Investigating Reciprocal Control of Adherence and Motility through the Lens of PapX, a Non-structural Fimbrial Repressor of Flagellar Synthesis

by

Daniel Jacob Reiss

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Microbiology and Immunology)
in the University of Michigan
2012

Doctoral Committee:

Professor Harry L.T. Mobley, Chair
Professor Michele S. Swanson
Associate Professor Maria B. Sandkvist
Associate Professor Mark A. Saper
Assistant Professor Nicole Koropatkin

For mom and dad

ACKNOWLEDGEMENTS

"WHAT IS THIS? WHAT HAVE YOU DONE? YOU'LL RUIN EVERYTHING!" –Keli'dan the Breaker

The investigation of PapX and its vagaries were alternatingly enticing, challenging, and rewarding (with a normal distribution of those three). Yet, the most important questions of my last four years were only peripherally related to the study of PapX. As anyone reading this probably knows, science is hard. The question whose answer will lastingly inform my life is essentially: "do I have what it takes to do science?" (answer pending).

I realize I am incredibly fortunate to have ended up under the tutelage and care of so many dedicated, supportive, and kind people. I do not know who instructed you to give me special treatment, but it totally worked. You have all made me feel special during my stay. I wish to show my gratitude.

Harry, I want to thank you first. Without you none of this would have been possible. It is clear to me that you made every effort to help me succeed. You've been tolerant of my foibles, encouraging when things were down, and a reliable pillar of support throughout my work. As they say on eBay: A+, would do business again.

I would like to thank mom and dad. Because of you I always felt free.

I want to thank the lab: Ariel, Mike, Sara, Steph, Chris, Chelsie, another Chelsie, Lauren, Sarguru, Rachel, Matt, Alejandra, Greta, Erin, Melanie, Rebecca, Rachel, Erika, Kelsie. We always had fun, and someone was always willing to help (patiently) with my struggles. I'll miss my lab family.

I want to thank Ron and Ellen and Penny of the MSTP. All of my friends in the MSTP and the staff in the MSTP office have made my stay here in the Big Scary Midwest feel like home. I feel much loved, thanks to you.

I want to send a lok'tar to Ael.

I want to thank our department for always treating me as a compatriot, instead of as a subordinate. I want to thank my committee for being (probably entirely too) kind to me, and for indulging my meandering ways. I want to thank Pat Schloss specifically for his patience in helping us with our last manuscript, and Oliver He for helping with the computational component of the JBC paper.

Lastly, I want to thank the money. Thanks, money! You were AI43363, AI59722, AI007528, and T32 GM07863 from the National Institute of Health.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF APPENDICES

ABSTRACT

Most uncomplicated urinary tract infections (UTIs) are caused by uropathogenic *Escherichia coli* (UPEC). Both motility and adherence are integral to UTI pathogenesis, yet they represent opposing forces. Therefore it is logical to reciprocally regulate these functions. PapX, a non-structural protein encoded by the *pheV*- but not *pheU*-associated *pap* operon encoding the P fimbria adherence factor of *E. coli* CFT073, represses flagella-mediated motility and belongs to a highly conserved family of winged-helix transcription factors. Thus, when P fimbriae are synthesized for adherence, synthesis of flagella is repressed. The mechanism of this repression, however, is not understood. *papX* is found preferentially in more virulent UPEC isolates, being significantly more prevalent in pyelonephritis strains (53% of isolates) than in asymptomatic bacteriuria (32%) or fecal/commensal (12.5%) strains. To examine PapX structure-function, we generated *papX* linker-insertion and site-directed mutants, which identified two key residues for PapX function (Lys^{54} and Arg^{127}) within domains predicted by modeling with I-TASSER software to be important for dimerization and DNA binding, respectively. SELEX in conjunction with high-throughput sequencing was utilized for the first time to determine the unique binding site for the bacterial transcription factor PapX in *E. coli* CFT073. It was necessary to write and implement novel software for the analysis of the results from this technique. The software, TFAST, is freely available (Appendix C) and has near-perfect agreement (κ =

0.89) to a gold standard in peak-finding software, MACS. Analysis of TFAST indicates that it correctly stratifies data to generate meaningful results, and successfully identified a 29 bp binding site within the *flhDC* promoter (TTACGGTGAGTTATTTTAACTGTGCGCAA), centered 410 bp upstream of the *flhD* translational start site. PapX bound the *flhD* promoter in gel shift experiments, which was reversible with the 29 bp sequence, indicating that PapX binds directly to this site to repress transcription of flagellar genes. Microarray, qPCR and promoter fusions indicate that PapX is not transcriptionally regulated itself. Co-precipitation studies indicate that PapX likely requires at least one cofactor for its repressive activity, and OmpA was identified as a promising candidate.

CHAPTER 1

INTRODUCTION

Urinary tract infections

The urinary tract consists of kidneys and ureters, referred to as the upper urinary tract, and bladder and urethra, referred to as the lower urinary tract. The kidneys filter blood and produce urine, which is transmitted via the ureters to the bladder. The bladder stores urine until it can be released from the body via the urethra [1]. Urinary tract infections (UTIs) occur when the upper or lower urinary tract is colonized by pathogenic bacteria, typically when a reservoir of a bacterial pathogen residing within the gastrointestinal tract [2-4] gains access to the urethra and ascends to the bladder [5]. Lower UTI, also called cystitis, is usually self-limiting [6] and rarely leads to severe complications. Upper UTI, also called pyelonephritis, attends greater severity and can lead to renal dysfunction, bacteremia or potentially fatal sepsis [7, 8]. Access of bacteria to the bloodstream is facilitated because there are only two layers of cells between the kidney tubules and host blood supply, the epithelial cells of the proximal tubule and the endothelial cells of the capillaries. UTI is a costly and potentially dangerous malady in the United States, associated with an estimated healthcare cost of $3.5 billion in 2000 [9] and 479,000 hospitalizations in 2006 [10].

Clinical diagnosis of bacterial UTI is frequently based on signs, symptoms and analysis of urine. Symptoms of lower UTI may include urinary urgency and pain associated with voiding. However, UTI can be asymptomatic in certain patients, either due to impaired sensory apparatus or due to infection with a strain of bacteria that causes asymptomatic bacteriuria (ABU). Signs of lower UTI may include frequent urination and altered urine characteristics, such as odor, color, or visible texture due to presence of blood or high concentrations of bacteria [11]. The signs and symptoms of UTI can vary; therefore, diagnosis is often heavily dependent on urinalysis or urine culture, and a diagnosis of UTI can be made based on the presence of $\geq 10^3$ bacteria/ml in a midstream clean-catch urine sample from a patient [7, 12]. The most common cause of upper UTI is lower UTI [5]. Upper UTI can present with all of the above signs, symptoms, and lab values, but additionally presents with flank pain or costovertebral tenderness, fever, chills, nausea and vomiting [13] as well as renal function impairment.

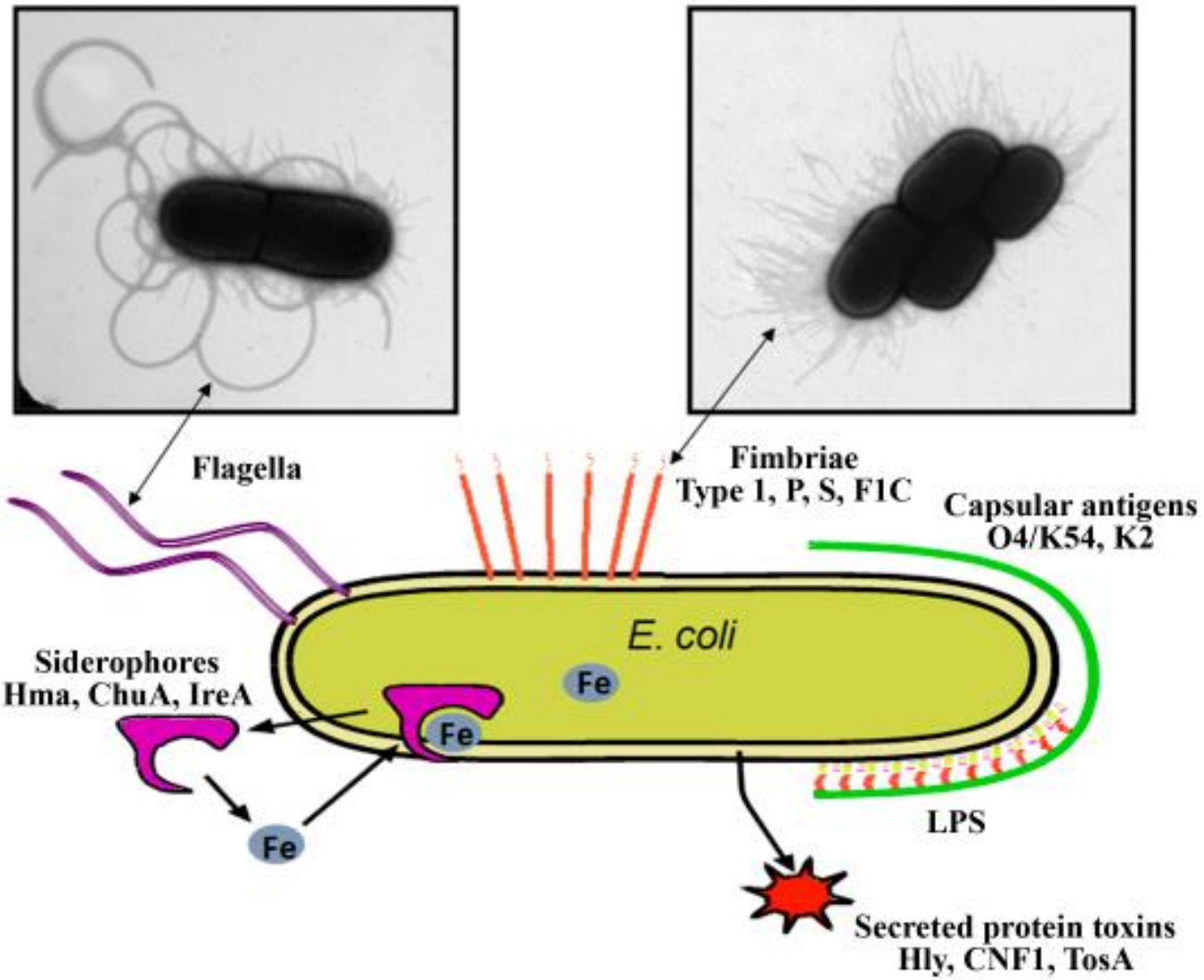
Microscopically, urine from a patient with UTI may reveal leukocytes and bacteria. Renal tubular casts consisting of cellular debris and leukocytes are diagnostic for upper urinary tract involvement [14]. Other laboratory values may indicate reduced ability of the kidneys to concentrate urine and general systemic inflammatory markers, including elevated leukocytes in the bloodstream, elevated erythrocyte sedimentation rate, and elevated C-reactive protein [15]. Histological examination of the urinary tract during pyelonephritis may reveal polymorphonuclear leukocytes (PMNs, also called neutrophils) extending from the renal papillae to the cortex, tubules filled with PMNs, and necrosis of proximal tubular epithelial cells. Glomeruli are frequently spared, even

when renal inflammation is severe [16]. The infection may spread beyond the kidneys, and bacteremia develops secondarily in 12% of cases of pyelonephritis [7, 8].

Uncomplicated UTI occurs in the setting of a healthy, non-pregnant adult woman with no urinary tract abnormalities, such as anatomical defects, impaired immune function, presence of devices, or obstructions [5]. 75-95% of all uncomplicated urinary tract infections are caused by *Escherichia coli*, specifically uropathogenic *E. coli* (UPEC) [17-27]. UTI-causing bacteria are studied using a combination of basic microbial techniques, murine, primate and canid animal models [28], and epidemiological and molecular analysis of samples derived from patients.

UPEC makes use of specialized virulence traits to cause UTI (**Figure 1-1)**. Iron is scarce in the urinary tract [29], so UPEC encodes numerous iron receptors and has developed multiple strategies for mediating iron acquisition [30] by encoding iron salvage and sequestration systems such as hemophores ChuA and Hma [31] and siderophores such as aerobactin, enterobactin, and salmochelin [32]. Extracellular polysaccharides may confer serum resistance or resistance to phagocytosis [33, 34], and disruption in synthesis of O4, K54, and K2 antigens led to fitness defects relative to parental strains in a murine model of UTI [35-38]. UPEC may also encode secreted toxins. α-hemolysin, encoded on the *hly* operon, causes enhanced damage and hemorrhaging in the uroepithelium and bladder, potentially liberating nutrients to expand UPEC's environmental niche within the urinary tract [39-41]. A related protein, TosA, is found much more commonly in UPEC compared to commensal strains of *E. coli*, acts as an adhesin, and is associated with persistence during bacteremia [42, 43]. Cytotoxic necrotizing factor 1 (CNF1) mediates changes in the cytoskeleton of neutrophils [44] and

Figure 1-1. UPEC virulence factors. UPEC strains encode numerous virulence factors that contribute to urovirulence. Shown are examples of adherence organelles (Type 1, P, S, and F1C fimbriae), motility apparatus (flagella), toxins (Hly, CNF1, TosA), extracellular polysaccharides (O4/K54, K2) and iron acquisition systems (Hma, ChuA, IreA). Insets show TEM images of flagellated and fimbriated bacteria.


Flagella
Fimbriae
Type 1, P, S, F1C
Capsular antigens
O4/K54, K2
E. coli
Fe
Siderophores
Hma, ChuA, IreA
Fe
Fe
LPS
Secreted protein toxins
Hly, CNF1, TosA

may reduce chemotaxis and antimicrobial activity in host phagocytes [44, 45]. Lateral gene transfer suggested by the existence of pathogenicity-associated islands (PAIs) suggests that many virulence traits may be shared, such as specific fimbriae, other adhesins, secreted toxins [46] and iron acquisition systems [47].

Ascension is the most common route of UTI [5]. To ascend, UPEC depends primarily on flagella [48-50], complex membrane-bound extracellular organelles that consist of a basal body, a hook, and a filament that propel the bacterium upwards through the urinary tract. UPEC strains unable to produce flagella are out-competed by parental strains in a mouse model of UTI [49, 50]. To resist being expelled from the urinary tract, bacteria are dependent on adherence. The presence of Type 1 fimbriae, which bind glycoproteins containing mannosides found in the lumen of the human urinary tract, is necessary for the development of UTI in a mouse model of infection [51-54]. P fimbriae are epidemiologically associated with UPEC strains relative to commensal strains [55, 56], and bind the α-D-Gal-(1→4)-β-D-Gal digalactoside moiety of the P blood group antigen, found on both erythrocytes [57] and uroepithelial cells [58].

Motility is the force that causes UPEC to move up the urinary tract, while adherence factors resist expulsion by the flow of urine. Thus, these forces should antagonize one another. Yet both are demonstrably important for urovirulence. Therefore, it would make sense that the expression of these features would be temporally coordinated during UTI. Biophotonic imaging studies indicate that flagellar expression is highly coordinated during UTI, with peak flagellar expression appearing during ascension of the ureters to the kidneys [59]. The plausibility of reciprocal control between these factors is underscored in work on Type 1 fimbria. When the genes for Type 1

fimbria are constitutively expressed, motility is greatly reduced, even if no functional Type 1 fimbria are produced on the surface of the bacterium [60]. The question of the underlying mechanisms mediating reciprocal control between the pathogenicity factors of motility and adherence in UPEC is central to understanding UPEC pathogenesis.

Adherence

UPEC strains can encode up to 15 epidemiologically and genetically distinct clusters of fimbriae, proteinaceous organelles primarily implicated in adherence; individual strains have been observed to encode up to 11 distinct fimbriae. Fimbriae project from the surface of bacteria and can be directly observed by transmission electron microscopy (**Figure 1-1**). Recent studies of fimbriae in UPEC support the model that adherence is an important component in UTI pathogenesis [61]. In these studies, the genomes of 303 *E. coli* strains were analyzed by multiplex PCR to assay for the presence of fimbrial operons. A bimodal distribution of fimbriae is encoded by uropathogenic and non-pathogenic *E. coli*: UPEC strains carry a high number and diverse array of fimbrial operons, whereas non-pathogenic *E. coli* tend to carry only a few. This suggests that acquiring or encoding a diverse array of fimbriae is an important characteristic of UPEC. Adherence is known to be important to prevent the flushing mechanism of voiding, and specific adherence organelles have been implicated in eliciting protective immune responses and are part of vaccine development [62, 63], highlighting their importance in the pathogenesis and study of uropathogenic bacteria. While UPEC strains can encode an array of distinct fimbriae, certain fimbriae have been identified as critical to the

development of a UTI. Among the best characterized fimbriae are P, Type 1, S and F1C fimbria.

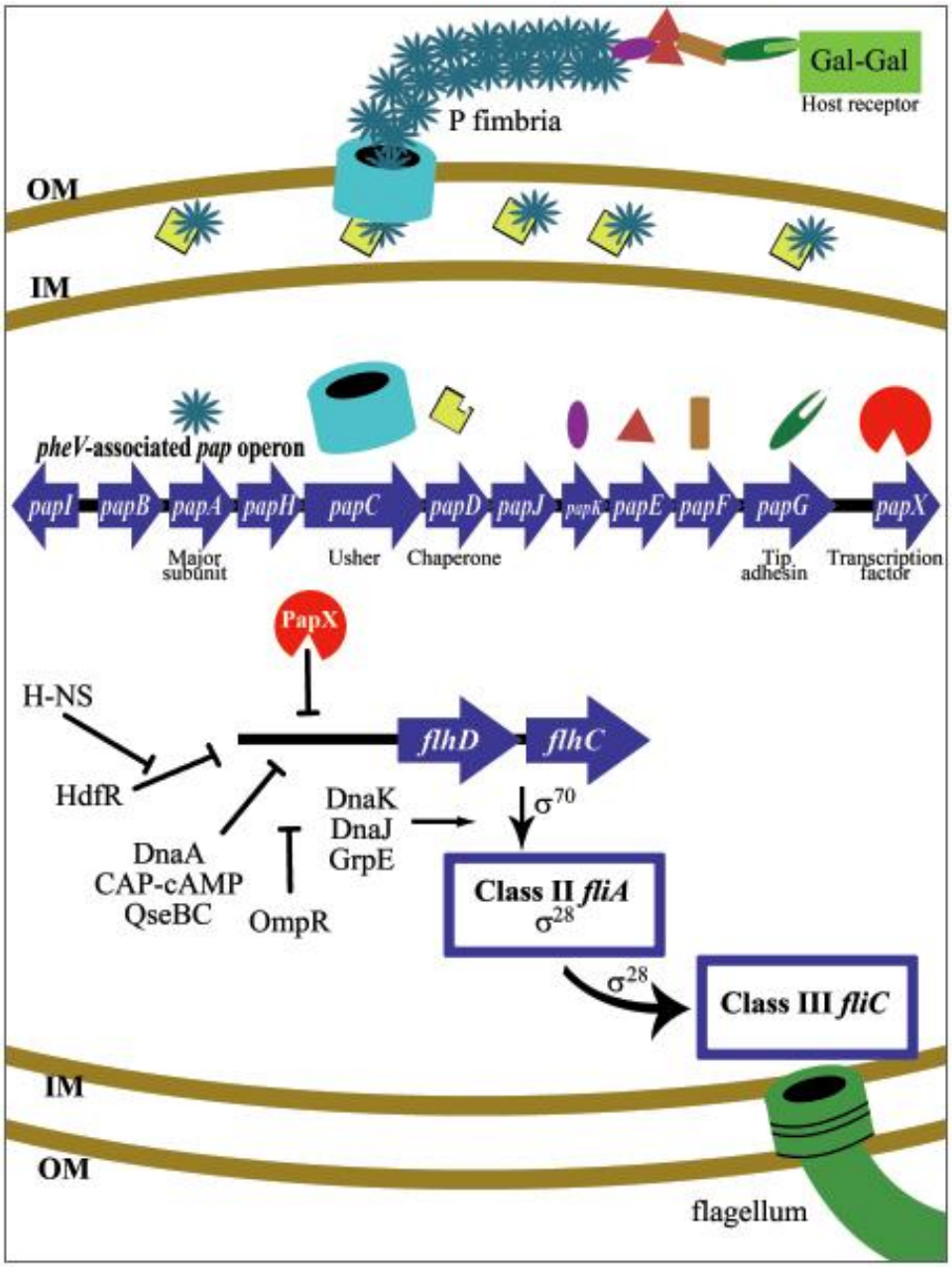
Early studies of fimbriae originated with examination of whole bacteria and their ability to agglutinate erythrocytes [64, 65]. Cross-linking of cells was used as a proxy for adherence, because bacteria binding to a given receptor should be able to bind to receptors expressed on the surface of multiple eukaryotic cells, bringing them into close proximity. Hemagglutination is still useful as a technique for assaying the phenotypic adherence properties of whole bacteria, as well as whole-cell agglutination and binding assays in general [56]. The use of different eukaryotic cells to assess adherence of bacteria is helpful in illuminating binding and substrate specificity [66-68]. In parallel to whole-cell phenotypic assays, electron microscopy techniques were employed to reveal the morphologic characteristics of adherent bacteria. Close examination of adherent bacteria revealed hair-like projections of cells, and the association of these extracellular structures, as well as the intuitive consistency that extracellular structures could be responsible for adherence, led investigators to posit that these structures, later identified as fimbria, were the factor mediating adherence of bacteria [69, 70].

The earliest virulence trait established for UPEC was P fimbria. In 1976, Svanborg-Eden *et al.* **[56]** made use of a whole-cell adherence assay with voided uroepithelial cells instead of erythrocytes, and discovered that pyelonephritis isolates of UPEC adhered to cells in greater numbers than did cystitis or fecal isolates. The adherence phenotype of those isolates was subsequently correlated with expression of fimbriae as assessed by transmission electron microscopy and hemagglutination [55]. These structures were ultimately designated P fimbriae because they adhered to the P

blood group antigen, common among most humans. Preparations of purified P fimbriae selectively adhered to uroepithelial cells [71] and mediated mannose-resistant hemagglutination (MRHA) of human type O erythrocytes [72]. Using erythrocytes, Svenson *et al.* [73] identified the receptor for P fimbria as the α–Gal(1→4)β–Gal moiety of the glycosphingolipid component of the P blood group antigen, a glycosphingolipid anchored in the cell membrane of erythrocytes and uroepithelial cells. Indeed, the free disaccharide is capable of inhibiting P fimbria from binding to its receptor on both erythrocytes [57] and uroepithelial cells [58]. In 1985, Uhlin *et al.* demonstrated that it was specifically the adhesive tip of P fimbria, PapG, that mediates binding to its substrate [74]. Parallel findings have since been repeated for other adherence systems, where only the terminal tip of a fimbria provides substrate binding specificity. The presence of P1 and P2 phenotypes of the P blood group antigen, which bind pyelonephritogenic *E. coli* [58], is elevated in women prone to UTI [75, 76] compared to non-UTI prone women, suggesting a causal link between carrier state of P blood group antigen and development of UTI.

