complex regulatory network and, amongst the factors that make the most connections, are CRP and Fis [50] . CRP activity increases in response to nutrient stress, notably absence of glucose, whilst Fis levels rise during rapid growth [20,21,59,60]. Hence, genes whose products assure catabolic functions often carry promoters that are activated by CRP, whilst the promoters of genes whose products are not required during rapid growth are often repressed by Fis. In this light, it is easy to understand why the expression of bacterial mucinases should be regulated by both CRP and Fis, especially if ancestral mucinase genes encoded secreted proteinases with relaxed specificity for proteins other than mucin. Note that, in addition to mucin, Pic can degrade diverse substrates, including gelatine, complement proteins and human leukocyte surface glycoproteins [25,26,31,32] and that expression of both *pic* promoters is maximal in minimal medium (Figure S4). Presumably, this is part of a larger more general bacterial foraging response to cope with starvation and slow growth, and the ancestral gene was exapted in pathogens, such as EAEC and UPEC, specifically to facilitate mucin digestion, without its transcriptional regulation being altered. It is of note that the DNA surrounding the *pic* genes from EAEC 042, *S. flexneri* and UPEC CFT073 contains many genes from insertion sequences (Figure S5), suggesting that these genes are readily mobile and can be swapped between bacteria. As both CRP and Fis are found in all *Enterobacteriaceae*, should *pic* gene exchange occur, regulation will be maintained, perhaps highlighting the choice of global regulators over "bespoke" virulence transcription factors.

In addition to Pic, many pathogenic *Enterobacteriaceae* carry additional virulence genes, which encode other SPATE proteases, such as the Pet, SigA and Sat cytotoxins from EAEC 042, *S. flexneri* and UPEC CFT073, respectively [61,62]. In each case, toxin expression is activated by CRP, rather than a dedicated virulence transcription factor, but in this case CRP activation requires Fis as a co-activator, rather than a repressor. It has been argued that this ensures that toxin is released in specific conditions where bacterial growth and nutritional status is not too good, but not too bad and is resemblant of the

"Goldilocks effect" [18,19]. Thus, the interplay between CRP and Fis appears to be recurrent theme in SPATE expression and is likely to be applicable to similar virulence determinants from other *Enterobacteriaceae*.

Acknowledgement

This work was supported by BBSRC research grants BB/R017689/1 and BB/W00285X/1 to D.F.B. and S.J.W.B. The authors extend their appreciation for the support of the Saudi Cultural Bureau in London (UK SACB) and Princess Nourah bint Abdulrahman University (PNU), KSA. for funding Munirah Alhammadi.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by the Biotechnology and Biological Sciences Research Council [BB/R017689/1 and BB/W00285X/1] and the Saudi Cultural Bureau in London (UK SACB) and Princess Nourah bint Abdulrahman University (PNU), KSA.

ORCID

Douglas F. Browning http://orcid.org/0000-0003-4672-3514

Data availability statement

All data relating to this article are present in the article and the accompanying Supplementary Material.