Since its initial discovery, the association of P fimbria with virulent uropathogenic strains has been confirmed by many investigators [61, 67, 72, 77-86]. In a meta-analysis of prevalence studies, Welch and Donnenberg reported that UPEC isolates from otherwise healthy patients with pyelonephritis have a six-fold elevated relative risk of encoding P fimbriae compared to *E. coli* fecal isolates, which translates to roughly 80% of pyelonephritogenic UPEC strains (compared to a mere 16% of fecal/commensal *E. coli* isolates). In humans, expression of P fimbriae by UPEC in the urine [87] and

Figure 1-2. **The *pap* operon and flagellar synthesis cascade**. The *pap* operon encodes 9 genes necessary for synthesis of intact, functional fimbriae (*papAHCDJKEFG*). Fimbaie are assembled via the chaperone-usher pathway, and many thousands of units of PapA oligomerize to form the shaft of the fimbria. Several hundreds of units of PapE are incorporated near the tip, and PapG provides substrate specificity of binding, and is essential for the ability of P fimbriae to adhere. The transcript for *papX* is encoded on the the *pheV*- but not *pheU*-associated *pap* operon in *E. coli* CFT073. PapX is a member of the 17kDa protein family and represses transcription of *flhD* and *flhC* by binding to the *flhDC* promoter. *flhD* and *flhC* encode the master regulators of motility in *E. coli*, and their transcription and activation initiates a cascade that results in the synthesis of functional flagella. Many factors are known to bind the *flhD* promoter to regulate flagellar synthesis and thus to control motility. PapX may serve to communicate between adherence and motility so that they may be reciprocally controlled in *E. coli* CFT073.


Gal-Gal
Host receptor
P fimbria
OM
IM
pheV-associated pap operon
papI
papB
papA
papH
papC
papD
papJ
papK
papE
papF
papG
papX
Major subunit
Usher
Chaperone
Tip adhesin
Transcription factor
PapX
H-NS
HdfR
DnaA
CAP-cAMP
QseBC
OmpR
DnaK
DnaJ
GrpE
flhD
flhC
σ70
Class II fliA
σ28
σ28
Class III fliC
IM
OM
flagellum

antibodies to P fimbriae in the serum [88] suggest that *E. coli* produce and display P fimbria *in vivo* in infected patients.

In 1981, Hull *et al*. [89] isolated a cosmid clone carrying the genes that encode P fimbria. Subcloning revealed that a minimal operon of 11 genes was necessary and sufficient to generate functional P fimbriae. The *pap* (pyelonephritis-associated pilus) operon (**Figure 1-2)** was crucial in the subsequent discovery and characterization of the chaperone-usher pathway for assembly of fimbria elucidated by Normark and Hultgren, a pathway of fimbrial biogenesis that is broadly applicable to a wide range of enterobacterial fimbriae [66, 74, 90-96] **(Figure 1-2)**. Thus, P fimbriae represented not only a UPEC-specific virulence determinant but also served as the cornerstone for understanding adherence in the *Enterobacteriaciae*. P fimbriae consist of a composite fiber coupled to a tip adhesin. P fimbriae fibers are primarily composed of PapA subunits, and the adhesive tip fibrillum is composed of repeated PapE subunits that terminate in PapG, the component of P fimbria responsible for specific binding to its receptor [92, 97]. P fimbriae bind the α-D-Gal-(1→4)-β-D-Gal digalactoside moiety of the P blood group antigen. Indeed, the free disaccharide can inhibit P fimbria binding to its receptor on both erythrocytes [57] and uroepithelial cells [58]. Binding studies of PapG identified *p*-methoxyphenyl galabioside [98] and heavily modified galactopyranoside derivatives [99] as inhibitors of substrate binding for P fimbria. Structural studies of PapG in complex with receptor homologs have refined and supported these findings [100]. Another *E. coli* fimbria, encoded by the *prs* (*pap* related sequence) operon, is serologically identical to P fimbriae, but the two do not share virulence characteristics as the tip adhesin, PrsG, is antigentically distinct from PapG and binds the Forssman antigen

at GalNAc-α-(1-3)-GalNAc moieties [101]. The comparison between *prs* and *pap* fimbriae highlights the essential nature of specificity in adherence as it relates to pathogenesis. That is, adherence factors with highly similar or even identical nucleotide sequences may have vastly different functional effects on bacterial morphology, phenotype, and urovirulence, and especially careful attention must be paid to the tip adhesin of an adherence factor. The functional significance of sequence similarity must always be investigated experimentally. Conversely, dissimilar sequences may exhibit functional complementation.

Expression of the *pap* operon is phase variable and regulated by a Dam-methylation system [102, 103]. PapB and PapI act in concert with Lrp to control the methylation state and binding of two GATC sites, GATC1128 and GATC1130, located in the *pap* regulator region located between *papI* and *papB* [104]. When GATC1128 is methylated, the operon is in the OFF state, whereas methylation of GATC1130 is required for the ON state. PapB also appears to repress expression of Type 1 fimbriae, and deletion of *pap*-like gene clusters (including *sfa*) in *E. coli* strain 536 increased *fim* expression [105]. Several environmental conditions alter *pap* expression. P fimbrial expression *in vivo* in mice is down-regulated as determined by microarray analysis when compared to static growth in LB [38]. However, *pap* expression does not appear to be up-regulated *in vivo* in humans [106]. In addition, the *pap* operon expressed in *E. coli* strain K12 is primarily phase-off in static liquid culture compared to growth on plates [107]. In contrast, Type 1 fimbrial expression is induced in static liquid broth culture compared to growth on plates [68].

In animal models, demonstrating the direct contribution of P fimbria to infection has proven elusive. Two P-fimbrial isogenic deletion mutants of clinical isolates have been tested for virulence, in CBA/J mice [108] and in cynomologus monkeys [109, 110]. In work done by Mobley *et al.*, [108] CBA/J mice were challenged with strain CFT073 or its P fimbria-negative mutant (both *pap* operons were disrupted by allelic exchange). At the end of a week-long infection using a wide range of infectious doses, no significant difference in pathogen concentration or histological findings were found between the parent and mutant strains in urine, bladder or kidney. Due to technical limitations at the time of the experiment, only discrete, cross-sectional data are available for those infections.

Continuous data ware recently collected in a study by Melican *et al.* [111] which used GFP-expressing *E. coli* and multiphoton-based live-animal imaging during ascending and descending UTI in a mouse model of infection. The purpose of the study was to identify phenotypes of P fimbrial and Type 1 fimbrial mutants during infection of kidney tubules and glomeruli. A *pap* knockout strain of CFT073 caused tubular infection in the kidney at one-third the rate of the parental strain. They also observed that *E. coli* strain K12 transformed with a plasmid encoding the *pap* operon achieved visible growth more rapidly in the kidney tubules than did a strain containing vector [112]. Together these studies suggested that P fimbriae play a subtle role in the pathogenesis of pyelonephritis in the CBA/J mouse model of UTI. The complex role of P fimbria in urovirulence has been supported in later work in which P fimbrial knockouts did not affect urovirulence in a Type 1 fimbria-negative background in UPEC [113]. It was found, however, that synthesis of the P fimbria-related F1C fimbriae was up-regulated in

P fimbrial knockouts, hinting that some complementation or functional compensation between fimbrial organelles may be possible. Roberts *et al.* [109, 110] studied the effects of a *papG* mutation on colonization of the bladder and development of pyelonephritis in a primate model. No difference was observed by analysis of the primate urine (which gauged leukocyte esterase and analysis of bacteria/ml, by culture, in the urine of infected monkeys) between parent and mutant strain after bladder inoculation, although it was found that the parent strain persisted longer than the mutant strain in the urinary tract, and only animals infected with the parent strain showed evidence of impaired renal function and decreased renal mass. However, in both studies the mutant strain was still capable of infecting the kidneys. Together with epidemiological data, these studies suggest that while P fimbriae are associated with and probably contribute to virulence in the urinary tract, the murine model may not be the optimal setting in which to test the role of P fimbria in urovirulence.

Nearly all *E. coli* strains including UPEC strains express the mannose-sensitive adhesin Type 1 fimbria [114, 115]. Type 1 fimbria has been well established as a virulence factor in the mouse model of UTI [51-54]. Type 1 fimbriae are encoded by a cluster of nine genes, *fimBEAICDFGH*, which are expressed from at least three transcripts [89, 116, 117]. One transcript encodes the main structural component of Type 1 fimbria, FimA. Two transcripts encode FimH, the tip adhesin for Type 1 fimbria, and accessory proteins for assembly of the fimbrial stalk [118]. FimA forms the principal component of the fimbrial shaft [64] serving as the chief structural subunit of Type 1 fimbriae. FimG and FimF form a complex with FimH to produce the tip adhesin, with FimH providing receptor binding specificity to glycolipids and glycoproteins containing

mannosides, such as uroplakin Ia. This receptor lines the interface between the luminal urothelium and urine within the urinary tract [119-128]. The binding and adherence function of Type 1 fimbriae, mediated by FimH, appears to be an essential factor in urovirulence. Point mutations altering the binding characteristics of FimH impacted urovirulence of select UPEC strains [129]. In addition, vaccine studies using FimH as the primary antigen demonstrated protection against UTI in mice [62] as well as in cynomologus monkeys [63].

Epidemiological data suggest that the adherence factors S and F1C fimbriae are associated with urovirulent strains. S and F1C fimbriae are nearly identical in their biogenesis, structure, and sequence, but possess distinct adhesive tips and thus differ in their substrate specificity. Pooled data suggest that F1C fimbriae are more common among pyelonephritis and cystitis isolates than among fecal strains of *E. coli* [130]. The presence of F1C fimbriae is especially common in O6-antigen serotypes of UPEC, which themselves are associated with pyelonephritis isolates and UPEC strains that encode P fimbriae and hemolysin [131-133]. Interestingly, FocB, a regulatory protein of the operon encoding F1C fimbriae, also appears capable of regulating both Pap and Type 1 fimbriae, indicating that F1C is involved in a complex genetic regulatory network with other fimbriae directly tied to urovirulence [134]. F1C fimbrial expression has also been observed to be elevated when Type 1 fimbriae and P fimbriae are deleted from the chromosome in UPEC, supporting the notion of coordinated regulation [113]. Mechanistically, F1C fimbriae could contribute to urovirulence by facilitating adherence to human distal tubular and collecting duct epithelial cells and vascular endothelial cells in kidney cross-sections by histology [135], and expression of F1C fimbriae *in vivo* has

been demonstrated using monoclonal antibodies to F1C in the urine of patients with UPEC UTI [136]. Using force-measuring optical tweezers, F1C fimbriae were shown microscopically to have a Young's modulus (stretch-to-strain ratio) similar to that of P fimbriae and Type 1 fimbriae, a range which has been proposed to be important in the urinary tract to suitably resist flushing by urine flow [137]. Though F1C fimbriae do not agglutinate erythrocytes, they bind to galactosylceramide and globotriaosylceramide [138], and this binding can be inhibited by GalNAcβ1→4Gal and GalNAcb1→4β1→4Glc [139]. The observed difference between whole-cell binding and substrate affinity may offer a clue to the mechanism by which F1C fimbriae bind to human epithelial and endothelial cells.

Regulation of flagellar synthesis

The pathogenesis of most UTIs proceeds via an ascending route [5]. Bacteria gain access to the urethra and ascend to the bladder and then to the kidneys. In *E. coli*, motility is primarily dependent on flagella [48] (**Figure 1-2**). Flagella are complex membrane-bound extracellular organelles that consist of a basal body, a hook, and a filament. Synthesis proceeds in an "inside-out" manner, so that the basal body is constructed first, then the hook, and finally the filament [140]. Synthesis of flagella proceeds in a highly ordered hierarchical manner and is dependent on proper coordination of three classes of genes: class 1, class 2 and class 3 **(Figure 1-2)** [141-146]. Class 1 genes include primarily transcription of *flhD* and *flhC*, whose gene products form the $FlhD_4C_2$ complex that acts as a transcription factor to activate transcription of class 2 genes. *flhD* and *flhC*

are located on a single operon and their expression is highly regulated by numerous factors [147]. Several transcription factors regulate *flhD* expression by binding to the *flhDC* promoter, such as DnaA [148], the CAP-cAMP complex [149], OmpR (which binds in a phosphorylation-dependent manner) [150], and the two-component QseBC system [151]. Other factors regulate transcription in more subtle ways, such as H-NS which binds the *flhDC* promoter to alter DNA topology, reversing the binding of HdfR, a transcriptional repressor of *flhD* [152]. Still other regulators act post-translationally. DnaK, DnaJ, and GrpE appear to act together to enhance the transcription-factor activity of $FlhD_4C_2$ by enhancing functional folding [153, 154].

Thus *flhD* and *flhC* act as a central point of regulation of motility for many systems and pathways, leading to their designation as the master regulators of motility. The multi-faceted regulation of the class 1 genes equips bacteria to regulate motility in response to a host of environmental factors. Subsequently activated class 2 gene products are responsible for assembly of the basal body and hook, and expression of FliA (sigma 28) and FlgM (anti-sigma 28) [142]. Once the basal body and hook structures are assembled, FlgM diffuses out of the bacterium [155], allowing activation of FliA. Active FliA is required for the transcription of the class 3 genes [156, 157]. Class 3 genes encode hook-associated proteins including FliC (flagellin), required to complete construction of functional flagella, and motility-associated proteins MotA, MotB, CheW and CheY [142].

Bacterial dependence on flagella-mediated motility in the pathogenesis of UTI has been illustrated in multiple experimental studies. UPEC *fliC* mutants are defective in motility but able to cause UTI in independent challenge in mice. However, the *fliC*

mutant was at a disadvantage relative to the motile parental strains in competition assays [49, 50]. In another study, non-motile mutants were highly attenuated for colonization of the kidneys even in independent challenge [59]. Interestingly, transcriptomic data from UPEC collected from the urine of mice with experimental UTI showed low levels of flagellar expression [38], suggesting that overall flagellar expression is repressed in the general population of *E. coli* infecting the urinary tract. Flagellar repression during UTI was confirmed in a study of human UTI, in which UPEC isolates collected directly from the urine of human subjects were subjected to transcriptional profiling by microarray. In general, the ranking of motility-related genes trended toward a lower-level of expression during human UTI [106]. Low ranking of motility genes may occur because flagellar expression is highly coordinated during UTI depending on location within the urinary tract, as has been shown by biophotonic imaging studies [59], or because UPEC shed in urine are not highly flagellated. Flagella are highly immunogenic in pathogenic *E. coli*, and are recognized by toll-like receptor 5 and induce IL-8 expression [158]. This may help to explain why UPEC tightly regulates expression of flagella to prevent excessive immune activation.

Reciprocal control

Ascension of the urinary tract by UPEC is an essential component of uropathogenesis to establish an infection of the bladder or kidneys during UTI. To ascend, however, bacteria must use adherence to resist being expelled by the flushing mechanism of voiding, while simultaneously engaging in upward motility. Motility, mediated by flagella, and adherence, mediated by fimbriae, are inherently antagonistic

forces. The immobilizing effect of fimbriae would limit the motility of flagella, and increased motility by flagella would reduce the ability of bacteria to adhere at one site, yet both motility and adherence contribute significantly to the development of UTI. Thus, it is logical that bacteria would have in place a mechanism to reciprocally control motility and adherence.

Hints at a mechanism of reciprocal control of motility by adherence arose in studies of a UPEC-related gram-negative enteric uropathogen, *Proteus mirabilis* strain HI4320. This strain encodes the mannose-resistant *Proteus*-like (MR/P) fimbria, which is a typical chaperone-usher fimbria similar in structure to P fimbria of *E. coli*. Indeed, MrpA shares homology with PapA, the major structural subunit of P fimbria [159]. Phenotypically, MR/P fimbriae also agglutinate erythrocytes in a mannose-resistant manner [159-161], although the precise receptor has not been identified. Epidemiologically, *P. mirabilis* strains that cause pyelonephritis tend to be enriched for MR/P fimbriae compared to cystitis isolates and controls [162]. MR/P fimbriae have been shown genetically to contribute to virulence in a murine model of UTI but are not required to cause UTI [161, 163, 164]. Like P fimbriae, MR/P fimbriae induce a strong immune response [159], and recent vaccine trials indicate that either MrpH or MrpA is protective against UTI in the mouse model of infection [165-167].

The last gene of the MR/P fimbrial operon is *mrpJ*, which encodes a transcription factor predicted to employ a winged-helix DNA-binding motif. MrpJ represses swimming and swarming motility by binding to the *flhD* promoter and repressing transcription of the flagellar master regulator, as shown by expression studies and electrophoretic mobility shift assays [168]. *P. mirabilis* HI4320 has 14 paralogues of

mrpJ, 12 of which also repress motility and 10 of which are encoded at the end of fimbrial-encoding operons [168]. Studies in UPEC demonstrated that PapX, a non-structural 183 amino-acid protein of the highly conserved 17 kDa family, is encoded as the last member of the PAI-CFT073$_{pheV}$ but not PAI-CFT073$_{pheU}$ –associated *pap* operon, down-regulates motility by repressing the expression of flagella [169]. Interestingly, there is up to 45% amino-acid identity between PapX and MrpJ in some of their domains, and *papX* functionally complements swimming and swarming phenotypes in isogenic *mrpJ* deletion mutants of *P. mirabilis* [170]. PapX also represses swimming motility in *E. coli* strain CFT073 (121). Both *papX* and *mrpJ* encode non-structural components of their respective fimbriae and are not required for fimbrial synthesis. Direct evidence in UPEC exists that PapX plays an integral role in reciprocal control of motility by adherence. In a series of studies investigating motility under conditions that affect adherence in *E. coli* strain CFT073, when the adherence factor Type 1 fimbria was constitutively expressed, motility was reduced. However, this repression of motility was partially reversed in *papX* deletion mutants [60, 169, 171]. Additionally, microarray and real-time PCR data indicate that the effect is mediated by repression of expression of *flhD* and *flhC*. Together with the homology to, and parallels with, *mrpJ*, these findings strongly support the model that *papX* encodes a transcription factor that plays an integral role in the regulation of motility by adherence in UPEC.

Methods for assessment

Transcription factors and their binding sites may be studied by a three-pronged approach: predictors of function, transcriptional profiling, and direct binding assays. Predictors of function are often dependent on associative measures and computations. For example, structural prediction and homology modeling software such as PHYRE [172] and I-TASSER [173, 174] can predict the behavior of a protein based on sequence similarity and integrated predictors of protein folding. Putative annotations (*i.e.,* in Pubmed or Genbank) serve much the same purpose. PapX is predicted to belong to the MarR-family of transcription factors, dimeric DNA binding proteins that alter downstream transcription. MarR family members typically bind to DNA using a winged-helix motif, in which a turn region flanked by short beta-pleated sheets ("wings") and a helix interact non-covalently with the major and minor grooves of the DNA helix to produce binding [175-179]. It varies by family member whether or not this interaction is dependent on cofactors, and whether the helix or the loop provides the sequence-specificity of binding. Frequently, though not universally, MarR-family proteins act near the -10 and -35 sequences of promoters to sterically regulate gene expression by altering DNA helix topology, typically binding a sequence motif between 21 and 45 bases in length [180].

Transcriptional profiling broadly describes techniques that uncover factor-dependent changes in transcription. This includes microarray, qPCR, and promoter-reporter fusions. The weakness of transcriptional profiling techniques is that they are heavily inductive. That is, limited inferences about the mechanism of transcriptional effect may be drawn when only transcriptional data are available, because a transcriptional effect may be the result of intermediate or dependent events. For this

reason, identification of a protein as a transcription factor typically requires direct evidence of protein:DNA interaction. This approach includes techniques in which the output is influenced by or dependent on the factor interacting persistently and meaningfully with fragments of DNA of specific sequence. Examples of this type of approach are DNase footprinting assays, electrophoretic motility shift assays (EMSA), co-precipitation techniques (such as ChIP-seq [181] and protein-binding arrays [182]), computationally intensive techniques such as SELEX and SELEX-seq [183-189], and screening techniques such as bacterial 1-hybrid systems [190]. In each of these techniques, proteins of interests are exposed to DNA fragments and changes in species behavior are used to gauge interactions.

Currently, programs exist to analyze high-throughput data from SELEX-seq, ChIP-seq and protein-binding arrays [186, 191]. All transcription factor binding-site prediction software requires substantially advanced knowledge of personal computing and tends to make heavy use of the command line or intermediary scripting languages and programs such as R or Matlab. The output of these techniques is frequently a list of sequences that were enriched against a target protein, ranging from a few sequences to thousands. Downstream analysis of these sequences can reveal binding motifs. The MEME suite of programs [192], for instance, can generate positional-weight matrices from a list of sequences. Positional-weight matrices are generated by aligning all similar sequences and quantifying how conserved a given nucleic acid is at any site within the sequence. These can be extremely useful for predicting likely binding sites and for directing downstream experiments to verify a candidate site.

Each of these techniques can be used to discover transcription factors, to determine whether or not a given protein is a transcription factor, and to identify the most likely site of interaction between a transcription factor and a genome. However, none of the techniques mentioned can identify the specific mechanism of protein:DNA interaction, though computational modeling may be useful for generating hypotheses. To explore mechanisms of DNA binding by proteins, biochemical and genetic techniques are essential, and a working knowledge of transcription-factor motifs and historical precedent for inferred mechanism are extremely useful. Biochemical and genetic techniques include making select mutations to a gene encoding the binding protein and assaying for binding capacity, solving the structure of the protein in the presence and absence of its DNA ligand (such as with crystallography or NMR), or activity assays under variable conditions. Currently, no technique allows an investigator to scrutinize DNA binding events and effects in real-time, so even direct techniques are inductive, though they are considered to be more reliable measures of protein function.

Statement of the problem

Although it was known that PapX regulates motility in UPEC, the precise mechanism of this control was unknown. There is a need to better understand the nuances of reciprocal control of adherence and motility in UPEC, and we reasoned that the study of PapX would be an excellent and useful lens through which to scrutinize the phenomenon.

Hypothesis 1: PapX is a transcription factor that regulates motility by repressing master regulator genes *flhD* and *flhC*. Determining the mechanism of action and binding site of this protein will shed light on a broad class of UPEC transcription factors.

Previous work by EMSA and transcriptional profiling has shown that PapX binds somewhere within the *flhD* promoter and that *flhD* undergoes a *papX*-dependent repression of expression. This led us to formulate the model that PapX is a transcription factor that represses motility in UPEC by repressing *flhD* expression. We wanted to elucidate the precise mechanism by which PapX exerts its effect, including investigation of the specific target binding site sequence and biochemical mechanism. Prior work with the functional PapX homolog, MrpJ, had shown that using linker-insertional mutagenesis could reveal active domains of the protein [168]. We reasoned that subjecting PapX to a similar approach would identify domains necessary for full function of the protein, and would lend insight into the mechanism of action and class of the PapX protein. We also reasoned that, as PapX binds to the *flhD* promoter, we could identify the specific sequence of interaction using unbiased techniques. Because PapX belongs to a broadly conserved family of proteins within UPEC, we anticipated that our work would shed light on the mechanism of control of motility across many uropathogens. We hypothesized that this information would give us crucial insight into the specific mechanism of reciprocal control in UPEC.

Hypothesis 2: *papX* is transcriptionally regulated by the *pap* operon

We were interested in the mechanism by which *papX* and its family members are regulated in UPEC. Prior work had demonstrated that *papX* is encoded on the *pap* operon [169], and that it is involved in reciprocal control of adherence and motility [60]. This led us to the simple model that PapX activity should be coeval (*i.e.*, co-regulated) with the transcriptional activity of the *pap* operon, so that when expression of the adherence factor P fimbria is increased, *papX* expression is increased and thus the activity of PapX is increased in the bacterium. This would provide a simple model of reciprocal control and an explanation of how adherence may reciprocally repress motility in UPEC. When the *pap* operon is phase off, both P fimbria (the adherence factor) and PapX are not synthesized. Flagellar synthesis is not repressed, flagella are assembled and the bacterium is motile. When the *pap* operon is phase on, both P fimbria and PapX are synthesized. Flagellar synthesis is repressed and the bacteria are not motile. Thus, the bacterium is either motile or adherent, but not both.

By testing these hypotheses, I will contribute to our understanding of the role of PapX in reciprocal control, the mechanism by which PapX suppresses motility, and the mechanism by which PapX activity is regulated. Because PapX and its homologs appear in many uropathogens, these studies will help to illuminate the pathways of reciprocal control of adherence and motility across a broad range of bacteria. The knowledge gained in this way could inform methods for disruption of this important pathway.

CHAPTER 2

DETERMINATION OF THE TARGET SEQUENCE BOUND BY PAPX, A REPRESSOR OF BACTERIAL MOTILITY, IN THE *FLHD* PROMOTER USING SELEX AND HIGH-THROUGHPUT SEQUENCING[*]

[*]**Published in modified form in the Journal of Biological Chemistry, 286(52):44726-38 (2011)**

Introduction

papX represses motility by inhibiting expression of *flhD* [169], but the mechanism remains unknown. We reasoned that a multi-pronged approach to studying *papX* action would yield productive insight into the structure, function and the specific nature of its interaction with *flhD*. We first wanted to determine whether the mechanism of *papX* action was virulence-specific, or more broadly effective. To examine this question, we carried out studies of the effect of *papX* expression on pathogenic and non-pathogenic strains of *E. coli*, reasoning that a difference in effect would suggest a specific relationship to virulence while a lack of difference would suggest that the role of *papX* may be broad and not limited to only uropathogenesis. We supported the results of these studies using molecular epidemiology to analyze the distribution of *papX* within a large *E. coli* strain collection. We hypothesized that if *papX* was found more commonly in strains of increased urovirulence as compared to strains of diminished urovirulence, such

as commensal strains, that would be suggestive of a specific role in uropathogenesis for *papX*.

To examine PapX mechanism directly, we designed molecular-genetic studies to uncover regions of PapX crucial for function. Homology modeling and mutagenesis approaches in the study of *mrpJ* have been informative for understanding the mechanism of its action [168], and because *papX* is a functional homolog of *mrpJ*, we concluded that this approach could be similarly useful in uncovering key regions and residues of the PapX protein. To further our understanding of the interaction between PapX and *flhDC* expression, we used systematic evolution of ligands by exponential enrichment (SELEX) in conjunction with high-throughput sequencing technology to identify a novel DNA binding site for PapX within the *flhDC* promoter. To the best of our knowledge, this is the first description of aptamer-free SELEX in conjunction with high-throughput sequencing technology in bacteria to study DNA binding proteins, and the first time these techniques have been used to identify a novel binding site instead of to confirm or refine known binding sites. Our application of SELEX with high-throughput sequencing provides for novel advantages over previous techniques for identifying bacterial transcription-factor binding sites, and may serve as the paradigm for future high-throughput screens. Because the 17 kDa family of genes is highly conserved, by elucidating the mechanisms underlying reciprocal control of motility by PapX in UPEC, we will be better equipped to manipulate and disrupt this important regulatory cascade and virulence property of a pathogenic microbe.

Results

PapX has homologs in Extraintestinal Pathogenic *E. coli* (ExPEC)

To assess the distribution and conservation of PapX, we compared the amino acid sequence of PapX homologs using Clustal-W. PapX has six closely related homologs of the 17 kDa protein family among strains of UPEC and meningitis-associated *E. coli*, with over 93% amino acid identity between them (**Figure 2-1**). Three of the ten non-identical residues are substitutions between Ala and Val or Ile and Val. Four non-identical sites contain either Ala or Thr. These features result in pairs of PapX homologs sharing up to 99% amino acid sequence identity.

Prevalence of PapX homologs among UPEC

To assess the prevalence of *papX* in UPEC strains, a collection of 294 *E. coli* strains representing a range of isolates from fecal/commensal strains (**n=88**) to strains isolated from clinical cases of asymptomatic bacteriuria (ABU) (**n=54**), complicated UTI (**n=39**), uncomplicated cystitis (**n=37**), and acute pyelonephritis (**n=76**), was examined by multiplex PCR for the presence of *papX* homologs. Primers were designed to unique flanking sequences of *papX* and *focX*, and to common regions of both to amplify fragments of specific size (**Figure 2-2A, Table 2-1**). Within this collection, 53% of pyelonephritogenic strains of *E. coli* contained at least one copy of *papX*, the highest prevalence within the strain collection and significantly more prevalent than the 32% of

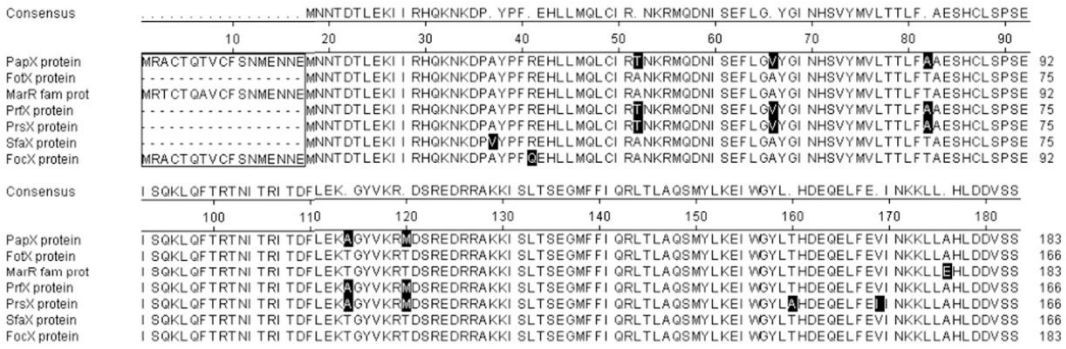


Figure 2-1. Alignment of PapX homologs in the 17 kDa family. PapX homologs of the 17 kDa protein family were aligned using Clustal-W using Megalign. Perfect consensus is shown above. Positions with imperfect amino acid identity between homologs are shown as ".". Amino acid residues that represented the minority for that position are shaded in black within each homolog's amino acid sequence. The 7 family members are ≥ 93% identical to one another at the amino-acid level past residue 17. N-terminal start-site annotation in homologs (boxed) is the result of *in silico* analysis and has not been experimentally verified (using *i.e.* primer extension).

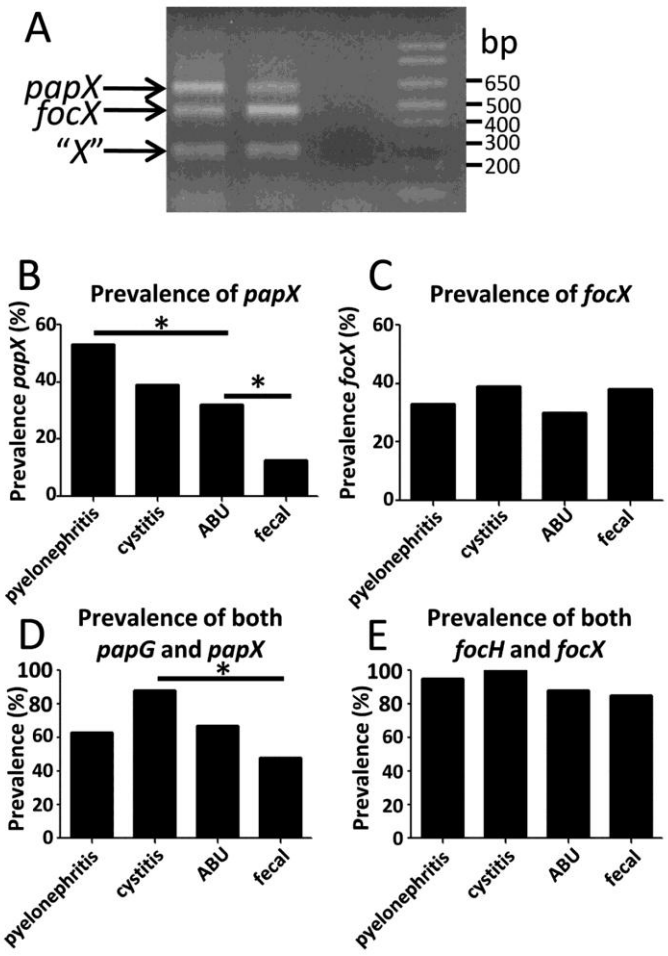


Figure 2-2. Multiplex PCR to analyze a 294 member library of UPEC and fecal/commensal isolates for the presence of *papX*, *focX* and non-specific 17 kDa family genes. (A) Three primer sets were designed to amplify a 607 bp fragment of *papX* and flanking region, a 449 bp fragment of *focX* and flanking region, and a 249 bp region common to all *papX* homologs. All primer sets shared a common reverse primer (**Table 2-1,** *papX*MltplxRev) with a unique forward primer (**Table 2-1,** *papX*MltPlxfwd, *focX*MltPlxfwd, *X*MltPlxfwd). *papX* and *focX* contain nearly identical nucleotide sequence, so unique primers for each originated in the sequence flanking each gene. Fragments were sequenced to confirm their composition. Each fragment was generated from extracted, pooled CFT073 WT genomic DNA and separated on 2% agarose gel. The prevalence of **(B)** *papX* and **(C)** *focX* within the library was scored by the appearance of their respective amplicon during multiplex PCR within each strain type**.** The percentage of **(D)** *papX*- or **(E)** *focX*-containing strains within *papG*- or *focH*-containing strains, respectively, is shown.

Table 2-1. Primers used in this study.

Primer	Sequence (5'→ 3')	Function
*papX*H1P1	...TGCCTGAAGCTATCCGGCATACTCAGGCATTTCACGCTTTGTGTAGGCTGGAGCTGCTTC	λ red mutants
*papX*H2P2	..AGAAATCAGGCAGGAATAAATTGTAGTGGAAAGTCGATGTATGGGAATTAGCCATGGTCC	λ red mutants
*lacZ*H1P1	……TTACGCGAAATACGGGCAGACATAGCCTGCCCGGTTATTA	λ red mutants
*lacZ*H2P2	……ATGACTATGATTACGGATTCTCTGGCCGTCGTATTACAAC	λ red mutants
*focX*H1P1	……ATGCGCGCTTGTACACAGACAGTGTGTTTCAGTAAGTGTAGGCTGGAGCTGCTTC	λ red mutants
*focX*H2P2	……AATGACTTCAAACAGTTCCTGTTCATCATGGGTCAATGGGAATTAGCCATGGTCC	λ red mutants
gapA 29bp fwd	……ACTCCACTCACGGCCGTTTCGACGGTACCG	Gel shift
gapA 29bp rev	……CGGTACCGTCGAAACGGCCGTGAGTGGAGT	Gel shift
papX 29bp fwd	……TTACGGTGAGTTATTTTAACTGTGCGCAAC	Gel shift
papX 29bp rev	……GTTGCGCACAGTTAAAATAACTCACCGTAA	Gel shift
EMSA154_F	……TCGCTACACTGGCAAGCAAA	Gel shift
EMSA154_R	……GTTGTGCGGTAAGTGTTTGT	Gel shift
*papX*RTfwd	……CGCATATCCCTTTCGGGAGCATTT	qPCR
*papX*RTrev	……TCCGCTGCGAATAATGTGGTGAGA	qPCR
*gapA*RTfwd	……CGTTAAAGGCGCTAACTTCG	qPCR
*gapA*RTrev	……ACGGTGGTCATCAGACCTTC	qPCR
*flhD*RTfwd	……TCCGCTATGTTTCGTCTCGGCATA	qPCR
*flhD*RTrev	……ACCAGTTGATTGGTTTCTGCCAGC	qPCR
*flhC*RTfwd	……AAACTGGCTTGTGTAATGGCGTCG	qPCR
*flhC*RTrev	……TCAACAAACCGCACCAATGTCCAG	qPCR
T^{21}Afwd	….GTAATATGGAGAATAATGAAATGAATAACGCAGACACATTAGAAAAAAATAATCAGACAC	Site-directed mutant
T^{21}Arev	……CATTATACCTCTTATTACTTTACTTATTGCGTCTGTGTAATCTTTTTTATTAGTCTGTG	Site-directed mutant
R^{51}Afwd	……GGAGCATTTGTTGATGCAACTCTGTATTGCCACAAATAAAAGAATGCAG	Site-directed mutant
R^{51}Arev	……CTGCATTCTTTTATTTGTGGCAATACAGAGTTGCATCAACAAATGCTCC	Site-directed mutant
K^{54}Nfwd	……AGATATATTATCCTGCATTCTGTTATTTGTGCGAATACAGAGTTGCA	Site-directed mutant
K^{54}Nrev	……TGCAACTCTGTATTCGCACAAATAACAGAATGCAGGATAATATATCT	Site-directed mutant
R^{55}Gfwd	……AGATATATTATCCTGCATTCCTTTATTTGTGCGAATACAGAGTTGC	Site-directed mutant
R^{55}Grev	……GCAACTCTGTATTCGCACAAATAAAGGAATGCAGGATAATATATCT	Site-directed mutant
H^{71}Afwd	……CTGAATTTCTGGGGGTGTATGGAATAAATGCCTCAGTATATATGGTTCT	Site-directed mutant
H^{71}Arev	……AGAACCATATATACTGAGGCATTTATTCCATACACCCCCAGAAATTCAG	Site-directed mutant
R^{101}Afwd	……TAAGCCAGAAACTTCAGTTTACCGCAACTAATATTACCCGCATTACAG	Site-directed mutant
R^{101}Arev	……CTGTAATGCGGGTAATATTAGTTGCGGTAAACTGAAGTTTCTGGCTTA	Site-directed mutant
R^{106}Afwd	……TCAGTTTACCAGAACTAATATTACCGCCATTACAGATTTTTTAGAAAAAGCC	Site-directed mutant
R^{106}Arev	……GGCTTTTTCTAAAAAATCTGTAATGGCGGTAATATTAGTTCTGGTAAACTGA	Site-directed mutant
K^{113}Nfwd	……TAATATTACCCGCATTACAGATTTTTTAGAAAACGCCGGATATGTAAAAAG	Site-directed mutant
K^{113}Nrev	……CTTTTTACATATCCGGCGTTTTCTAAAAAATCTGTAATGCGGGTAATATTA	Site-directed mutant
K^{118}Afwd	……TTACAGATTTTTTAGAAAAAGCCGGATATGTAGCAAGGATGGATAGCAGGG	Site-directed mutant
K^{118}Arev	……CCCTGCTATCCATCCTTGCTACATATCCGGCTTTTTCTAAAAAATCTGTAA	Site-directed mutant
R^{119}Afwd	……AGATTTTTTAGAAAAAGCCGGATATGTAAAAGCGATGGATAGCAGGGAGG	Site-directed mutant
R^{119}Arev	……CCTCCCTGCTATCCATCGCTTTTACATATCCGGCTTTTTCTAAAAAATCT	Site-directed mutant
R^{123}Afwd	……GTAAAAAGGATGGATAGCGCGGAGGATCGCCGTGCTAA	Site-directed mutant
R^{123}Afwd	……TTAGCACGGCGATCCTCCGCGCTATCCATCCTTTTTAC	Site-directed mutant
R^{126}Afwd	……AAAAAGGATGGATAGCAGGGAGGATGCCCGTGCTAAAAAA	Site-directed mutant
R^{126}Arev	……TTTTTTAGCACGGGCATCCTCCCTGCTATCCATCCTTTTT	Site-directed mutant
R^{127}Gfwd	……TAGCAGGGAGGATCGCGGTGCTAAAAAAATCAGT	Site-directed mutant
R^{127}Grev	……ATCGTCCCTCCTAGCGCCACGATTTTTTTAGTCA	Site-directed mutant
R^{127}Kfwd	……GATGGATAGCAGGGAGGATCGCAAGGCTAAAAAAATCAGTCTGACAT	Site-directed mutant
R^{127}Krev	……ATGTCAGACTGATTTTTTTAGCCTTGCGATCCTCCCTGCTATCCATC	Site-directed mutant
K^{129}Afwd	……CAGGGAGGATCGCCGTGCTGCAAAAAATCAGTCTGACATCTG	Site-directed mutant
K^{129}Arev	……CAGATGTCAGACTGATTTTTGCAGCACGGCGATCCTCCCTG	Site-directed mutant
K^{130}Afwd	……GGGAGGATCGCCGTGCTAAAGCAATCAGTCTGACATCTGAA	Site-directed mutant
K^{130}Arev	……TTCAGATGTCAGACTGATTGCTTTAGCACGGCGATCCTCCC	Site-directed mutant
*papX*MltplxRev	……TCCGCTGCGAATAATGTGGTGAGA	Multiplex PCR
XMltPlxfwd	……GCGCGCTTGTACACAGACAGTGTG	Multiplex PCR
*focX*MltPlxfwd	……CAGGCGTGTGGAAGCTGATG	Multiplex PCR
*papX*MltPlxfwd	……CCTGACCATCGGCAGTCGCCTC	Multiplex PCR
DNase*flhD*rev	……TCAGCAACTCGGAGGTATGC	DNase protect assay
DNase*flhD*Afwd	……ACCAAAAAGGTGGGTCTGCT	DNase protect assay
DNase*flhD*Bfwd	……AAAATCGCAGCCCCCCTCCG	DNase protect assay
DNase*flhD*Cfwd	……TTTGCTTGCCAGTGTAGCGA	DNase protect assay
DNase*flhD*Dfwd	……GTTGTGCGGTAAGTGTTTGT	DNase protect assay
DNase*gapA*fwd	……ACGTGACTGATTCTA	DNase protect assay
DNase*gapA*rev	……CAGCTATTTGTTAGTG	DNase protect assay
*flhD*EHrev	……TCAGCAACTCGGAGGTATGC	*lacZ*-fusions
*flhD*Afwd	……ACCAAAAAGGTGGGTCTGCT	*lacZ*-fusions
*flhD*Bfwd	……AAAATCGCAGCCCCCCTCCG	*lacZ*-fusions
*flhD*Cfwd	……TTTGCTTGCCAGTGTAGCGA	*lacZ*-fusions
*flhD*Dfwd	……GTTGTGCGGTAAGTGTTTGT	*lacZ*-fusions
*flhD*Efwd	……ATGTACTGATTTCCCATATT	*lacZ*-fusions
*flhD*Ffwd	……TGGAGAAACGACGCAATCCC	*lacZ*-fusions
pap_{GX}_Eco_fwd	……CCGGAATTCATGGTTTATCCGGAGCCGGAT	*lacZ*-fusions
Pap_{GX}_BamHI_rev	……CGCGGATCCCAGTCATCCGGTAAACATCGAC	*lacZ*-fusions
pap_{BA}_Eco_fwd	……CCGGAATTCCGCTTTATTTGTTCAATTTAGTGAATTTGC	*lacZ*-fusions
pap_{BA}_BamHI_rev	……CGCGGATCCGCGCCTTGTTTCCCCCTTCTGTC	*lacZ*-fusions

ABU strains (P = 0.0192). Only 12.5% of fecal/commensal *E. coli* contained at least one copy of *papX*, the lowest prevalence within the strain collection and significantly less prevalent than in ABU strains (P = 0.0081) (**Figure 2-2B**). *focX*, a close homolog of *papX*, did not significantly vary in prevalence within the UPEC strain collection (P = 0.7927) (**Figure 2-2C**). This was not surprising as *focX* is closely associated with genes encoding F1C fimbria, which appears more closely associated with meningitis-associated strains than with UTI-associated strains [193].

The association of the *pap* operon, which encodes P-fimbriae, with pyelonephritogenic strains of UPEC is well documented, with 77% of pyelonephritogenic isolates containing P fimbriae, as compared to only 23% of cystitis, 20% of ABU, and 16% of fecal/commensal isolates [73, 75]. 59% of all strains tested carrying at least one copy of *papG* also carried at least one copy of *papX*. Therefore, to differentiate the influence of the increased representation of the *pap* operon in UTI-causing UPEC from the role of *papX* alone, we quantified the prevalence of *papX* in strains containing *pap* operons for each type of strain within the collection, using the prevalence of *papG* as a proxy for the prevalence of the *pap* operon [194].

Within strains that contained both the *pap* operon and *papX*, the relative prevalence of *papX* compared to *papG* was significantly higher in cystitis-causing strains of *E. coli* as compared to fecal/commensal strains (P =0.0165) (**Figure 2-2D**). The prevalence of strains containing both *focX* and *focH*, a proxy for the presence of F1C fimbriae, was not significantly different between UTI-causing strains of *E. coli* and fecal/commensal strains of *E. coli* (P = 0.2222) (**Figure 2-2E**). These data indicate that the presence of *papX*, but not *focX*, is strongly associated with UPEC. Taken together

with the highly conserved nature of the protein sequence between PapX homologs, these data suggest that PapX contributes to UPEC virulence, but that encoding *papX* alone is not sufficient to convey virulence to *E. coli* strains.

Functional characterization of PapX

We previously demonstrated that PapX down-regulates motility [60, 169]. In strain CFT073, deletion of the single copy of *papX*, which resides on the PAI-CFT073$_{pheV}$ but not PAI-CFT073$_{pheU}$-associated *pap* operons [169], results in increased motility as compared to wild-type. On the contrary, over-expressing *papX* from the leaky, inducible plasmid pPxWT repressed motility. *papX* overexpression in CFT073 (pPxWT) has a specific transcriptional effect, with 38 of 42 differentially expressed genes being associated with motility and the downstream effects of *flhD* and *flhC* regulation, the master regulator of motility [169]**.** Coupling this finding with the discovery that *papX* appears to be strongly associated with uropathogens, we introduced *papX*, carried on pPxWT, into strain CFT073 and the prototypical non-pathogenic fecal/commensal *E. coli* K12 strain MG1655. Motility assays, performed on the resulting transformants, demonstrated that CFT073 had reduced motility in the presence of *papX* when compared to vector control (**Figure 2-3A)**, but that *papX* had no effect on the motility of the non-pathogenic K12 strain MG1655 (**Figure 2-3B**). This informed us that the role of *papX* is pathogen-specific.