References

[1] Dorman CJ. Genetics of bacterial virulence. Cambridge (MA): Blackwell Scientific Publications; 1994.
[2] Sansonetti PJ. Bacterial virulence basic principles, models and global approaches - infection biology. Cambridge (MA): Wiley Blackwell; 2010.
[3] Nataro JP, Mai V, Johnson J, et al. Diarrheagenic *Escherichia coli* infection in Baltimore, Maryland, and New Haven, Connecticut. Clin Infect Dis. 2006;43(4):402–407. DOI:10.1086/505867
[4] Wilson A, Evans J, Chart H, et al. Characterisation of strains of enteroaggregative *Escherichia coli* isolated during the infectious intestinal disease study in England. Eur J Epidemiol. 2001;17(12):1125–1130. DOI:10.1023/A:1021224915322
[5] Okeke IN, Lamikanra A, Czeczulin J, et al. Heterogeneous virulence of enteroaggregative *Escherichia coli* strains isolated from children in Southwest Nigeria. J Infect Dis. 2000;181(1):252–260.
[6] Franca FL, Wells TJ, Browning DF, et al. Genotypic and phenotypic characterisation of enteroaggregative *Escherichia coli* from children in Rio de Janeiro, Brazil. PLoS One. 2013;8(7):e69971. DOI:10.1371/journal.pone.0069971
[7] Elias WP Jr, Czeczulin JR, Henderson IR, et al. Organization of biogenesis genes for aggregative adherence fimbria II defines a virulence gene cluster in enteroaggregative *Escherichia coli*. J Bacteriol. 1999;181(6):1779–1785.
[8] Sheikh J, Czeczulin JR, Harrington S, et al. A novel dispersin protein in enteroaggregative *Escherichia coli*. J Clin Invest. 2002;110(9):1329–1337. DOI:10.1172/jci16172
[9] Nishi J, Sheikh J, Mizuguchi K, et al. The export of coat protein from enteroaggregative *Escherichia coli* by a specific ATP-binding cassette transporter system. J Biol Chem. 2003;278(46):45680–45689. DOI:10.1074/jbc.M306413200
[10] Morin N, Santiago AE, Ernst RK, et al. Characterization of the AggR Regulon in enteroaggregative *Escherichia coli*. Infect Immun. 2013;81(1):122–132. IAI.00676-12 [pii]. doi:10.1128/IAI.00676-12
[11] Abdelwahab R, Yasir M, Godfrey RE, et al. Antimicrobial resistance and gene regulation in enteroaggregative *Escherichia coli* from Egyptian children with diarrhoea: similarities and differences. Virulence. 2021;12:57–74. DOI:10.1080/21505594.2020.1859852
[12] Nataro JP, Yikang D, Yingkang D, et al. AggR, a transcriptional activator of aggregative adherence fimbria I expression in enteroaggregative *Escherichia coli*. J Bacteriol. 1994;176(15):4691–4699.
[13] Yasir M, Icke C, Abdelwahab R, et al. Organization and architecture of AggR-dependent promoters from enteroaggregative *Escherichia coli*. Mol Microbiol. 2019;111(2):534–551. DOI:10.1111/mmi.14172
[14] Estrada-Garcia T, Navarro-Garcia F. Enteroaggregative *Escherichia coli* pathotype: a genetically heterogeneous emerging foodborne enteropathogen. FEMS Immunol Med Microbiol. 2012;66(3):281–298.
[15] Eslava C, Navarro-Garcia F, Czeczulin JR, et al. Pet, an autotransporter enterotoxin from enteroaggregative *Escherichia coli*. Infect Immun. 1998;66(7):3155–3163.
[16] Henderson IR, Hicks S, Navarro-Garcia F, et al. Involvement of the enteroaggregative *Escherichia coli* plasmid-encoded toxin in causing human intestinal damage. Infect Immun. 1999;67(10):5338–5344.
[17] Ruiz-Perez F, Nataro JP. Bacterial serine proteases secreted by the autotransporter pathway: classification, specificity, and role in virulence. Cell Mol Life Sci. 2014;71(5):745–770.
[18] Rossiter AE, Browning DF, Leyton DL, et al. Transcription of the plasmid-encoded toxin gene from enteroaggregative *Escherichia coli* is regulated by a novel co-activation mechanism involving CRP and Fis. Mol Microbiol. 2011;81(1):179–191. DOI:10.1111/j.1365-2958.2011.07685.x
[19] Rossiter AE, Godfrey RE, Connolly JA, et al. Expression of different bacterial cytotoxins is controlled by two global transcription factors, CRP and Fis, that co-operate in a shared-recruitment mechanism. Biochem J. 2015;466(2):323–335.
[20] Gosset G, Zhang Z, Nayyar S, et al. Transcriptome analysis of CRP-dependent catabolite control of gene