To investigate structure-function relationships within PapX, we constructed a library of linker insertion mutations within *papX* in pPxWT. 15-nucleotide insertions

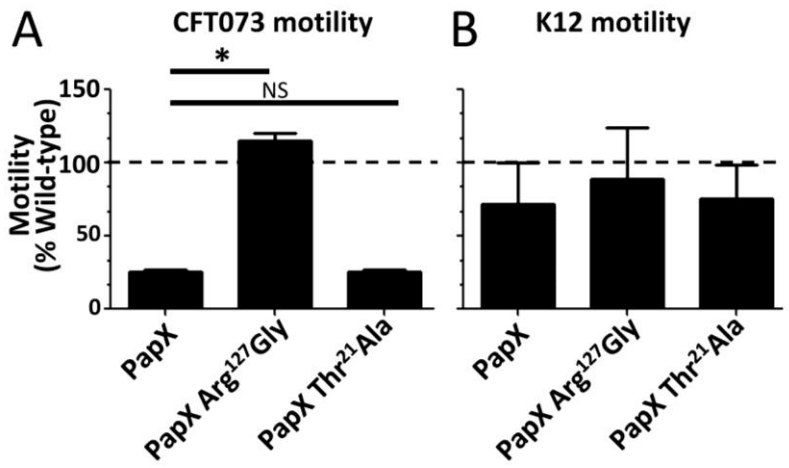


Figure 2-3. Effect of *papX* and *papX*_Arg^{127}Gly mutant on motility of UPEC and a non-pathogenic *E. coli.* *E. coli* CFT073 and K12 MG1655 were transformed with pPxWT (WT *papX*), pR127G (*papX* Arg^{127}Gly mutant**)** or pT21A (*papX* Thr^{21}Ala mutant). Constructs were then stabbed into soft-agar plates and incubated at 30^{o}C. Diameter of swimming motility was recorded after 16 hours. For each construct, assays were performed on four biological replicates in triplicate. **(A)** Compared to vector control (dashed line) in CFT073, pPxWT and pT21A reduced motility significantly, but pR127G did not. **(B)** None of the three constructs influenced motility significantly differently from vector control in K12 MG1655.

were randomly introduced along the coding sequence of *papX*. Thirty unique in-frame insertions were generated, corresponding to amino acid residues 15-162 of PapX. Each of the 30 insertional mutations was electroporated into CFT073 Δ*papX*. The growth rates of the transformants were not significantly different from CFT073 Δ*papX* complemented with either wild-type *papX* or vector control (**Supplementary Figure 2-1**). Motility conferred by each construct was measured in soft agar, side-by-side with motility assays of CFT073 Δ*papX* complemented with either pPxWT (encodes *papX*) or pVector (vector control) (**Figure 2-4A**). We found that mutations in two regions of *papX* (encoding residues 51-74 and 103-133) produced constructs no longer able to repress motility (**boxes, Figure 2-4A**). Insertions into either of these regions abolished PapX activity, resulting in motility similar to that observed for the CFT073 Δ*papX* construct containing vector alone. The motility results indicated that these two regions were critical for PapX function.

We reasoned that structural modeling would provide insight into the structure and function of this DNA binding protein. Indeed, structural models of PapX generated using I-TASSER [173, 174] and PHYRE [172] suggested that PapX is likely to be a member of the MarR-family of helix-wing-helix homodimeric transcriptional regulators (**Figure 2-4B).** Size-exclusion chromatography and semi-native SDS-PAGE confirmed that PapX is a dimer (**Supplementary Figure 2-2**). The MarR family of transcription factors tends to rely on specific positively charged residues, often Arg, to mediate DNA binding interactions [175-178, 180] and dimerization. When the predicted structure is compared to existing MarR family proteins [175, 176], the two regions of PapX sensitive to

Figure 2-4. Motility index of 15 bp insertional mutants of *papX*. (A) Thirty in-frame 15 bp insertions were made in *papX* in pPxWT. Constructs, pPxWT (WT *papX*) and pVector (empty vector) were electroporated into CFT073 Δ*papX*. Motility was assessed in soft agar and compared relative to CFT073 Δ*papX* electroporated with pPxWT. Vertical boxes indicate regions of loss of function. Top and bottom horizontal boxes represent motility of CFT073 Δ*papX* electroporated with pVector or with pPxWT, respectively. Presence of positively charged residues is indicated along the bottom of the figure. **(B)** I-TASSER software was used to generate a predicted structural model based on the amino acid sequence of PapX. Arrows indicate putative DNA binding (Arg^{127}) and dimerization (Lys^{54}) regions of the protein. **(C)** Site-directed mutants of *papX* were made in pPxWT and transformed into a CFT073 Δ*papX* background. Soft-agar motility assays were performed on the resulting constructs, performed in triplicate using biological triplicates. Boxes indicate regions sensitive to mutation determined in **(A)** by linker-insertional mutagenesis. Solid bars are not significantly different from vector control (dashed line). **(D)** Western blot showing pPxWT (WT PapX), pR127G (PapX Arg^{127}Gly mutant) and pK54N (PapX Lys^{54}Asn mutant) transformed into *E. coli* strain K12 either uninduced or induced with 300 μM IPTG. Bands migrate with the same electrophoretic mobility for mutants and wiltype PapX by semi-native SDS-PAGE (12% acrylamide).

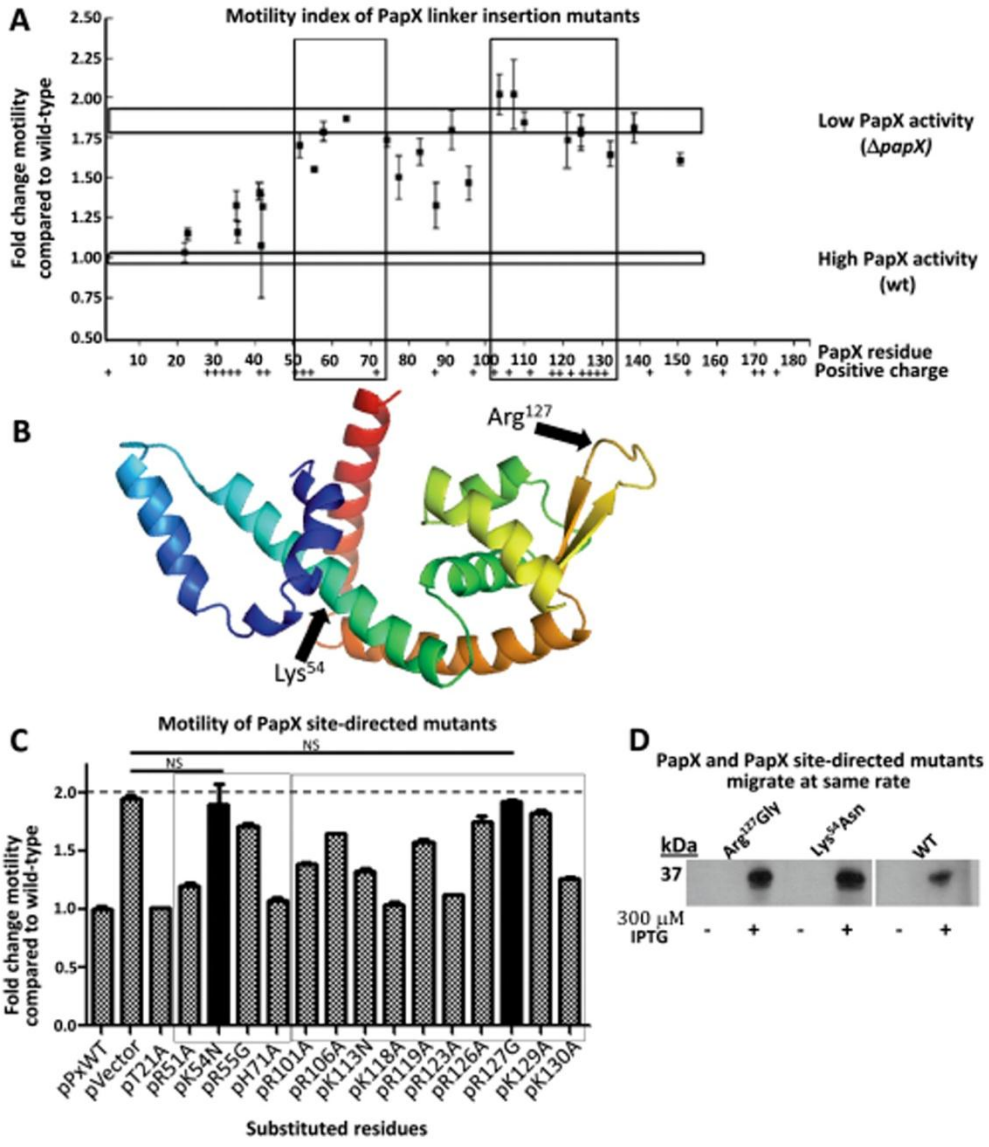

A
Motility index of PapX linker insertion mutants
Fold change motility compared to wild-type
Low PapX activity (ΔpapX)
High PapX activity (wt)
PapX residue
Positive charge
B
Arg127
Lys54
C
Motility of PapX site-directed mutants
NS
Fold change motility compared to wild-type
pPxWT
pVector
pT21A
pR51A
pK54N
pR55G
pH71A
pR101A
pR106A
pK113N
pK118A
pR119A
pR123A
pR126A
pR127G
pK129A
pK130A
Substituted residues
D
PapX and PapX site-directed mutants migrate at same rate
kDa
37
Arg127Gly
Lys54Asn
WT
300 μM IPTG

Table 2-2. Bacterial strains and plasmids used in this study.

Plasmid or strain name	Function
pMCSG7	Recombinant PapX-His_6
pMCSG7_R127G	PapX-His_6 including an $Arg^{127}Gly$ point mutant
pMCSG7_K54N	PapX-His_6 including an $Lys^{54}Asn$ point mutant
pMCSG7_T21A	PapX-His_6 including an $Thr^{21}Ala$ point mutant
pRS551	Promoterless *lacZ* containing vector
$pRS551_A$	*lacZ* fused to the first 130 bases of *flhD* promoter
$pRS551_B$	*lacZ* fused to the first 215 bases of *flhD* promoter
$pRS551_C$	*lacZ* fused to the first 360 bases of *flhD* promoter
$pRS551_D$	*lacZ* fused to the first 501 bases of *flhD* promoter
$pRS551_E$	*lacZ* fused to the first 597 bases of *flhD* promoter
$pRS551_F$	*lacZ* fused to the first 718 bases of *flhD* promoter
$pRS551_{GX}$	*lacZ* fused to the 262 bases between *papX* and *papG*
$pRS551_{BA}$	*lacZ* fused to the 330 bases between *papI* and *papB*
pLX3607 or pVector	Empty IPTG-inducible vector
pDRM001 or pPxWT	Wildtype *papX* made in pLX3607
pDRM002 or pR127G	*papX*_$Arg^{127}Gly$ constructed in pDRM001
pDRM003 or pK54N	*papX*_$Lys^{54}Asn$ constructed in pDRM001
pDRM004 or pR127K	*papX*_$Arg^{127}Lys$ constructed in pDRM001
pDRM005 or pT21A	*papX*_$Thr^{21}Ala$ constructed in pDRM001
pDRM006 or pR51A	*papX*_$Arg^{51}Ala$ constructed in pDRM001
pDRM008 or pR55G	*papX*_$Arg^{55}Gly$ constructed in pDRM001
pDRM009 or pH71A	*papX*_$His^{71}Ala$ constructed in pDRM001
pDRM010 or pR101A	*papX*_$Arg^{101}Ala$ constructed in pDRM001
pDRM011 or pR106A	*papX*_$Arg^{106}Ala$ constructed in pDRM001
pDRM012 or pK113N	*papX*_$Lys^{113}Asn$ constructed in pDRM001
pDRM013 or pK118A	*papX*_ $Lys^{118}Ala$ constructed in pDRM001
pDRM014 or pR119A	*papX*_$Arg^{119}Ala$ constructed in pDRM001
pDRM015 or pR123A	*papX*_$Arg^{123}Ala$ constructed in pDRM001
pDRM016 or pR126A	*papX*_$Arg^{126}Ala$ constructed in pDRM001
pDRM017 or pK129A	*papX*_ $Lys^{129}Ala$ constructed in pDRM001
pDRM018 or pK130A	*papX*_ $Lys^{130}Ala$ constructed in pDRM001
pKD3	λ red encoding chloramphenicol resistance cassette
pKD4	λ red encoding kanamycin resistance cassette
pCP20	λ red encoding recombinase to remove cassettes
pKD46	λ red encoding recombinase to insert cassettes
CFT073	Prototypical strain of pyelonephritogenic UPEC
CFT073 Δ*papX*	Deletion of *papX* in strain CFT073
CFT073 Δ*focX*	Deletion of *focX* in strain CFT073
CFT073 Δ*papX*Δ*focX*	Deletion of *papX* and *focX* in strain CFT073
CFT073 Δ*papX*Δ*lacZ*	Deletion of *papX* and *lacZ* in strain CFT073
CFT073 Δ*lacZ*	Deletion of *lacZ* in strain CFT073
K12 MG1655	Fecal/commensal strain K12 MG1655

insertional mutagenesis (amino acid residues 51-74 and 103-133) were predicted to occur in dimerization and DNA-binding domains, respectively (**Figure 2-4B arrows**).

To investigate the role of the two regions in PapX highlighted by insertional mutagenesis, site-directed mutants were constructed, in which positively-charged residues were substituted with non-positively charged residues in pPxWT (**Table 2-2**). Site-directed mutant constructs were transformed into CFT073 Δ*papX* and motility assays were performed for the resulting transformants (**Figure 2-4C**). We found that two mutations, $Lys^{54}Asn$ and $Arg^{127}Gly$, abolished PapX activity, producing the same motility phenotype as vector control (**solid bars**). We confirmed that each of these mutants produced PapX of the predicted apparent molecular weight as assessed by western blot **(Figure 2-4D**) by over-expressing each construct from an inducible vector in the K12 background, which lacks a native *papX* or *papX* homolog. These results indicated that either of these single amino acid substitutions alone is capable of ablating PapX activity, suggesting that Lys^{54} and Arg^{127} are key residues necessary for PapX function.

By homology modeling, Lys^{54} and Arg^{127} occur in the predicted dimerization domain and DNA binding domain of PapX, respectively. As discussed in the introduction, MarR family proteins often depend on an Arg in the wing of the winged-helix domain in DNA-binding [175-179]. Our alignment suggested that this conserved Arg corresponds to Arg^{127} in PapX. To further investigate the Arg^{127} mutant, we used site-directed mutagenesis to construct an $Arg^{127}Lys$ mutant containing a different positively charged residue. Surprisingly, the $Arg^{127}Lys$ mutant also disrupted PapX function as assessed by motility assay (**Figure 2-5A**), indicating that Arg^{127} is a key

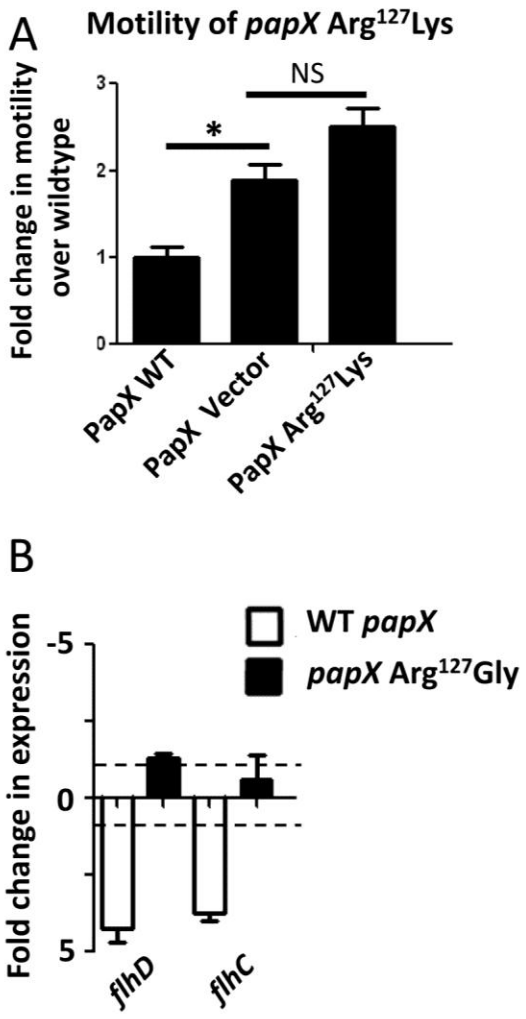


Figure 2-5. PapX Arg^{127}Gly mediates motility and flagellar transcription. (A) CFT073 Δ*papX* was complemented with pVector, pPxWT (WT *papX*) or pR127K (PapX Arg^{127}Lys mutant), and motility was assayed for each construct using soft-agar motility assays. Motility of CFT073 Δ*papX* electroporated with either pVector or pR127K was not significantly different from one another (NS, $P > 0.05$), and both were significantly more motile (*, $P < 0.05$) than CFT073 Δ*papX* electroporated with pPxWT. **(B)** pR127G (PapX Arg^{127}Gly mutant), pPxWT (WT *papX*) and pVector (empty vector) were electroporated into CFT073 Δ*papX*. RNA was isolated and analyzed using quantitative PCR, using *gapA* as the normalizer. Bars indicate fold-change in expression of *flhD* and *flhC* compared to CFT073 Δ*papX* electroporated with vector control. Dashed lines indicate 1 and -1 fold (no differential change in expression). Compared to the presence of pVector, both *flhD* and *flhC* transcript levels were decreased in the presence of pPxWT, but levels were unchanged in the presence of pR127G.

residue in PapX that is characteristic of the MarR family's DNA binding protein fold [175, 176] and that positive charge at position 127 alone is not sufficient for activity.

To better understand the motility phenotype of the Arg^{127} mutant, we compared the transcriptional profile of our Arg^{127}Gly mutant expressed from pR127G, wild-type PapX expressed from pPxWT, and vector control expressed from pVector in CFT073 Δ*papX* by quantitative real-time PCR (**Figure 2-5B**). Expression of *flhD* and *flhC* in the presence of the pR127G mutant was not significantly different from vector control, while pPxWT significantly reduced transcription of *flhD* and *flhC* (P = 0.0091 and P = 0.0469, respectively) (**Figure 2-5B**). Taken together, these results suggest that Arg^{127} is a key residue required for full PapX function, and plays a residue-specific role in the repression of motility by regulating expression of *flhD* and *flhC*.

PapX binds to a specific DNA motif within the *flhDC* promoter

Transcriptional data assessed by microarray, quantitative real-time PCR, and homology modeling suggested that PapX is a DNA-binding protein that acts directly on the *flhDC* promoter ([169]**, Figure 2-4B, Figure 2-5B**). As previously assessed by electrophoretic mobility shift assay (EMSA), PapX interacts with the 500 bases upstream of the *flhD* translational start site [169]. However, we had no data regarding sequence specificity of the target. To validate these results and to hone in on the region of the *flhDC* promoter recognized by PapX, we constructed nested deletions of the *flhDC* promoter and fused them to a promoterless *lacZ* in pRS551. Fusion constructs were transformed into CFT073 Δ*lacZ*Δ*papX* and CFT073 Δ*lacZ*. β-galactosidase activity was

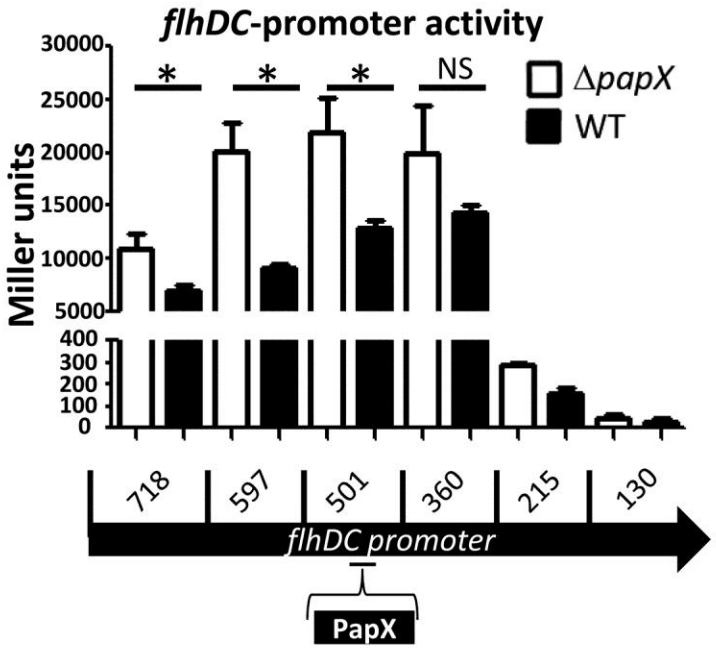


Figure 2-6. PapX repression of *flhDC-lacZ* promoter fusions. Nested deletions of the *flhDC* promoter were fused to promoterless *lacZ* in pRS551. Deletions contained progressively less of the 5' end of the promoter. pRS551-*flhDC* promoter constructs were transformed into CFT073 WT (black bars) or CFT073 Δ*papX* (white bars). No significant difference (NS, $P > 0.05$) was observed in β-galactosidase activity as assessed by Miller assay between CFT073 WT and CFT073 Δ*papX* using one-tailed *t*-test when only the first 360 bases of the *flhDC* promoter was included. Significant differences (*, $P < 0.025$) were observed when 501 bases or more of the *flhDC* promoter was included in the fusion construct.

measured using a modified Miller assay as a surrogate for promoter activity [195] (**Figure 2-6)**. A significant PapX-dependent difference in β-galactosidase activity was observed in all fragments containing at least 501 bases of the *flhDC* promoter ($P < 0.025$), but not in the fragment containing $\leq$ 360 bases of the *flhDC* promoter ($P > 0.13$). Thus it is likely that PapX exerts its effect between -360 and -501 bases upstream of the *flhDC* start site. Additionally, DNase protection and EMSA assays on the first 360 bases of the *flhDC* promoter failed to demonstrate sequence-specific effects (**Supplementary Figure 2-3**). This is consistent with previous findings in EHEC, a relative of UPEC, in which the *flhDC* promoter has a drastically reduced ability to drive gene expression when fewer than 300 bases of the 3' end of the promoter are present [151], as is the case with our fusions. The promoter fusion data suggest positive factors interact upstream of 300 bases that could be antagonized by a suppressor of *flhDC* transcription such as PapX.

PapX binds to a short motif in the CFT073 chromosome

To determine the exact binding site of PapX, we used SELEX methodology, an unbiased whole-genome approach [185], in conjunction with high-throughput sequencing technology. This combined two robust methods for identifying subtle, but specific, binding motifs [187]. To identify such a motif, wild-type CFT073 genomic DNA was sheared to generate 80-100 bp fragments. The resulting pool represented every possible 80-100 bp fragment from the CFT073 genome (confirmed after construction by high-throughput sequencing), or roughly $5x10^6$ unique members. Adaptors of approximately 50-bp in length were paired-end ligated onto each fragment to complete library synthesis.

Table 2-3. Recombinant PapX purity by LC-MS/MS

Protein	Ontology	emPAI*	Coverage
PapX	Transcriptional regulator	28.21	60%
L-glutamine:D-fructose-6-phosphate aminotransferase	Metabolism	.69	44%
cAMP receptor protein	Transcriptional regulator	2.27	53%
FKBP-type peptidyl-prolyl cis-trans isomerase	Protein folding	.55	17%
Aconitate hydrase B	Metabolism	.03	1%
30S Ribosomal subunit S3	Translation	.13	5%

***Exponentially Modified Protein Abundance Index (emPAI) is an estimate of the prevalence of a species in a sample analyzed by LC-MS/MS.**

Recombinant PapX-His_6 was purified and analyzed by mass spectroscopy to confirm that it was the predominant species after purification (**Table 2-3**). PapX-His_6 was then bound to Ni-NTA resin and a sample (200 μl) of resin-bound PapX-His_6 was incubated with library DNA (150 ng, 30 min, 37^{o}C). Unbound DNA was removed with three washes in 20 mM imidazole column buffer; bound PapX-DNA complexes were eluted using 500 mM imidazole elution buffer. DNA, isolated from the eluted PapX-DNA complexes by phenol:chloroform-extraction, was PCR-amplified using primers specific for the paired-end adapters to ensure that only the library, and not contaminating DNA, was amplified. The enriched library DNA was quantified with real-time PCR using primers specific to the adapters (but internal to the amplification primers) and then applied to a new Ni-NTA column containing bound PapX-His_6. The entire process of binding the library to Ni-NTA-bound PapX-His_6, amplifying bound members and re-applying the enriched library to Ni-NTA-bound PapX-His_6 was repeated for four cycles. The amounts of PapX-His_6 and DNA used were empirically determined to be the least quantities that still allowed for low-cycle (*i.e.* unbiased) amplification [196] of each cycle's DNA output to regenerate sufficient template for the next cycle of SELEX.