expression in *Escherichia coli*. J Bacteriol. 2004;186 (11):3516–3524.
[21] Green J, Stapleton MR, Smith LJ, et al. Cyclic-AMP and bacterial cyclic-AMP receptor proteins revisited: adaptation for different ecological niches. Curr Opin Microbiol. 2014;18:1–7. DOI:10.1016/j.mib.2014.01.003
[22] Lawson CL, Swigon D, Murakami KS, et al. Catabolite activator protein: DNA binding and transcription activation. Curr Opin Struct Biol. 2004;14(1):10–20.
[23] Busby S, Ebright RH. Transcription activation by catabolite activator protein (CAP). J Mol Biol. 1999;293 (2):199–213.
[24] Browning DF, Busby SJ. Local and global regulation of transcription initiation in bacteria. Nat Rev Microbiol. 2016;14(10):638–650.
[25] Henderson IR, Czeczulin J, Eslava C, et al. Characterization of Pic, a secreted protease of *Shigella flexneri* and enteroaggregative *Escherichia coli*. Infect Immun. 1999;67(11):5587–5596.
[26] Parham NJ, Srinivasan U, Desvaux M, et al. PicU, a second serine protease autotransporter of uropathogenic *Escherichia coli*. FEMS Microbiol Lett. 2004;230 (1):73–83.
[27] Harrington SM, Sheikh J, Henderson IR, et al. The Pic protease of enteroaggregative *Escherichia coli* promotes intestinal colonization and growth in the presence of mucin. Infect Immun. 2009;77(6):2465–2473. IAI.01494-08 [pii]. doi:10.1128/IAI.01494-08
[28] Flores-Sanchez F, Chavez-Dueñas L, Sanchez-Villamil J, et al. Pic protein from enteroaggregative *E. coli* induces different mechanisms for its dual activity as a mucus secretagogue and a mucinase. Front Immunol. 2020;11:564953.
[29] Liu L, Saitz-Rojas W, Smith R, et al. Mucus layer modeling of human colonoids during infection with enteroaggragative *E. coli*. Sci Rep. 2020;10(1):10533. DOI:10.1038/s41598-020-67104-4
[30] Navarro-Garcia F, Gutierrez-Jimenez J, Garcia-Tovar C, et al. Pic, an autotransporter protein secreted by different pathogens in the Enterobacteriaceae family, is a potent mucus secretagogue. Infect Immun. 2010;78 (10):4101–4109.
[31] Abreu AG, Fraga TR, Granados Martínez AP, et al. The serine protease Pic from enteroaggregative *Escherichia coli* mediates immune evasion by the direct cleavage of complement proteins. J Infect Dis. 2015;212 (1):106–115. DOI:10.1093/infdis/jiv013
[32] Ruiz-Perez F, Wahid R, Faherty CS, et al. Serine protease autotransporters from *Shigella flexneri* and pathogenic *Escherichia coli* target a broad range of leukocyte glycoproteins. Proc Natl Acad Sci USA. 2011;108 (31):12881–12886. DOI:10.1073/pnas.1101006108
[33] Ayala-Lujan JL, Vijayakumar V, Gong M, et al. Broad spectrum activity of a lectin-like bacterial serine protease family on human leukocytes. PLoS One. 2014;9: e107920.
[34] Sambrook J, Russell DW. Molecular cloning : a laboratory manual. 3rd ed. Cold Spring Harbor (NY): Cold Spring Harbor Laboratory Press; 2001.
[35] Kolb A, Kotlarz D, Kusano S, et al. Selectivity of the *Escherichia coli* RNA polymerase Eσ 38 for overlapping promoters and ability to support CRP activation. Nucleic Acids Res. 1995;23(5):819–826.
[36] Islam MS, Bingle LE, Pallen MJ, et al. Organization of the LEE1 operon regulatory region of enterohaemorrhagic *Escherichia coli* O157: H7 and activation by GrlA. Mol Microbiol. 2011;79(2):468–483.
[37] Sarkar G, Sommer SS The "megaprimer" method of site-directed mutagenesis. Biotechniques. 1990 8 (4):404–407.
[38] Osuna R, Finkel SE, Johnson RC. Identification of two functional regions in Fis: the N-terminus is required to promote Hin-mediated DNA inversion but not lambda excision. Embo J. 1991;10(6):1593–1603.
[39] Ghosaini LR, Brown AM, Sturtevant JM. Scanning calorimetric study of the thermal unfolding of catabolite activator protein from *Escherichia coli* in the absence and presence of cyclic mononucleotides. Biochemistry. 1988;27:5257–5261.
[40] Godfrey RE, Lee DJ, Busby SJW, et al. Regulation of *nrf* operon expression in pathogenic enteric bacteria: sequence divergence reveals new regulatory complexity. Mol Microbiol. 2017;104(4):580–594.
[41] Browning DF, Cole JA, Busby SJ. Regulation by nucleoid-associated proteins at the *Escherichia coli nir* operon promoter. J Bacteriol. 2008;190(21):7258–7267. JB.01015-08 [pii]. doi:10.1128/JB.01015-08
[42] Miller J. Experiments in molecular genetics. Cold Spring Harbor (NY): ColdSpring Harbor Laboratory; 1972.
[43] Browning DF, Lee DJ, Spiro S, et al. Down-Regulation of the *Escherichia coli* K-12 *nrf* promoter by binding of the NsrR Nitric Oxide-Sensing Transcription Repressor to an upstream site. J Bacteriol. 2010;192 (14):3824–3828. JB.00218-10 [pii]. doi:10.1128/JB.00218-10
[44] Leyton DL, de Luna M, Sevastsyanovich YR, et al. The unusual extended signal peptide region is not required for secretion and function of an *Escherichia coli* autotransporter. FEMS Microbiol Lett. 2010;311 (2):133–139. DOI:10.1111/j.1574-6968.2010.02081.x
[45] Zuker M. Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Res. 2003;31 (13):3406–3415.
[46] Browning D, Savery N, Kolb A, et al. Assays for transcription factor activity. Methods Mol Biol. 2009;543:369–387.
[47] Heimer SR, Rasko DA, Lockatell CV, et al. Autotransporter genes Pic and *tsh* are associated with *Escherichia coli* strains that cause acute pyelonephritis and are expressed during urinary tract infection. Infect Immun. 2004;72(1):593–597.
[48] Subashchandrabose S, Smith SN, Spurbeck RR, et al. Genome-Wide detection of fitness genes in uropathogenic *Escherichia coli* during systemic infection. PLoS Pathog. 2013;9(12):e1003788.
[49] Busby SJW. Transcription activation in bacteria: ancient and modern. Microbiology. 2019. Epub ahead of print. DOI:10.1099/mic.0.000783
[50] Martinez-Antonio A, Collado-Vides J. Identifying global regulators in transcriptional regulatory networks in bacteria. Curr Opin Microbiol. 2003;6(5):482–489.