The initial CFT073 sheared genome library, as well as the second, third and fourth libraries enriched by serial rounds of SELEX, were sequenced using a high-throughput Illumina sequencing. On average, $24 \pm 2.8 \times 10^6$ 80-bp reads were obtained for each SELEX round, representing 400-fold coverage of the CFT073 genome. Sequences for each round were aligned to the CFT073 genome using a local BLAST (**Figure 2-7A**), showing a clear pattern of site-specific enrichment. A dot-plot of the frequencies of hits at each position in the chromosome is shown for a single

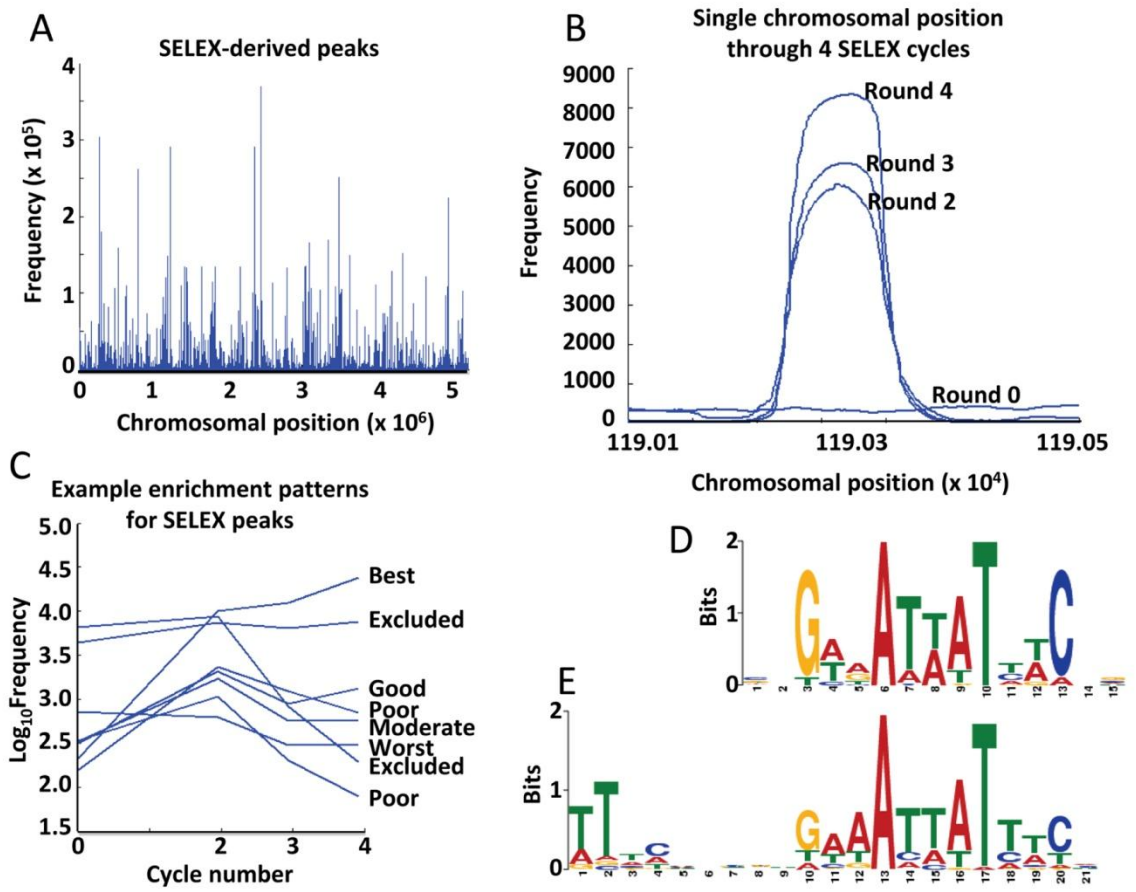


Figure 2-7. Analysis of SELEX Illumina sequencing results. (A) The fourth enrichment of SELEX was sequenced and results were aligned to the genomic sequence of strain CFT073 using a local BLAST. Frequencies of a sequence read at every chromosomal position were plotted using a dot-plot. **(B)** Dot-plot of frequencies across SELEX rounds zero, two, three and four in the chromosomal vicinity of a prototypical low-frequency best-grade peak. **(C)** Representative example of chromosomal peak positions and their frequencies plotted against SELEX rounds zero, two, three and four. Grades of peaks are based on their pattern of frequencies across rounds and appears to the right (defined in the footnote of **Table 2**). The MEME suite of sequence analysis tools were employed to determine if motifs were present in the list of 366 Best-grade peaks, generating two significant motifs. **(D)** Position weight matrix present in 362 of the 366 Best-grade peaks, *E*-value 4.5 x 10^{-142}. **(E)** Position weight matrix present in 269 of the 366 Best-grade peaks, *E*-value 4.7 x 10^{-140}.

representative peak across all sequenced rounds of SELEX (**Figure 2-7B**). The width of peaks (160-200 bp) corresponds to roughly twice the size of an average member of the sheared genomic library, suggesting that PapX-His_6 interacts with the CFT073 genome in a strongly sequence-specific manner. Peak positions generated from the library of the final round of SELEX were overlaid onto each sequenced round of enrichment, and the frequency of occurrence at that chromosomal position was plotted against the round of enrichment (**Figure 2-7C**). Peaks were graded based on their rate of enrichment across rounds (**Table 2-4)**, with heavy weight given to positions that enriched under the most stringent environment (that is, the final rounds of enrichment), and discarded if they were too heavily represented in the control library at greater than two standard deviations above the mean positional frequency across the entire genome of the unselected initial library. Sufficient selective enrichment occurred as indicated by the background representation of library members dropping to zero between peaks (**Figure 2-7B**). These data were sufficiently rich to generate bell-shaped curves of predicted width (160-200 bp) and appropriate distribution from dot-plots, indicating that the protocol was successful at amplifying only regions that contained short, PapX-specific binding sequences. Of the sequences generated from the fourth round of SELEX, frequencies at 94% of the chromosome were below the mean of the unselected library, and the frequency at 64% of chromosomal positions was 0. This demonstrates that the majority of the chromosome was successfully counter-selected and did not interact with PapX as assessed by SELEX.

Of 4,407 total peaks in the final round of enrichment, 366 satisfied the "best" criteria for peaks (**Table 2-4**). Sequence flanking each best peak by 30 bases both up and downstream in the chromosome was analyzed using the online MEME suite of tools

Table 2-4. Positional-weight matrix motif derivatives prevalence in graded peaks and randomly selected sequences

Motif	All peaks[1]	Best[2]	Good[3]	Moderate[4]	Poor[5]	Worst[6]	Control[7]
Derivatives of GWWAWWWTWWC	%	%	%	%	%	%	%
AAAAT	50	62	49	41	49	46	18
AAATT	34	48	34	23	34	23	11
AAATTT	7	10	6	3	7	7	2
AATTTC	12	22	12	5	11	7	3
ATTTC	35	52	39	32	32	23	12
TTATT	47	69	48	38	45	30	12

[1] 4407 peaks

[2] "Best" quality peaks are the 366 peaks that increased in frequency each round of SELEX

[3] "Good" quality peaks are the 932 peaks that increased in frequency in the final round of SELEX, and had a fourth round frequency greater than 2,800

[4] "Moderate" quality peaks are the 269 peaks that increased in frequency in the final round of SELEX, and had a fourth round frequency less than 2,800

[5] "Poor" quality peaks are the 2827 peaks that increased in frequency only in the first round of SELEX

[6] "Worst" quality peaks are the 13 peaks that decreased in frequency in the first round of SELEX

[192]. 362 of the 366 (99%) best peaks contained a single core motif represented by positional weight matrix (**Figure 2-7D**), with an *E*-value of $4.1x10^{-142}$. [*E*-value is an estimate of the expected number of a particular motif one would find in a similar, random set of sequences; thus, an *E*-value of $4.1x10^{-142}$ indicates a $4.1x10^{-142}$ probability the motif arose by chance.] A second, modified version of this motif was found present in 269 of the 366 (73%) best peak sequences and represented by positional-weight matrix with an *E*-value of $4.7x10^{-140}$ (**Figure 2-7E**). Because motifs were calculated presuming positional independence, palindromes were more difficult to identify using the bits scoring system. Within the 269 sequences, the second motif appeared as a palindrome, described using the International Union of Pure and Applied Chemistry (IUPAC) nomenclature as TT(7n)GWWAWWWTWWC(7n)AA (where W is A or T). Database analysis using Microsoft Access verified that the palindromic form of the second, modified core motif appeared significantly more often than if by chance within the 269 sequences in which the second motif was present (**Table 2-4**). Altogether these represented the motifs with the lowest *E*-values within the set; the next lowest value motif was represented by fewer than 4 members with *E*-values greater than 10^{-5}.

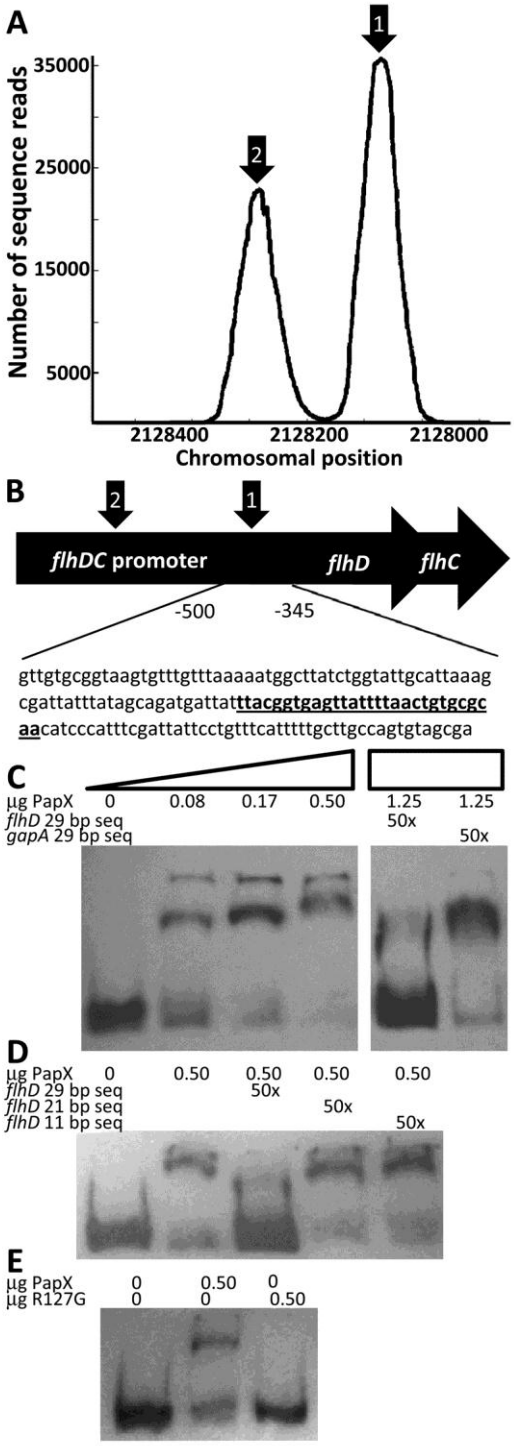
To verify that these motifs were not emerging at random, two methods of validation were employed. The input sequences were shuffled randomly and analyzed for motifs, which produced no motifs with *E*-values lower than 1,000 (that is, motifs generated from the shuffled set were very likely to have arisen by chance). Additionally, six groups of 366 random peaks were generated using Matlab and analyzed in the same manner as the best peaks. The rates of the predicted motif (**Figure 2-7D**) and several derivatives were significantly less common in the random peaks using Microsoft Access

(**Table 2-4)** and MEME could detect no significantly represented motifs from the random peak sets. Together, these data suggested that PapX binds to the CFT073 chromosome in a sequence-dependent manner at a motif 11-29 bases long.

PapX interacts with a short sequence within the *flhDC* promoter of CFT073

Two best-quality peaks appear within the *flhDC* promoter (**Figure 2-8A**). The larger of the two peaks, centered 410 bp upstream of the *flhDC* translational start site, centers around a perfect match to both predicted motifs, matching 11 of 11 and 29 of 29 bases (**Figure 2-8B**). The second, smaller peak at -624 bases upstream of the *flhDC* translational start site is a degenerate version of the motif, matching only 9 of 11 and 25 of 29 bases. To validate our findings, we used EMSA to shift the 154-bp fragment, -501 to -346 bp upstream of the *flhD* translational start site, containing the best version of the predicted motif (**Figure 2-8C**). The motility of the fragment through a 5% acrylamide gel was retarded in the presence recombinant PapX-His_6. This shift was dose-dependent with increasing concentrations of PapX. In a competition assay, we competed the DNA-binding reaction with unlabeled short sequences. A 29-bp sequence (**Figure 2-8B, bold, underlined)** found within the 154 bp target fragment (representing a 29 of 29 nucleotide match with the second, modified core motif (TT(7n)GWWAWWWTWWC(7n)AA)) was added in 50-fold excess by mass (250-fold by molarity) to the DNA binding reaction between PapX and the 154 bp fragment, and was able to inhibit the shift. A second 29 bp fragment derived from *gapA* was unable to reverse the shift under the same conditions (**Figure 2-8C**). The 29 bp fragment found in the *flhDC* promoter appears only

Figure 2-8. PapX binds to the *flhDC* promoter. (A) Dot-plot of chromosomal frequencies by SELEX revealed two significant peaks within the *flhDC* promoter (arrows 1 and 2). The peak under arrow 1 includes the predicted PapX binding site. **(B)** Schematic of the *flhDC* promoter, indicating the region -346 to -501 bases upstream of the *flhD* translational start used in subsequent gel-shifts. A perfectly identical (29/29) predicted PapX binding site is indicated in bold underline and is centered 410 bp upstream of the *flhD* translational start site. **(C)** Increasing amounts of PapX protein were added to 3.2 ng of DIG-UTP-end-labeled 154 base pair fragments derived from the *flhDC* promoter, the sequence shown in **(B**). Unlabeled competitor DNA at 50-fold excess by mass was added to the protein-DNA binding reaction for each of the last two lanes. "*flhD* 29 bp seq" is the sequence indicated in bold underline in **(B)**. "*gapA* 29 bp seq" is the sequence ACTCCACTCACGGCCGTTTCGACGGTACCG from *gapA*. **(D)** The competition assay performed in **(C)** was repeated using only the 0.50 μg amount of PapX-His_6 protein. Additional sequences derived from the 29 bp competitor (TTACGGTGAGTTATTTTAAC, indicated by "*flhD* 21 bp seq" in the figure and GTTATTTTAAC, indicated by "*flhD* 11 bp seq" in the figure), determined by sequence analysis in **(Fig. 2-7D)** and **(Fig. 2-7E)** were also used in 50-fold excess by mass. **(E)** Gel shifts were repeated with the same 154 bp fragment as in **(C)** and with 0.50 μg amounts of wildtype PapX-His_6 protein and PapX-His_6 containing the $Arg^{127}Gly$ mutation.


A
Number of sequence reads
35000
25000
15000
5000
2128400
2128200
2128000
Chromosomal position
1
2
B
2
1
flhDC promoter
flhD
flhC
-500
-345
gttgtgcggtaagtgtttgtttaaaaatggcttatctggtattgcattaaag
cgattatttatagcagatgattattttacggtgagttattttaactgtgcgc
aacatcccatttcgattattcctgtttcatttttgcttgccagtgtagcga
C
µg PapX 0 0.08 0.17 0.50 1.25 1.25
flhD 29 bp seq 50x
gapA 29 bp seq 50x
D
µg PapX 0 0.50 0.50 0.50 0.50
flhD 29 bp seq 50x
flhD 21 bp seq 50x
flhD 11 bp seq 50x
E
µg PapX 0 0.50 0
µg R127G 0 0 0.50

once in the *E. coli* strain CFT073 genome, and no similar sequences appear in the TRANSFAC database by TfBlast.

We also tested the ability of the motifs predicted by SELEX (**Figure 2-7D, 2-7E**) to bind PapX using the same competition assay approach. Oligos matching the 11 bp sequence and 21 bp sequence were assayed for their ability to compete with DNA binding to the 154 bp *flhD* promoter fragment. All competing fragments were added in

50-fold excess by mass. Results indicated that only the 29 bp fragment was able to bind PapX, demonstrated by its unique ability among competitor DNA fragments to reverse the shift observed with wildtype recombinant purified PapX and the 154 bp *flhD* promoter fragment (**Figure 2-8D**).

As mentioned earlier, structural modeling predictions based on I-TASSER and alignments suggested that Arg^{127} in PapX could represent a key residue required for DNA binding [175-179]. To test this hypothesis, recombinant purified PapX-His_6 containing the Arg^{127}Gly point mutation was used in an EMSA assay against the 154 bp fragment of the *flhD* promoter (**Figure 2-8E**). Wildtype recombinant PapX-His_6 was able to shift the fragment, indicating binding, whereas the binding of the Arg^{127}Gly mutant was ablated (**Figure 2-8E**).

Taken together, these results suggest that PapX binds to the *flhDC* promoter in a sequence-specific manner at a novel, unique motif predicted by SELEX, at a position consistent with the findings of our *flhDC* promoter-fusion experiment.

Discussion

Previously, we have shown that PapX regulates motility via transcriptional regulation of *flhD and flhC* in UPEC [169]. The current study expands upon this finding, defining the mechanism by which PapX, a 43 kDa non structural dimeric protein encoded by an operon that encodes P fimbria, mediates its effect. Our structural and functional results indicate that PapX is a dimeric, helix-turn-helix DNA-binding protein of the 17 kDa family, a MarR homolog that relies on Lys^{54} and Arg^{127} for full activity. Our novel use of SELEX in conjunction with high-throughput sequencing in bacteria indicates that PapX exerts its effect by binding to a unique, specific, previously unknown sequence, TTACGGTGAGTTATTTTAACTGTGCGCAA, within the *flhDC* promoter, centered at a position 410 bp upstream of the *flhD* translational start site. Because PapX homologs are so closely related (>93% amino acid sequence identity), these findings have implications across a broad range of ExPEC strains and their likely regulation of motility and adherence.

Evaluating the interaction of PapX and the *flhDC* promoter

MarR family proteins tend to bind motifs between 21 and 45 bases long [180]. In addition, we demonstrated that neither the 21 bp motif (**Figure 2-7E**) nor the 11 bp motif (**Figure 2-7D)** are alone sufficient to bind PapX (**Figure 2-8D**). The 29 bp palindromic form of the 21 bp motif, however was sufficient to bind PapX, whereas a 29 bp sequence that did not contain either the 21 bp or 11 bp motifs generated from *gapA* was unable to bind PapX (**Figure 2-8C**). These data indicate that both the specific sequence content, as

well as the length of the sequence, are together necessary to mediate PapX binding to DNA. This may explain why the palindromic form of the 21 bp motif did not appear in the Best-grade peaks; the software used to identify motifs, MEME [192], analyzes sequences for conserved nucleotides at given positions, but cannot predict important regions that are necessary for topological reasons or that contain degenerate sequence. Both motifs (**Figure 2-7D, 2-7E**) appear to be necessary and specific sequences, as indicated experimentally (**Figure 2-8C, 2-8D**). This is supported as well *in silico* by their comparably low *E*-values, which were by far lower than any other motif identified. This suggests that SELEX with high-throughput sequencing was successful at identifying necessary binding elements for PapX.

Interestingly, a degenerate version of the 11 bp motif exists at the 5' end of the 360 bp fragment used in the *lacZ* fusions (**Figure 2-6**). This may have produced a small degree of PapX interaction with the fragment, and may explain the elevated variance seen between replicate assays with the 360 bp fragment. The variance appeared to be biological in origin, persisting across three repeats of the experiment, each using biological triplicates and technical duplicates. It is also possible that PapX acts by antagonizing an activator, so a PapX binding site may be present in the 360 bp fragment but reduced upstream activator binding may explain the difference in β-galactosidase activity.

Over half of the 366 best peaks fall within predicted promoter regions for genes. As mentioned, over-expression and deletion of *papX* has a transcriptionally specific effect [169], with transcriptional regulation confined to a small list of genes or gene-classes. There are several possible explanations for this mismatch phenomenon. First,

SELEX is a molecular technique that helps to make inferences about binding energy between a ligand and a nucleic acid sequence. When aptameric libraries are used, limited inference about biological relevance of the strength of *in silico* ligand-protein interactions can be drawn. SELEX is best used to identify target motifs, but follow-up, such as by EMSA or competition assay, is necessary to demonstrate biological significance.

We have confirmed a direct interaction of PapX with a specific sequence within the *flhDC* promoter, and demonstrated a transcriptional effect on the downstream genes. By investigating a binding site known to produce an effect, we may discover additional clues to the mechanism whereby PapX regulates gene expression. When theoretical sites are determined using aptameric libraries, there must be follow-up to confirm biological relevance, as the fact of DNA-protein interaction by SELEX is not sufficient to guarantee physiologically-relevant action. Our studies have successfully demonstrated the utility of SELEX to identify a short, specific sequence within which a protein acts to modulate transcriptional regulation.

Pathogen specificity of PapX-*flhDC* promoter interaction

Our multiplex results (**Figure 2-2**) and motility assays (**Figure 2-3**) indicated that the effect of PapX is pathogen-specific. However, the binding site identified by SELEX (**Figure 2-8B, bold, underlined**) appears as a 29/29 nucleotide match in the *flhDC* promoter of strain K12. In this work and previously, control of motility in UPEC by PapX is mediated by *flhD* and *flhC* transcriptional repression. Taken together, this suggests that there are additional factors present in strain CFT073 and absent in strain

K12 that affect the activity of PapX on *flhD* and *flhC* transcription, and that the mere presence of a nucleotide sequence with likely PapX affinity is insufficient to regulate downstream transcription in the commensal strain. For example, PapX may be regulated by other factors (such as small molecules, as with salicylate derivatives and MarR [176]) that locally affect PapX activity at the site of regulation. This may help to explain the discrepancy between the transcriptional effect of PapX described previously [169], which is limited to 42 differentially regulated genes (38 of which are downstream of *flhD* and *flhC*), and the list of hundreds of peaks by PapX with SELEX. It is possible that PapX binds most or all of the peak regions identified, but that other factors are necessary to produce a differential transcriptional effect. Models of such an interaction include PapX binding resulting in displacement of activators, PapX acting as a scaffold for repressors, or PapX undergoing allosteric and conformational changes that would modify DNA topology and modify the binding properties of the *flhD* promoter for *i.e.* RNA polymerase. Additionally, the presence of a cofactor may reduce the promiscuity of afSELEX-seq selection for fragments. The 29 bp fragment found in the *flhDC* promoter appears only once in the *E. coli* CFT073 and *E. coli* K12 genomes, offering further evidence of the specificity of transcriptional effect by PapX.

Functional characterization of PapX

Our data suggest that Arg^{127} in PapX is crucial for DNA binding (**Figure 2-8E**). Taken together with the loss of wildtype PapX phenotype of the Arg^{127}Gly mutant seen by motility assay (**Figure 2-3, Figure 2-4**) and loss of transcriptional repression (**Figure**

2-5), suggesting that it is essential for PapX to bind DNA to exert its motility-repressing phenotype.

In addition, within the PapX homologs, Ala to Thr substitutions are found commonly (**Figure 2-1**), occurring at residue positions 52, 82, 114 and 160. Ala-Thr substitutions also affect protein activity in other proteins [197]. Our $Ala^{21}Thr$ mutation had full PapX activity by motility assay suggesting that at least at some positions this substitution is inconsequential for PapX activity. It would stand to reason that substitutions that have a small or no impact on protein function will emerge by chance and not remain conserved, offering a potential explanation of why the Ala-Thr substitutions are common among the homologs.

The combination of SELEX and Illumina high-throughput sequencing provided a high resolution map of putative PapX binding sites. A single unique sequence within the genome of *E. coli* CFT073, a prototype of UPEC strains, was identified within the *flhDC* promoter, which controls flagellar gene expression. Direct binding of the target sequence by PapX was demonstrated. Repression of motility was observed in the pathogenic strain but not a non-pathogenic strain, suggesting the presence of additional factors required for this regulation. Such accessory factors will be the target of future studies, which will also aim to address the discrepancy between the specific list of *papX* differentially regulated genes seen previously [169] and the more extensive number of peaks determined by SELEX.

CHAPTER 3

INVESTIGATING THE REGULATION OF *PAPX*

Introduction

To understand the mechanism of reciprocal control of motility by adherence, in Chapter 2 we focused on the mechanism of action of PapX and its role in suppressing motility. Though we have begun to understand the mechanism of action of PapX, what remains to be understood is how PapX itself is regulated. Underscored in the model of reciprocal control is that motility-suppressing mechanisms should be up-regulated in the context of adherence. Thus, we sought to determine the relationship between PapX and adherence. As discussed in Chapter 2, uropathogens encoding *papX* also encode the *pap* operon up to 80% of the time. This suggests that there is an association between PapX and adherence. To further explore the relationship, we examined the levels of expression of *papX* mRNA and compared them to the levels of transcription of members of the *pap* operon. If *papX* expression is tied to that of the *pap* operon, we would expect analysis of the transcription of each under identical circumstances to be illuminating about the mechanism of *papX* regulation. We also created fusions of the region upstream of *papX* to a promoterless *lacZ* and assayed for β-galactosidase activity to determine whether or

not *papX* may have its own promoter and own set of transcriptional controls. Results from transcriptional profiling and fusions lead us to consider other mechanism of control of PapX.

As discussed in Chapter 2, an $Arg^{127}Gly$ mutant of PapX fails to inhibit motility of *E. coli* CFT073, and it is unable to bind the *flhD* promoter as judged by EMSA. Thus a simple model of PapX inactivation is that Arg^{127} is a crucial residue for establishing contact between PapX and DNA. Disruption of this contact would inhibit PapX repression of *flhD*. We also noted that a $Lys^{54}Asn$ mutant of PapX is unable to repress motility in *E. coli* CFT073. We sought to determine the contribution of Lys^{54}, to shed light on the mechanism of action of wild type PapX. We performed DNA binding assays using the $Lys^{54}Asn$ mutant and found it bound *flhD* promoter DNA. This led us to consider an additional feature of PapX apart from merely binding of *flhD* DNA, and led us to postulate that post-binding events are important in the regulation of PapX activity. To attempt to determine if PapX interacts with cofactors that might explain the $Lys^{54}Asn$ inactivation phenotype, we performed co-precipitation experiments, in which immobilized, purified PapX was used to bind putative cofactors present in pooled wildtype *E. coli* strain CFT073 lysates. We predicted that by examining the enriched products of a co-precipitation, we would discover factors interacting with PapX to help explain how PapX's post-binding activity is regulated.

Results

Investigating *papX* transcriptional regulation

We wanted to investigate how *papX* and its effects on the cell were regulated. Previous work has shown that *papX* mRNA can be amplified from a transcript of the *pap* operon [169]. That result suggested the simple model that regulation of *papX* may proceed via transcriptional control and that *papX* and *pap* expression should be coincident. This model is appealing because it provides a direct explanation of how an adherence factor might suppress motility. Microarray studies have compared CFT073 collected from the urine of infected mice (*in vivo* condition) to CFT073 cultured in LB, either cultured statically or with aeration (*in vitro* conditions) [38]. In that work, most genes within the *pap* operon varied significantly between *in vivo* and *in vitro* conditions, but *papX* transcript levels did not change significantly **(Figure 3-1A**), suggesting that regulation of *papX* expression is influenced by other factors than merely the transcript levels of the *pap* operon. Additionally, the comparison between groups indicated that the *pap* operon is repressed *in vivo* compared to *in vitro* conditions. Later work scrutinizing the expression of mRNAs from human UTI isolates indicated that *papX* is generally up-regulated *in vivo* [106]. Consistent with *fliC in vivo*, this further supports the notion that *papX* may be transcriptionally regulated at least in part independently of the *pap* operon, or perhaps that the phenotypic effects of *papX* on the cell are not regulated at the level of transcription.

Past work by our lab had shown that suppression of motility in *E. coli* strains constitutively expressing Type 1 fimbriae (L-ON) is dependent upon *papX*, because

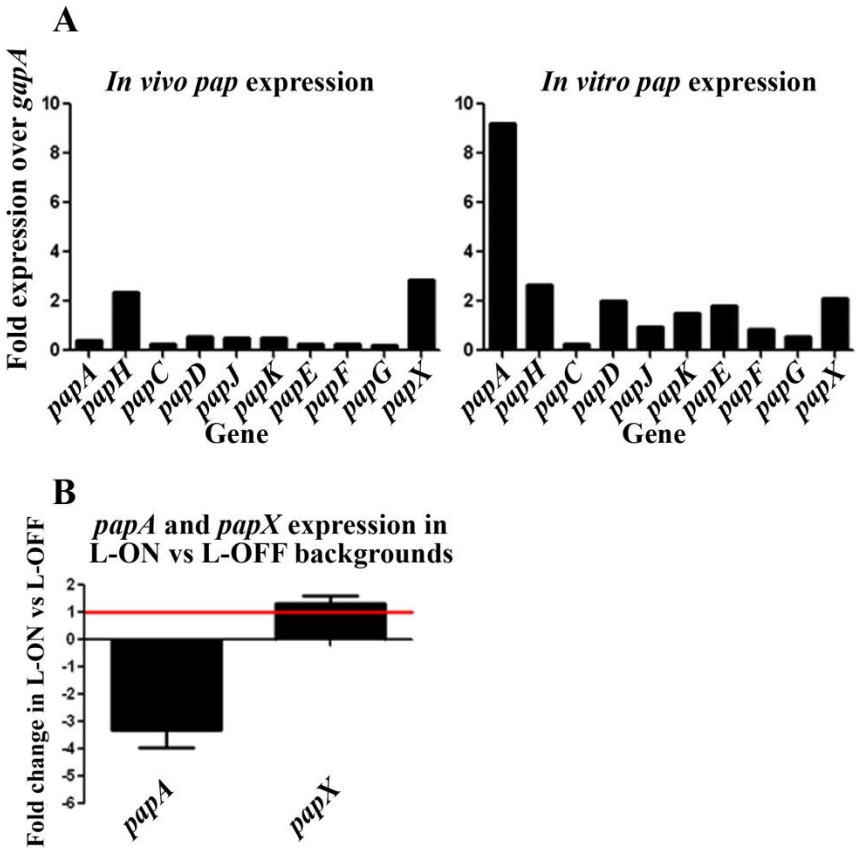


Figure 3-1. *papX* transcription is not regulated by the *pap* operon. (A) Transcript levels of the *pap* operon in *E. coli* CFT073 derived from mouse urinary tracts (**left**) or from culture in LB **(right**) as determined by microarray. *papX* transcript levels do not vary significantly between the two conditions. *papAHDJKEF* all vary. Data extracted from [38]. **(B)** *papA* and *papX* transcript levels as determined by qPCR in *E. coli* CFT073 L-ON compared to L-OFF. *papA* transcript levels are 3.5 fold lower in *E. coli* CFT073 L-ON as compared to L-OFF ($p < 0.05$). *papX* transcript levels did not vary significantly from a fold change of 1 between L-ON as compared to L-OFF.

deletion of *papX* in L-ON strains of *E. coli* CFT073 partially rescues the reduced-motility phenotype [169, 171]. Prior data also indicate that the Type 1 fimbria L-ON and L-OFF strains affect *pap* operon expression [113]. When incubated statically in liquid broth, known to induce Type 1 fimbrial expression [68] and suppress P fimbrial expression [107], L-ON strains repressed *pap* operon expression, especially *pap_2*, when compared to L-OFF strains of *E. coli* CFT073.

We decided to test whether *papX* and *pap* operon expression would vary under conditions of L-ON compared to L-OFF strains of *E. coli* CFT073 under conditions of growth in aerated liquid culture. Using quantitative reverse-transcriptase PCR we directly compared the levels of expression of *papA* and *papX* mRNAs between the L-OFF and L-ON strains. In our hands, *papA* expression was increased 3.5-fold in the L-ON strain compared to the L-OFF strain as judged by quantitative PCR (**Figure 3-1B**). Under the same conditions, *papX* expression was not significantly affected, as indicated by quantitative PCR (**Figure 3-1B**). The difference in trend of *pap* expression and type 1 fimbrial expression is likely due to the difference in culturing conditions. We cultured the strains under aeration while previous work had studied the transcriptional effect under static growth conditions. This confirmed the trend that the phase-state of Type 1 fimbriae influences *pap* expression, and highlights the interesting feature that culture conditions influence fimbrial expression. In all cases, variation in expression of the *pap* operon appears to occur with no significant impact on the expression of *papX*. These data support the model that *papX* expression is regulated independently of *pap* operon expression.

To test this model, we cloned the DNA sequence between *papG* and *papX* within the *pap* operon in front of a promoterless *lacZ* in plasmid $pRS551_{GX}$. The same technique was applied to the promoter of the *pap* operon to generate $pRS551_{BA}$. Both constructs $pRS551_{GX}$ and $pRS551_{BA}$ drove *lacZ* expression compared to empty vector, suggesting that each of these regions contains at least minimal promoter elements. $pRS551_{BA}$ had roughly two-fold increased strength according to Miller assay as compared to $pRS551_{GX}$ (**Figure 3-2A**). In the L-ON background, the $pRS551_{BA}$ demonstrated a 2-fold increase in expression of *lacZ* compared to the L-OFF strain (**Figure 3-2B, 3-2D**). Both vector and $pRS551_{GX}$ showed no significant difference in *lacZ* expression between either L-ON or L-OFF backgrounds (**Figure 3-2C, 3-2D**). Together, these findings suggest that *papX* has its own promoter, and that expression of *papX* is neither dependent on Type 1 nor P fimbrial expression.

Other regulators of PapX

In Chapter 2, we demonstrated that PapX binds to the *flhD* promoter, and that it is likely that this binding is necessary for activity. We found that a $Lys^{54}Asn$ mutant of PapX does not repress motility in *E. coli* strain CFT073 by motility assay (**Figure 2-4C**). However, this mutant PapX does bind the *flhD* promoter by EMSA (**Figure 3-3A**). Binding without repression of motility suggests there are additional processes downstream of binding relevant to the motility-repressing effect of PapX. We were curious whether there were cofactors involved in full PapX activity. Several parallel

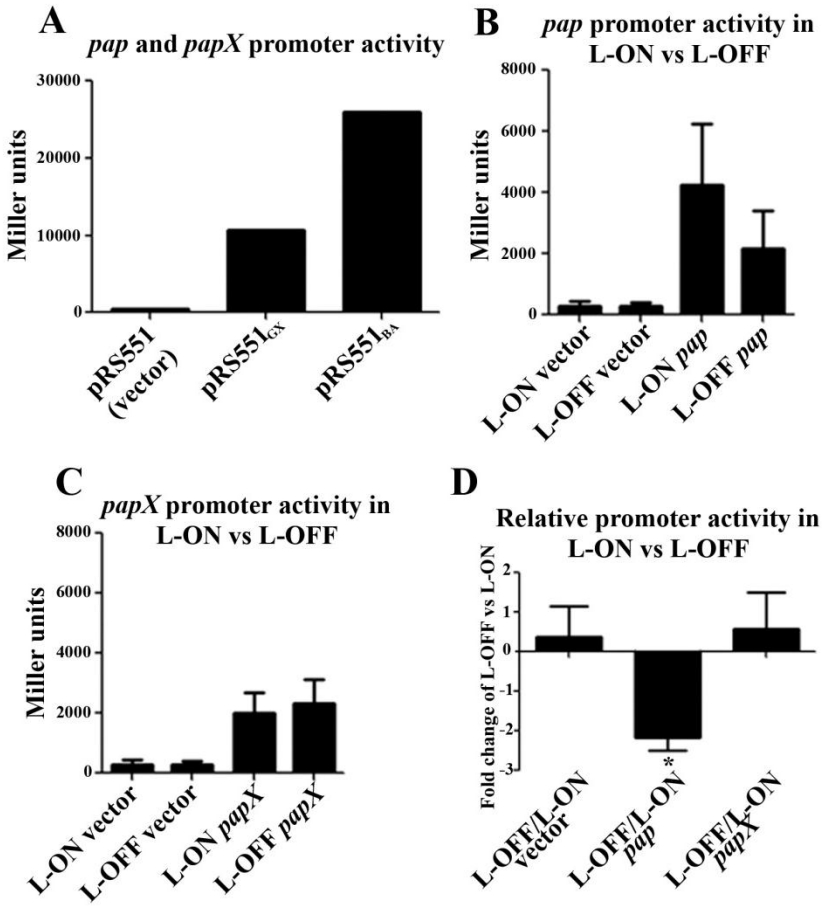


Figure 3-2. Type 1 phase state does not influence *papX* expression. (A) β-galactosidase activity was measured and calculated in Miller Units as a surrogate for promoter strength in the promoterless pRS551. $pRS551_{GX}$ and $pRS551_{BA}$ each drive *lacZ* expression, indicating that the fusions contain promoter elements. Vector did not express *lacZ* significantly over background. **(B)** $pRS551_{BA}$ drives statistically significantly ($p < 0.05$) increased *lacZ* expression in L-ON strains than in L-OFF strains of *E. coli* CFT073. Vector control does not vary significantly from background in either background. **(C)** $pRS551_{GX}$ drives equivalent *lacZ* expression in L-ON strains than in L-OFF strains of *E. coli* CFT073. Vector control does not vary significantly from background in either background. **(D)** Fold change of β-galactosidase activity attributable to $pRS551_{BA}$ and $pRS551_{GX}$ in L-ON relative to L-OFF backgrounds of *E. coli* strain CFT073. $pRS551_{BA}$ demonstrated significantly repressed β-galactosidase in L-OFF compared to L-ON strains of *E. coli* CFT073. $pRS551_{GX}$ and vector did not demonstrate significantly repressed β-galactosidase in L-OFF compared to L-ON strains of *E. coli* CFT073.

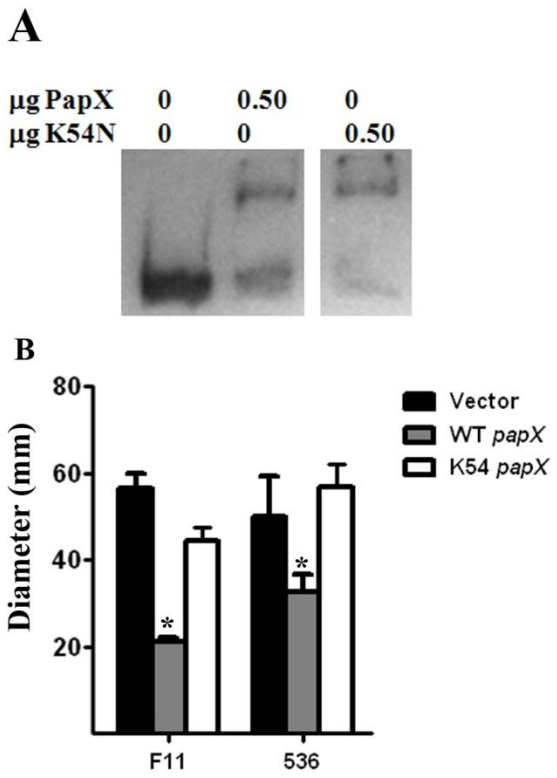


Figure 3-3. PapX Lys^{54}Asn binds *flhD* promoter but doesn't repress motility. (A). 0.50-µg of recombinant PapX-His6 protein shifts 3.2 ng of DIG-UTP-end-labeled 154 bp fragment of the *flhD* promoter. An equal amount of the Lys^{54}Asn mutant form of PapX is also able to shift the fragment under identical conditions. **(B)** WT *papX* encoded on pPxWT significantly ($p < 0.05$) represses motility of the UPEC strains F11 and 536 as compared to vector control. Lys^{54}Asn encoded on pK54N does not significantly influence motility as compared to vector control.

approaches were employed to aid in discovering such a factor. First, we reasoned that strains of *E. coli* that display repressed motility in the presence of *papX* should have the factor present. As demonstrated in Chapter 2 (**Figure 2-3**), *E. coli* CFT073 displays repressed motility in the presence of *papX*, but *E. coli* K12 does not, suggesting that, if there is a factor, it is present in the former but not the latter. This experiment was repeated using other uropathogenic *E. coli* strains F11 and 536, both of which displayed repressed motility in the presence of *papX* (**Figure 3-3B**).

To further refine the list of candidate proteins that may act in concert, we performed a co-purification. We reasoned that if a cofactor or cofactors exist, they should be associated with PapX. Nickel-NTA resin, either without modification (mock group) or loaded with recombinant, purified, His-tagged PapX (PapX group), was incubated with wild type *E. coli* strain CFT073 lysate. His-tagged PapX has been shown to be phenotypically active by motility assay (data not shown). Eluate was subjected to tandem LC-MS/MS mass spectroscopy, and the resulting species of each group were compared to highlight factors from the lysate of *E. coli* CFT073's lysate present or enriched in the presence of PapX-His_6 but not by Nickel-NTA resin alone. Proteins were identified by searching an *E. coli* CFT073 database. We searched only for proteins with at least 5 spectra using Assigned Spectra in Scaffold (Proteome Software Inc. 2005-2010) with a minimum peptide value of 95% and protein value 99.9%. 21 proteins had more spectra in the PapX group as compared to the mock group, and only 6 proteins had no spectra in the mock and at least 5 in the PapX group, which we took to be the enriched proteins. We scrutinized the enriched proteins for members that were different between *E. coli* K12

Table 3-1 Tandem LC-MS/MS spectroscopy results

Identified Protein	Molecular Weight	# spectra[1] in mock	# spectra[1] in PapX
PapX	19 kDa	2	161
PapX	19 kDa	0	14
OmpA	37 kDa	0	5

[1]# spectra refers to the total number of spectra in the LC-MS/MS analysis that matched an identified protein. This is an estimate for the presence of a given protein species.

and *E. coli* CFT073, F11 and 536. Three proteins were not 100% identical between K12 and the pathogenic strains (**Table 3-1**). The first two were identified as PapX and are likely two products of the same species, which was enriched 88-fold in the PapX group as compared to mock. The other two was the outer membrane protein OmpA, which was absent in the mock as determined by LC-MS/MS.

In *E. coli* strains CFT073, 536 and F11, OmpA is 96%, 98% and 36% identical to that of K12. The OmpA of *E. coli* CFT073 has an additional 5 amino acid domain between positions 160-165, whereas the OmpA of *E. coli* strain 536 has only sporadic substitutions and the OmpA of *E. coli* strain F11 almost wholly distinct. The deviations of OmpA in the uropathogenic strains from OmpA of the commensal strain were not conserved.

Discussion

PapX is not regulated in the same manner as the *pap* operon

papX resides at the 3' end of the *pap* operon, so the simplest model of regulation of PapX activity put forward was that *papX* expression would increase when *pap* expression increased, increasing PapX activity and reducing motility when adherence was up-regulated. However, using qPCR, we found that *papX* transcript levels were unaffected when expression of the *pap* operon varied significantly. This result supported

prior microarray findings (**Figure 3-1B**). Even though expression state of Type 1 fimbriae (L-ON vs L-OFF) has been shown to mediate motility in a partially *papX*-dependent manner [169], we found that *papX* expression was no different in *E. coli* strain CFT073 when Type 1 fimbirae were L-ON as compared to L-OFF by qPCR (**Figure 3-1B**). Our results suggested that *papX* has its own promoter in addition to being expressed on a transcript that includes the rest of the *pap* operon, and indeed, the region between the end of *papG* and the start of *papX* was able to drive *lacZ* expression in the $pRS551_{GX}$ construct (**Figure 3-2A-D**). In Type 1 fimbrial L-ON and L-OFF *E. coli* CFT073 backgrounds, these fusions also showed no difference in *lacZ* expression, indicating that the putative promoter before *papX* does not vary in strength of expression in a Type 1 fimbrial-dependent manner. We concluded that a model was likely in which *papX* has its own promoter, and that PapX activity is not regulated at the level of transcription. The *papX* homolog, *prsX*, shares an identical putative promoter region, whereas the other homologs of *papX* in UPEC strains, namely *focX*, *sfaX*, *fotX*, and *prfX*, contain unique putative promoter regions preceding the genes. The shared promoter elements indicate that our findings pertinent to *papX* transcriptional regulation are applicable to several members of the 17 kDa-protein family in UPEC. Future studies could be performed to determine whether or not *focX*, *sfaX*, *fotX*, or *prfX* are regulated at the transcriptional level.

This left us with the question of how PapX activity may be regulated. Previous work has demonstrated that MarR-type transcription factors can be regulated post-translationally. Our afSELEX-seq data predicts that PapX can bind at numerous regions within the CFT073 genome *in vitro*, as many non-identical fragments were enriched

during afSELEX-seq **(Figure 2-7A).** However, as mentioned, over-expression and deletion of *papX* has a transcriptionally specific effect [169]. In addition, the *E. coli* strain K12 used in Chapter 2 (**Figure 2-3**) has an identical *flhD* promoter region to that of *E. coli* CFT073, yet *papX* exerts no phenotypic effect on the motility of *E. coli* strain K12. If the predictions from afSELEX-seq are accurate and PapX is able to bind promiscuously, but only regulates the expression of a much smaller number of genes, then another factor or condition must be invoked to explain the discrepancy. One possible explanation is that a local (*i.e.*, non-diffusible) factor interacts with *papX* at certain promoters, acting as a cofactor in transcriptional regulation.

Other factors may interact with PapX

In Chapter 2 we demonstrated that $Arg^{127}Gly$ and $Lys^{54}Asn$ mutations of PapX are incapable of repressing swimming motility at 30^{o}C (**Figure 2-3**). $Arg^{127}Gly$ does not bind DNA by EMSA (**Figure 2-8**), and so presumably is incapable of targeting the *flhD* operon for repression. However, $Lys^{54}Asn$ is able to shift the same fragment of the *flhD* promoter as wildtype PapX (**Figure 3-3A**), indicating that it can bind the *flhD* promoter under certain conditions. This result suggests that there are downstream events that follow PapX binding that also influence motility, and that binding alone is insufficient to repress motility. The Lys^{54} is found within the putative dimerization domain of PapX, but we have shown that the $Lys^{54}Asn$ mutant still dimerizes. pK54N does not increase motility in WT *E. coli* CFT073, indicating that the mutation does not display a dominant negative phenotype. This domain could also potentially serve as a binding domain for

other similar proteins, so the mutation may disrupt the interaction of PapX with a cofactor. MarR family proteins may be allosterically regulated by small molecules and proteinaceous cofactors instead of, or in addition to, being regulated by transcriptional mechanisms controlling expression [176-178, 180].