[51] Marteyn B, West NP, Browning DF, et al. Modulation of *Shigella* virulence in response to available oxygen in vivo. Nature. 2010;465:355–358. DOI:10.1038/nature08970

[52] Boisen F, Ruiz-Perez JP, Scheutz N, et al. Short report: high prevalence of serine protease autotransporter cytotoxins among strains of enteroaggregative *Escherichia coli*. Am J Trop Med Hyg. 2009;80 (2):294–301.

[53] Rasko DA, Webster DR, Sahl JW, et al. Origins of the *E. coli* strain causing an outbreak of hemolytic–uremic syndrome in Germany. N Engl J Med. 2011;365 (8):709–717. DOI:10.1056/NEJMoa1106920

[54] Zdziarski J, Brzuszkiewicz E, Wullt B, et al. Host imprints on bacterial genomes—Rapid, divergent evolution in individual patients. PLoS Pathog. 2010;6(8): e1001078. DOI:10.1371/journal.ppat.1001078

[55] Reeves PR, Liu B, Zhou Z, et al. Rates of mutation and host transmission for an *Escherichia coli* clone over 3 years. PLoS One. 2011;6:e26907. DOI:10.1371/journal.pone.0026907

[56] Reister M, Hoffmeier K, Krezdorn N, et al. Complete genome sequence of the gram-negative probiotic *Escherichia coli* strain Nissle 1917. J Biotechnol. 2014;187:106–107. DOI:10.1016/j.jbiotec.2014.07.442

[57] Behrens M, Sheikh J, Nataro JP. Regulation of the overlapping pic/set locus in *Shigella flexneri* and enteroaggregative *Escherichia coli*. Infect Immun. 2002;70 (6):2915–2925.

[58] Bhullar K, Zarepour M, Yu H, et al. The serine protease autotransporter Pic modulates *Citrobacter rodentium* pathogenesis and its innate recognition by the host. Infect Immun. 2015;83(7):2636–2650. DOI:10.1128/iai.00025-15

[59] Ball CA, Osuna R, Ferguson KC, et al. Dramatic changes in Fis levels upon nutrient upshift in *Escherichia coli*. J Bacteriol. 1992;174(24):8043–8056.

[60] Ali Azam T, Iwata A, Nishimura A, et al. Growth phase-dependent variation in protein composition of the *Escherichia coli* nucleoid. J Bacteriol. 1999;181 (20):6361–6370.

[61] Guyer DM, Henderson IR, Nataro JP, et al. Identification of Sat, an autotransporter toxin produced by uropathogenic *Escherichia coli*. Mol Microbiol. 2000;38(1):53–66. mmi2110[pii]. DOI:10.1046/j.1365-2958.2000.02110.x

[62] Yang F, Yang J, Zhang X, et al. Genome dynamics and diversity of *Shigella* species, the etiologic agents of bacillary dysentery. Nucleic Acids Res. 2005;33(19):6445–6458. 33/19/6445[pii]. DOI:10.1093/nar/gki954