To narrow the list of candidate factors, we expanded the number of strains that did or did not experience repressed motility in the presence of *papX*. Motility assays performed on uropathogenic *E. coli* strains F11 and 536 in the presence or absence of pPxWT (encoding wildtype PapX), pVector, and pK54N indicated that motility of these strains was repressed in the presence of *papX* (**Figure 3-3B**). This result helps us to narrow the candidate list of factors that may influence PapX activity. If a single unique factor explains the difference in response, then only factors common to *E. coli* strains CFT073, F11 and 536, but not present in *E. coli* strain K12, should be responsible for the difference in response to *papX*. A simple model to explain the difference in response to PapX between pathogenic and non-pathogenic strains is that one or more proteins could be cofactors for PapX.

To discover factors that interact directly with PapX, we performed a co-precipitation experiment. PapX was 88-fold enriched in the PapX group compared to the mock, confirming that there was a sharp difference in the conditions. PapX levels were the most different between mock and the PapX group, suggesting that the differences between groups can be most directly attributed to PapX itself. 21 Proteins were enriched in the PapX group compared to the mock group, indicating that we had concentrated factors that are associated with the presence of PapX. Of those proteins, only one, OmpA,

was enriched that was common to *E. coli* strains CFT073, F11 and 536, but was not identical in *E. coli* strain K12. OmpA has been tenuously linked to motility and flagellar rotation, so it is possible that it serves as a cofactor or is acted upon by PapX to produce a phenotype [198].

OmpA is also involved in acid stress response [199], which is also known to affect motility, so it is plausible that OmpA could be an environmental sensor coupled to PapX. Acid stress may be related to motility because it affects proton motive force and proton transport across the membrane, which is necessary for flagellar rotation [147]. OmpA is an appealing candidate because it has been shown that *ompA* is one of the most highly up-regulated genes during UTI [38]. Unfortunately, non-identical regions of OmpA of uropathogenic strains and OmpA of the commensal strain are not completely conserved. If changes were uniquely linked to PapX, we would expect some of the changes to remain consistent between strains. However, given that single amino acid substitutions in PapX are sufficient to abolish activity, it is possible that the differences in OmpA sequence are meaningfully linked to activity. It is possible that any disruption of OmpA is sufficient to change its interaction with PapX. It is also possible that uropathogenic strains express OmpA at different times or to different levels than commensal strains of *E. coli*. Future work could consider examining the effects of pPxWT on strains lacking OmpA and assaying for an impact on motility. OmpA from a uropathogenic strain could also be transformed into a commensal strain of *E. coli*, and motility could then be assayed.

OmpA is an outer membrane that acts as a large porin, capable of transmitting many compounds. If PapX affects OmpA activity, one might hypothesize that compounds transmitted into the periplasm and intracellular space by OmpA (such as ions or small molecules) could influence PapX binding behavior, impacting downstream PapX activity negatively or positively. However, OmpA is located at the outer membrane, whereas PapX is cystosolic. Cytosolic proteins are typically separated from direct interaction with outer membrane proteins by the cell membrane, and this physical separation of compartments raises a question over how these proteins might interact. Our co-purification study should assay for direct interactions, so either recombinant PapX-His_6 interacts with the outer membrane, or it could be interacting with OmpA as it is being synthesized in the cytosol. I have already presented evidence that PapX is a direct DNA binding protein and that DNA binding is important to activity, so the most parsimonious model is that PapX interacts with OmpA in the cytosol. Only small quantities of incomplete OmpA peptide should be present in the cytosol relative to the outer membrane, perhaps explaining why so few spectra of OmpA were seeing in the co-purification. One hypothesis is that PapX may affect OmpA folding or trafficking to the membrane, which in turn affects *E. coli* cell membrane permeability to outside compounds, which themselves interact with PapX to modulate activity.

Further experimental investigation of this model is required. Studies investigating the affinity of PapX for OmpA could be performed using a yeast two-hybrid screen. Our co-purification is performed *in vitro*, and thus falls prey to the pitfall of potentially not representing the biologically relevant and physiological conditions of a bacterial cell. Alternatively or in parallel to the *in vitro* co-purification, antibodies to PapX could be

synthesized and an *in vivo* co-precipitation could be done to attempt to discover proteins interacting with PapX under native conditions. Preliminary results suggest that PapX is not present in large quantities within the cell, so this approach could have the advantage of not misrepresenting stoichiometry of native interactions. However, in turn, the quantity of starting material to be co-precipitated would have to be large so that sufficient sample could be analyzed by LC-MS/MS or other protein identification and quantitation methods.

It is possible that no single factor can explain the difference in responsiveness between commensal and uropathogenic strains of *E. coli* in response to PapX. If a group of factors act in tandem, or if a scaffold that associates factors together is causative, for example, then the results of a co-precipitation will be more difficult to interpret. Proteins and factors may be identical between uropathogenic strains and commensal strains, but the ratios, combinations, or other conditions (such as pH or ion concentration) in the cytosol may be the explanatory factor. These would not have appeared as single enriched species in a co-precipitation. In addition, though we made efforts to generate stoichiometrically physiologic ratios of PapX to other *E. coli* strain CFT073 cystosolic proteins, expression is often difficult to balance precisely. PapX in excess of native levels may bind promiscuously or saturate binding too quickly on a physiologically relevant target. We know from past experience that PapX is prone to precipitating out of solution at inconvenient concentrations and times, which suggests it may bind non-specifically under the proper conditions.

CHAPTER 4

A NOVEL APPROACH FOR TRANSCRIPTION FACTOR ANALYSIS USING SELEX WITH HIGH-THROUGHPUT SEQUENCING (TFAST)

Introduction

Systematic Evolution of Ligands by Exponential Enrichment (SELEX) is a technique for determining nucleotide binding sites of transcription factors ([183, 184] and reviewed in [185]). SELEX is an iterative method where the products of one cycle are used to generate the input for the next (**Figure 4-1A**), enriching strongly binding sequences in the output. In the past, only the terminal cycle was analyzed, and typically 50 or fewer members of the cloned output library were sequenced, which typically meant that SELEX had to be performed so that the vast majority of the final library consisted of a single or very few species [183, 184]. However, doing so runs the risk of eliminating physiologically relevant, but less strongly binding, DNA targets. High-throughput sequencing has made SELEX-seq possible, in which intermediate cycles are analyzed and millions of members of each cycle are sequenced. Using SELEX-seq, it is possible to identify the cycle at which all non-binding species have been selected against (when the background frequency drops below that of the initial library), and to halt iterative cycles before potentially biologically important species have been eliminated from the pool. In this way, SELEX-seq permits the accumulation of many more potential target species

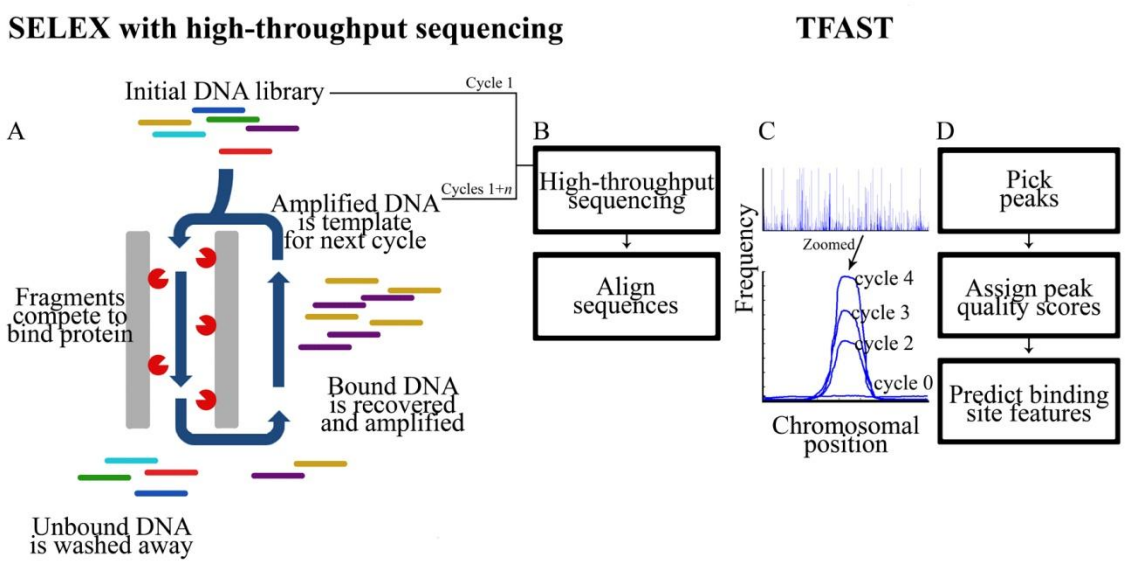


Figure 4-1. Schematic of the TFAST workflow. TFAST analyzes data produced using SELEX and high-throughput sequencing. (**A**) An overview of SELEX. Members of an input DNA fragment library compete to bind a protein of interest. Out-competed fragments are washed away and removed. Fragments that bind competitively are recovered and separated from the protein (*e.g.,* by phenol-chloroform extraction). Recovered fragments are amplified using low-cycle PCR, and the resultant library becomes the input for the next cycle. *n* cycles are repeated to enrich for strongly binding fragments. (**B**) DNA inputs for each cycle are subjected to high-throughput sequencing. Sequence reads are aligned to the relevant target genome, producing (**C**) frequency-position plots for each input sequenced. Shown is a magnified example of a region of the chromosome that behaves as a true binding site, enriching in frequency with each cycle. "Cycle 1" refers to the initial DNA library. (**D**) TFAST identifies and evaluates peaks and compares peak characteristics across all sequenced inputs. TFAST assigns quality scores to peaks and predicts binding site features. TFAST also generates quality scores on randomly selected regions of the chromosome to act as background controls, to improve downstream motif analysis.

than SELEX as it has been classically performed. In addition, SELEX-seq represents an improvement over SELEX in the detection and analysis of binding sites because behavior of the library can be scrutinized in million-fold greater detail.

The elevated level of detail in SELEX-seq compared to SELEX is similar to the increase in detail between terminal (discrete) transcriptional profiling, such as PCR, and real-time (continuous) methods, such as qPCR. Determining DNA binding sites has traditionally been a labor-intensive and intricate process. The improved resolution of SELEX-seq over SELEX represents an increase in dynamic range and a concurrent reduction in type II error, which improves confidence in the results of SELEX-seq analysis and makes SELEX-seq less subject to *in vitro* biases because library behavior can be tracked across cycles, so loss of binding candidates can be tracked to specific environments. The resulting increased level of detail allows the results of SELEX-seq to be used for inferences about binding site affinity and length, where SELEX can only approximate species with discrete high or low affinity, and cannot speak to site length or genomic context.

SELEX-seq can detect the presence of library members and quantify their representation, making it possible to estimate binding affinities of many sequences and to compare sequences to one another to generate motifs [186-189]. This advance also obviates the need to run the protocol until only a single sequence predominates in the terminal cycle. In fact, successful SELEX-seq experiments depend on multiple species being detectable in the terminal sequenced cycle to accurately infer binding characteristics [186].

Usually, the library of DNA to be enriched consists of random-sequence 10-20 bp oligonucleotides, called aptamers. This type of library is especially useful when a specific subset of aptamers is itself the desired end-product. Inter-cycle computational comparators (see [186, 187, 200] for examples**)** utilize the iterative nature of SELEX-seq to calculate binding affinity and activity of individual aptamers or aptamer sets, which is useful in designing therapeutic aptamers [201-203]. However, use of aptameric SELEX-seq for discovery of genomic binding sites may be problematic. Aptameric libraries are by definition random, so a given sequence may not be found within a target genome and thus discovered sequences may lack physiological relevance. Aptamers also lack genomic context, so that binding behavior may obscure *in vivo* findings by too strongly reflecting *in vitro* conditions [204].

To address these potential pitfalls of SELEX-seq, we designed an aptamer-free SELEX-seq protocol (afSELEX-seq) that uses sheared genomic dsDNA as the input library, with which we successfully identified a novel and single-copy transcription factor binding site [188]. Our approach incorporates the advantage of multiple-round enrichment with a physiologically relevant target library. In afSELEX-seq, the results of every sequenced cycle are aligned to a target genome (though not every cycle needs to be sequenced), and alignments are compared between rounds to predict binding sites. To process our results, it was necessary to develop software capable of both analyzing chromosomal alignments and acting as an inter-cycle comparator. Transcription Factor Analysis using SELEX with High-Throughput sequencing (TFAST) was written in Java, and once sequences have been aligned to the genome of the target organism (for example, with BLASTn [205] or BOWTIE [206]), TFAST completes analysis in minutes on a

personal computer. TFAST was designed to be useful to a broad range of biologists, so it employs an uncomplicated graphical interface to streamline its use. It is easy to install, requiring no special expertise in bioinformatics.

Materials and methods

Peak identification

afSELEX-seq is performed using sheared genomic fragments, and the products of each of multiple cycles are subjected to high-throughput sequencing. Sequence data are aligned to the target chromosome using either BLASTn or any alignment method capable of producing output in SAM format (**Figure 4-1B**). TFAST processes the same number of aligned reads from each cycle, thus reducing the impact of variability in sample read quality and completeness, to generate single column tables representing frequencies of aligned tags at every position within the target genome. TFAST then uses a modified sliding window algorithm (see [207] for a review of sliding window algorithms) to identify peaks in the final round of afSELEX-seq (**Figure 4-1C**). The sliding window width is set to twice the estimated fragment length of an average library member and proceeds along the chromosome in single-nucleotide increments to identify strict maxima. To be considered peaks, maxima must be greater than two standard deviations above the mean frequency of the control (unselected) cycle within a range around the peak (distance set by user). Minimum spacing to identify individual peaks is set at a single fragment length, with the exceptions that deep valleys (<50% adjacent peak value)

between peaks or adjacent peaks within 75% frequency of one another are both delimited as separate.

Peak scoring

Frequencies at each peak position are compared across all cycles of afSELEX-seq, allowing the program to score peaks based on rate of enrichment as a proxy for binding affinity consistent with existing models of enrichment during SELEX (**Figure 4-1C and D**) [185, 186, 189, 208, 209]. Sequences are given a cumulative score for each round in which they are able to enrich, so that sequences that remain strongly competitive for binding throughout afSELEX-seq achieve the highest scores. To reflect the increasing stringency of competition for binding in later rounds, the user should set scores for enrichment between cycles to be directly proportional to the bulk affinity of the library within each cycle. This causes TFAST to weigh later-round (*i.e.,* more stringent) enrichment more heavily than early-round enrichment, improving the discrimination of strongly-binding sequences.

TFAST then generates information on the relative representation of each peak, annotated as a raw final frequency and fraction of the area under a given peak in comparison to the total area under all other peaks in that cycle. TFAST also prints the genomic sequence under each peak (range determined by user) and outputs all data in tab-delimited format for downstream analysis. This format is designed to dovetail smoothly with methods for the discovery of DNA binding-site motifs. TFAST also automatically produces and analyzes sets of random sequence positions denoted “spoof”

peaks, which can be used to improve motif discovery through counter-selection and background modeling. TFAST does not generate negative peaks, which are sometimes used to refine motif discovery, because counter-selection in afSELEX-seq produces a background of zero frequency between peaks. This means that negative peaks are artifacts of variation in control cycle sequencing and not indicative of especially poor regions of protein:DNA interaction.

Binding site sequence length prediction

TFAST is designed to estimate the length of a putative region of protein:DNA interaction. Members of a dsDNA library should only have specific affinity for the protein of interest if they include the binding sequence. The full width of any peak over a single binding site will be twice the length of an average library member less the length of the binding site. TFAST calculates the full width of a peak as the range along the chromosome that includes 99% of the area under the peak, and then uses the width to calculate a predicted length for any binding site. Predicted length is included with the output of peak features.

Results and discussion

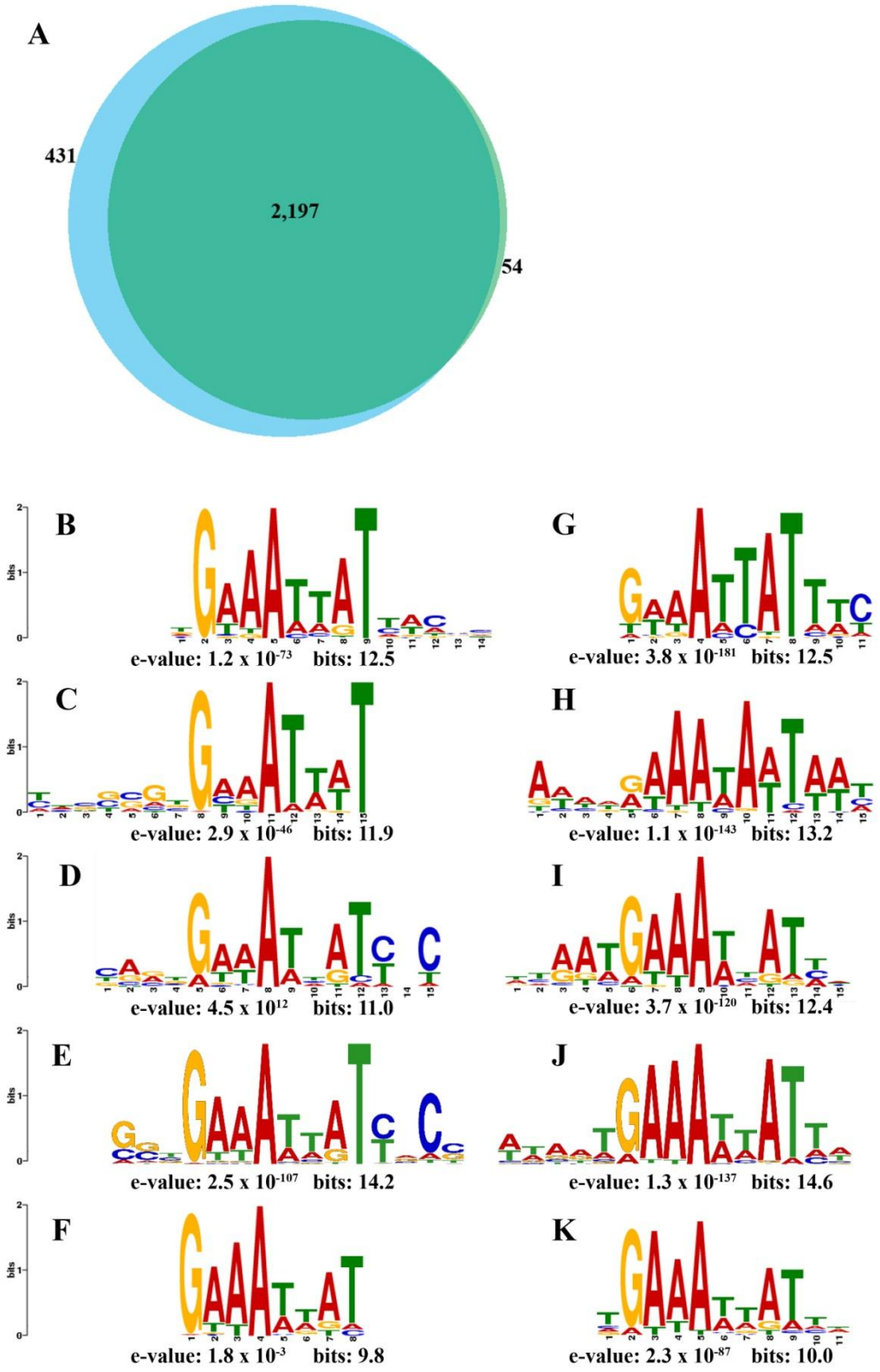
TFAST accurately discriminates peaks in afSELEX-seq data

To validate our peak calling method, we compared peak detection between TFAST and MACS [191] in an afSELEX-seq dataset generated in prior work [188]. We chose MACS because it is an established and well-vetted peak finding program designed for use with aligned high-throughput sequence data. TFAST called 96% of regions identified by MACS as peaks, and subdivided large peaks into multiple smaller ones (**Figure 4-2A**). In detection of enriched regions, TFAST and MACS had a simple agreement of 0.98 and a κ value of 0.89, values generally considered to indicate near perfect agreement. [Footnote: The simple agreement of 0.98 and the κ of 0.89 between MACS and TFAST were calculated under the simplifying assumption that the 5,231,428 bp genome of *E. coli* strain CFT073 is divided into 26,157 candidate peak regions of 200 bp. Analysis assuming 13,079 candidate peak regions of 400 bp yields a simple agreement of 0.96 and a κ value of 0.88.]

TFAST peak grading correlates with informational density of discovered motifs

When TFAST was used to process our dataset, 2,628 total peaks were identified, of which 457 achieved the highest weight ("Best" peaks), or 17.4% of the total peaks called. When peaks are called without using the last or last two enrichment cycles, 4,255 and 7,274 peaks are identified, respectively, with a concurrent loss in specificity in identifying the "Best" peaks (31% and 100%, respectively) (**Table 4-1**). This demonstrates TFAST's unique advantage of using iterative cycles over simply the final set of peaks. Additionally, we instructed MEME [192] to discover 15 bp positional weight matrices (motifs) from peaks picked by TFAST or MACS (**Figure 4-2B-K**).

Figure 4-2. TFAST identifies peaks with discoverable motifs from afSELEX-seq data. TFAST and MACS were used to pick and evaluate peaks from our data set. (**A**) TFAST picked a total of 2,628 peaks, of which 2,197 covered 96% of the peaks identified by MACS in the final cycle of afSELEX-seq. Positional weight matrices generated in MEME instructed to search for a 15 bp motif using 200 sequences from (**B**) the 457 "Best" weight (most enriched) peaks, (**C**) the 888 next-best weight (second most enriched) peaks, (**D**) the 1,283 worst weight (least enriched) peaks, (**E**) all peaks called by TFAST pooled together and (**F**) 200 peaks called by MACS with the lowest false discovery rate (FDR). Sets of peaks from (**B-F**) were subjected again to analysis by MEME under the similar conditions but with the inclusion of a zero-order background Markov model to generate (**G-K**). E-value (the chance that a motif arose from a dataset by chance) and bit score (the total information content of a positional weight matrix) are shown below each logo.


A
431
2,197
54
B
e-value: 1.2 x 10-73 bits: 12.5
G
e-value: 3.8 x 10-181 bits: 12.5
C
e-value: 2.9 x 10-46 bits: 11.9
H
e-value: 1.1 x 10-143 bits: 13.2
D
e-value: 4.5 x 1012 bits: 11.0
I
e-value: 3.7 x 10-120 bits: 12.4
E
e-value: 2.5 x 10-107 bits: 14.2
J
e-value: 1.3 x 10-137 bits: 14.6
F
e-value: 1.8 x 10-3 bits: 9.8
K
e-value: 2.3 x 10-87 bits: 10.0
bits

Peaks with the best, second best and worst weight scores identified by TFAST had e-values (an estimate for likelihood of a motif arising by chance) of 1.2×10^{-73}, 2.9×10^{-46} and 4.5×10^{12}, respectively, and informational content of 12.5, 11.9 and 11.0 bits, respectively (**Figure 4-2B-E**). This indicates that TFAST was successful at stratifying classes of peaks over consistent sequences in the genome, and that positional weight matrix reliability was greatest in the "Best" peaks.

All peaks discovered by TFAST pooled together generated a positional weight matrix with an e-value of 2.5×10^{-107}, an intermediate value that was less significant than "Best" peaks selected by TFAST (**Figure 4-2E**). Furthermore, pooled peaks picked by MACS produced a logo with an e-value of 0.54, which was not significant. Using the 200 MACS peaks with the lowest false discovery rate (FDR), MEME was able to generate a positional weight matrix with an e-value of 1.8×10^{-3} and informational content of 9.8 bits (**Figure 4-2F**). No other subset of MACS peaks (*i.e.*, peaks with lowest p-value, greatest fold enrichment, etc) generated a logo with an e-value less than 1. This is a substantially larger e-value than that generated from the peaks discovered in TFAST, likely because the average width of peaks picked by MACS was 503 ± a standard error of 9.71, whereas all peaks picked by TFAST were under 200 bp wide. These findings support the method employed by TFAST of picking many narrow peaks based on absolute local maxima, coupled with fragment length, rather than statistical modeling of aligned sequences.

Table 4-1. Incorporation of increased SELEX cycle number improves peak discrimination

Number of cycles used[a]	Total peaks called[b]	Number of "Best" peaks[c]	Percent "Best" peaks of total
3	**2,628**	**457**	**17.4**
2	**4,255**	**1,350**	**31.7**
1	**7,274**	**7,274**	**100**

a. "Number of cycles used" does not include control

b. "Total peaks called" refers to peaks called using TFAST's algorithm

c. "Best" peaks are peaks with the highest weight scores for that set, as calculated by TFAST.

Background models derived from TFAST improve motif discovery

To validate the use of spoof peaks (described in **methods**) to improve motif discovery, we generated a zero-order Markov model from the spoof output of TFAST. When we incorporated the model into our motif discovery using MEME, the information content of the motifs rose by an average of 0.78 bits or 6.3%, and e-value fell by an average of 8.2×10^{-133} for peaks picked by TFAST (**Figure 4-2G-J**). Additionally, the Markov-corrected 15 bp motif from the "Best" weight sequences picked by TFAST most accurately predicted the 15 bp core of the previously validated binding site GTTATTTTAAC [188]. Use of the Markov model also strengthened the motif generated from MACS, improving information content from 9.8 to 10.0 bits or 2.0% and the e-value from 1.8×10^{-3} to 2.3×10^{-87} (**Figure 4-2F, 4-2K**). Overall, this supports the notion that using an accurate background model generated from TFAST can improve the ability to discover significant motifs in downstream peak analysis, and improves total motif informational content.

TFAST can accurately predict binding site sequence length

In our data set, TFAST predicted a binding site width for the "Best" peaks of 20.88 bp ± a standard deviation of 10.33 bp, consistent with the motifs discovered. Additionally, TFAST predicted the binding site for the peak in the *flhD* promoter of CFT073 for PapX to be 28 bp. Experimentally, PapX binds a 29 bp fragment of the *flhD* promoter, but not a truncated 21 bp fragment [188]. The consistency between the

prediction made by TFAST and the experimentally verified binding site indicates that TFAST is capable of predicting binding site length from afSELEX-seq data.

Advantages of afSELEX-seq over SELEX and SELEX-seq

afSELEX-seq has several advantages over SELEX and SELEX-seq. Instead of using an aptameric library [185], our library was constructed directly from sheared CFT073 genomic DNA, generating paired-end ligated 80-100 bp fragments. A typical SELEX-seq aptameric library contains 10^{13}-10^{15} members of 10-20 nucleotides in length [210], whereas our library contained only 5 x 10^6 unique members (and 10^6 of each unique member per cycle of SELEX), each 80-100 bp in length, resulting in a high-stringency selection throughout. Because we had substantial evidence that the putative transcription factor under investigation, PapX, exerted its effect on genes within the CFT073 genome [169] (**Figure 2-3**), the best binding site was, of course, likely to be present in this library. Additionally, because aptameric libraries are usually generated at random and aptamers may be shorter than a complete binding site, a problem that SELEX and SELEX-seq encounter is questionable physiologic relevance of a determined motif [209]. Therefore, a maximally bound motif may not actually be present in any organism. afSELEX-seq circumvents this weakness. In addition, some physiologically relevant consensus sequences rarely appear as perfect matches; for example, the Shine-Dalgarno sequence, a ubiquitous ribosomal binding site found in *E. coli*, frequently appears as an imperfect, 4/6 or 5/6 match, yet still remains functional. In fact, varying the degree to which a given Shine-Dalgarno site matches the consensus is a means of regulating

translation [211]. Indeed, any positive hits from our library should be, by definition, present in the CFT073 genome and are thus more likely to be physiologically relevant.

Additionally, because afSELEX-seq uses longer probes than SELEX-seq (80-100 bp and 10-20 bp, respectively), information about neighboring binding sites is retained in afSELEX-seq that might have been lost in SELEX-seq. For example, if the transcription factor under investigation frequently binds in the proximity of other associated chromosomal target regions of other factors, those sites may be revealed and provide insight into novel interactions between that transcription factor and other species. It is more difficult to precisely contextualize aptamers in an organism's genome because their short length reduces the strength of an alignment.

To identify target sequences using SELEX and high-throughput sequencing, a sufficient coverage of the target genome must be used in the initial library. For each cycle of SELEX, roughly 3×10^{12} 80-bp fragments (or 2.4×10^{14} nucleotides) were exposed to immobilized protein. The quantity of fragments selected suggest that this technique could be adapted to eukaryotic genomes and still preserve hundred-thousand ranges of representation at each nucleotide position in the probed sample. The 2.5×10^{7} reads per lane generated in high-throughput sequencing limits the use of the control library in establishing quality scores for peak enrichment. However, the peak scoring algorithm applies between enriched rounds as well as between the control library and the first enrichment, and so a few rounds of SELEX should be sufficient to generate initial peaks from which to base subsequent scoring in a eukaryotic schema.

Conclusions

Here we present TFAST, an easy-to-use tool for rapidly and accurately analyzing data generated by afSELEX-seq to discover and characterize transcription factor binding sites. Currently, no other software is designed to analyze afSELEX-seq. The peak finding component of TFAST compares favorably to that of MACS. For use with afSELEX-seq, TFAST outperforms MACS in generating significant motifs. The scheme TFAST employs for grading and analyzing peaks uniquely leverages the cyclical nature of SELEX and direct protein:genomic dsDNA interactions to accurately and sensitively predict binding site sequence and length, as demonstrated by the consistent concurrence between predictions generated by TFAST and our experimental findings [188].

afSELEX-seq offers an alternative to current methods of discovering DNA binding sites, and requires no antibody generation (unlike ChIP-seq (reviewed in [181] and references therein)) or complex biological screens (unlike bacterial 1-hybrid systems [190]). Large libraries of purified, tagged bacterial proteins of unknown function currently exist at institutions participating in the Protein Structure Initiative [212]. This is sufficient material to run afSELEX-seq, and with only minor modifications the binding sites of hundreds if not thousands of transcription factors could be quickly elucidated.

CHAPTER 5

CONCLUSIONS AND FUTURE DIRECTIONS

Summary of results

The studies presented here address the role of PapX in repressing motility as a component of reciprocal control of motility and adherence. Specifically, we investigated the mechanism by which PapX exerts its effects, and made progress in identifying mechanisms by which PapX may be regulated. The major findings of this work are described briefly:

*****PapX and its numerous homologs are commonly present in ExPEC and UPEC strains.

*****Uropathogenic strains of *E. coli* encode *papX*, but not the close homolog *focX*, more commonly than commensal strains of *E. coli*, suggesting a role in urovirulence for *papX*.

*****The effects of PapX are pathogen-specific, indicated by its suppression of

motility in uropathogenic strains of *E. coli* but not commensal strains of *E. coli.*

*****PapX is dependent on Arg^{127} and Lys^{54} for its activity.

*****PapX binds to a specific sequence of DNA found uniquely within the *flhD* promoter of the CFT073 genome.

*****Binding of PapX to the *flhD* promoter appears to be necessary, but not sufficient, for its activity, as indicated by the fact that an Arg^{127}Gly mutant of PapX is unable to bind DNA and unable to repress motility of *E. coli* CFT073, and that a Lys^{54}Asn mutant binds to the *flhD* promoter, but does not repress motility.

*****PapX activity is not regulated at the level of transcription of the rest of the *pap* operon.

*****PapX likely interacts with environmental and metabolic triggers or factors.

Conclusions and perspectives

In this work, we elucidated the mechanism whereby an adherence-associated factor, PapX, negatively regulates motility. To answer the fundamental questions of how PapX is able to accomplish this feat, we deployed novel computational approaches and engineered software designed to be useful for any bench-top scientist investigating related fields of study. The findings here expand and enrich our understanding of reciprocal control of motility by adherence, and add valuable new tools that can help illuminate the unresolved questions about protein:DNA interactions at a molecular level of detail. We employed a combination of bench and computational techniques to answer

questions of potentially broad significance within the world of microbiology, adherence, and motility.

UPEC strains can encode up to 15 epidemiologically and genetically distinct clusters of fimbriae, some of which are important components of UTI pathogenesis [61]. Our work has supported observations that the clusters encoding fimbriae exhibit cross talk with other systems within the bacterium. Other work has indicated this as well, such as the observation that PapB, a regulatory protein involved in the expression of the *pap* operon, suppresses Type 1 fimbria from remaining in the L-ON state [111, 213]. FocB, of the *sfa* operon in meningitis-causing *E. coli*, also shares this property, and the binding site of FocB was recently demonstrated crystallographically [214]. Interactions between various systems, both within adherence-encoding systems and without, suggest that the interplay between fimbriae, adherence, and numerous regulatory networks within the cell are extensive and merit further scrutiny.

The conditions under which the numerous other fimbriae of UPEC strains are expressed remains an active field of inquiry. Why UPEC encodes so many fimbriae is an interesting question to which there is currently no definitive answer. However, prior data has linked the presence of fimbriae on bacteria in the context of an appropriate substrate to clinically significant phenotypes, such the reduced urovirulence of UPEC strains in mice under the influence of pilicides [215], or the complementable defects in urovirulence when specific fimbriae are knocked out [51-54]. In addition, the receptors for fimbriae have been identified in the environments in which those fimbriae are thought to influence urovirulence, such as uroplakin Ia or P blood group antigen [73, 126-128]. Results of this type suggest that the other, as-yet uncharacterized fimbriae of UPEC

strains must also have substrate specificity, and thus imply the existence of environmental niches in which their presence or absence meaningfully contributes to UPEC survival or phenotype.

There is good reason to believe that UPEC require many kinds of adherence organelles, and that UPEC must regulate them carefully. Even within the urinary tract, UPEC faces different environmental niches. Studies have indicated that even the variable oxygen content can affect *pap* expression [38], and within perturbed urinary tracts there are obvious changes in urine characteristics, such as glucose levels in diabetes or protein levels in kidney failure. Osmolarity of urine, too, appears to be relevant to UPEC pathogenesis as OmpA, an outer-membrane osmolarity regulator, has been shown to be important for uropathogenesis [216]. Osmolarity of urine does not even remain consistent over the course of hours, and so we expect the genes involved in environmental responsiveness within the urinary tract and beyond to be precisely and coordinately regulated, to be expressed when needed and to incur minimal costs to avoid the disadvantage of expressing proteins that will not assist in (or may even hinder) survival. As UPEC does not disseminate from one urinary tract to the next, there are obligate intermediate niches through which it must pass, and for which UPEC must be able to adapt in short order.

One area in which some inference can be made about the role of uncharacterized fimbriae is that of fimbrial operon expression. Currently, we lack knowledge of the conditions that induce many fimbriae to express. Data do suggest that not all fimbriae are expressed simultaneously, and, as mentioned, some fimbria appear to antagonize others, such as P fimbria and Type 1 fimbria [112]. The triggers that regulate the expression of

fimbrial systems remain largely unknown, however, and exploring these regulatory pathways may help to further our understanding of a vast, unknown aspect of UPEC environmental and pathogenic existence. The persistent existence of many kinds of fimbriae in UPEC, especially as compared to commensal strains [194], suggests that these fimbriae contribute to urovirulence. UPEC strains do not only encode more fimbriae than commensal strains, but each individual UPEC strain tends to carry a vast array of fimbriae.

As it appears that only a few fimbriae are expressed highly *in vivo* [106] and thus putatively related to urovirulence within the urinary tract, suggesting that there are extra-urinary-tract roles for these fimbriae, and that those roles are conserved among uropathogens. We do know that other ExPEC strains rely on different adherence profiles to attain full virulence, such as meningitis strains of *E. coli* seemingly depending on *sfa*-encoded fimbriae in certain aspects of their virulence cycle [217]. Even in the case of the distinct meningitis-causing *E. coli* strains of ExPEC, an apparently similar regulatory mechanism is conserved among UPEC strains, in the form of the SfaX protein which shares many characteristics of PapX [213]. It would be logical that the mechanisms of crosstalk and fimbrial regulation were conserved across ExPEC strains, and the furthering of our knowledge of the mechanism of a UPEC strain in regards to this specific pathway should improve our understanding of the influence of adherence on other characteristics and pathways across a broad range of ExPEC.

Given the numerous uncharacterized adherence systems, and the likelihood that they are influenced by environment and play roles in the UPEC cycle of dissemination and pathogenesis, we believe that our tools for rapid characterization of protein:DNA

interaction are exciting and highly useful, rife with potential. The utility of our approach is that it is broadly adaptable to any organism and permits the rapid discovery of meaningful interactions. In recent years, computation and molecular techniques in biology have grown in complexity and nuance. The sheer volume of data generated in any given molecular biology experiment has expanded exponentially. The increase in data volume attends a need for useful tools for analysis, and questions of DNA:protein interaction are ideally suited to be addressed using these techniques. Characterizing new fimbrial systems is exciting because it may lead to a rapid increase in knowledge of UPEC physiology and biology. Better understanding of UPEC biology will hopefully lead to targeted therapies and may help to structure approaches for disrupting UPEC virulence. If a detailed enough model of UPEC fimbrial interactions can be generated, it is possible that novel treatment methods could be designed, and that perhaps subtle variations in environment could reduce UTI substantially. SELEX-seq itself has been used to design therapeutic aptamers [201-203] and so it is even possible that by understanding the molecular mechanisms of fimbrial interactions, we may advance the design of therapeutic aptamers against UPEC.

Future directions

Two related questions focused on PapX beg further investigation. Subsequent to binding, how does PapX effect suppression of motility, and how is PapX activity regulated? Preliminary data indicate that the answers to these questions may be complementary.

afSELEX-seq indicated that PapX may bind to several degenerate sites *in vitro*. However, there is a transcriptionally specific effect of *papX* knockout and overexpression as illuminated by microarray, though the microarray was performed only under a single set of conditions. Additionally, *papX* is able to influence the swarming and swimming motility of *P. mirabilis* HI4320, but this strain contains in its *flhD* promoter a degenerate version of the binding site found within the *E. coli* strain CFT073 *flhD* promoter. If the microarray results are correct in revealing a transcriptionally specific effect of PapX, then further explanation is required to resolve this puzzle. Some answers may be found in the results of our co-precipitation, which indicate OmpA may be a candidate to explain the uropathogen-specific nature of the repression of motiliy by PapX. To further examine and investigate the downstream effects of PapX binding, this candidate gene should be knocked out in *E. coli* strain CFT073 and the effect of PapX on the motility of the resulting contruct should be examined to determine whether or OmpA plays a role in the repression of motility by PapX. In parallel to determine whether or not OmpA can make K12 responsive to the repression of motility by PapX, the OmpA encoded by *E. coli* CFT073 should be inserted into the chromosome of *E. coli* strain K12 and to determine whether or not either or both of these genes can make K12 responsive to the repression of motility by PapX.

Concurrently, a genetic approach could be taken to seek out factors encoded by the *E. coli* strain CFT073 chromosome that might cause it to be sensitive to PapX. Currently, there exist reagents to insert *fliC-lux* reporter fusions into the chromosomes of *E. coli*. These reagents were necessary to perform the biophotonic imaging studies to examine the expression of *fliC* as a surrogate for flagella and motility in *E. coli* strain

CFT073 [59]. These same reagents could be used to generate *fliC-lux* reporter fusions in the chromosome of *E. coli* strain K12 ($K12_{fliC\text{-}lux}$). That construct could then have *papX* inserted into its chromosome ($K12_{flic\text{-}lux\text{-}papX}$). One could then transform $K12_{flic\text{-}lux\text{-}papX}$ with a cosmid library of *E. coli* strain CFT073's chromosome and screen for colonies that demonstrate reduced luminescence as a marker for reduced *fliC* expression. The technique could be repeated with smaller fragments or only genes identified on the sequenced cosmid clones. Candidates would be discarded if they suppressed luminescence in $K12_{fliC\text{-}lux}$ because the motility-suppressing effects of those genes would presumably be independent of the presence of PapX. Colonies in which PapX was specifically attributable as the factor suppressing the expression of *fliC* could identify a cofactor or cofactors crucial for full PapX activity and may identify the salient difference between *E. coli* strains K12 and CFT073 when it comes to PapX activity.

PapX affects both swimming and swarming motility in *P. mirabilis* HI4320 [218]. Prior data and the work presented herein scrutinized the effect of *papX* on swimming motility in *E. coli* strain CFT073, but recent preliminary data indicate that PapX also influences swarming motility in *E. coli* strain CFT073 **(Figure 5-1**). Interestingly, it appears that the homlog *focX* also influences swarming motility. Though this is unsurprising because the sequences of *focX* and *papX* are nearly identical, we had not pursued investigation of *focX* for several reasons. First, it was not identified in a screen for genes that de-repressed motility in Type 1 fimbrial L-ON strains [60]. Second, *focX* mutants of *E. coli* strain CFT073 have only a slight effect on the motility of the resulting contructs. Third, single knockouts of *papX* have a phenotypic and readily discernible genetic effect on *E. coli* strain CFT073, so we believed that the two proteins, PapX and

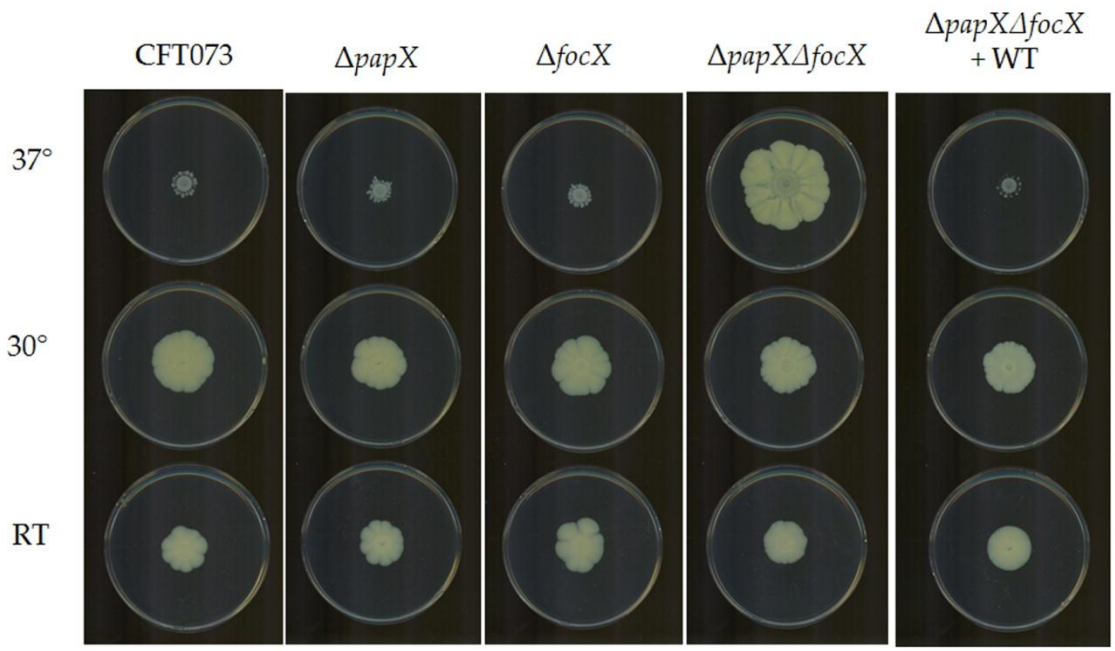


Figure 5-1. ***papX* and *focX* affect swarming motility in *E. coli* in a temperature-dependent manner.** Single *papX*, *focX*, and double *papX focX* mutants of *E. coli* CFT073 were assessed for swarming motility at 23^{o}C, 30^{o}C and 37^{o}C. Preliminary results suggest that *papX* and *focX* are each able to repress swarming motility at 37^{o}C, as indicated by the increase in swarming motility when both are deleted. *papX* and *focX* had no discernible effect on motility at 23^{o}C and 30^{o}C. *papX* encoded on pPxWT was able to repress motility of the double *papX focX* knockout strain of *E. coli* CFT073.

FocX, were not perfectly or even highly redundant. Finally, our molecular epidemiological data from Chapter 2 (**Figure 2-2C, E**) indicates that the coding status of *focX* does not appear to track with pathogenicity of *E. coli* strains. The *focX* homolog, *sfaXII*, does appear to be important for the pathogenesis of another ExPEC strain, however [213], and our preliminary data indicate that *focX* is important to the swarming behavior of *E. coli* strain CFT073. The effect of *focX* and *papX* together on swarming is temperature-dependent, a feature that has never been described for either protein. This exciting result may also help to shed light on the questions of how *papX* is regulated and what factors are involved in its activation and suppression of motility. Temperature dependence may relate to the question of why uropathogenic strains, but not commensal strains, of *E. coli* demonstrate repressed motility in response to PapX. An intermediary temperature-dependent process may be invoked, or potentially PapX is itself altered by temperature For example, the affinity of PapX for DNA may be altered under varying temperatures, and indeed, I observed that PapX had higher affinity for *E coli* CFT073 genomic DNA during SELEX at 37^{o}C as compared to 30^{o}C and 23^{o}C (unpublished observation) as reflected in a larger portion of the library fragments recovered in outputs at higher temperatures than in the lower temperatures in the presence of PapX-His_6 compared to the input libraries. Though single *papX* knockout mutants showed no effect on urovirulence of mutant strains, it could be productive to repeat the study using double *focX papX* knockout strains of *E. coli* strain CFT073. Additionally, transcriptional profiling of *flhD* and *flhC* could be repeated in the double-kockout background to eliminate any lingering background effect of *focX* on expression. This experiment has not yet been repeated and so preliminary data on that subject are unreliable.

That the relationship between *papX* and swarming motility may be temperature-sensitive may be used to further characterize and investigate the downstream effects of PapX binding. The swarming assay may provide an ideal system in which to examine the $Lys^{54}Asn$ mutant of PapX. Swarming assays should be performed as described in Appendix A comparing the effects of pPxWT, pVector, and pK54N on *E. coli* strains CFT073 and K12, with a careful eye to the effects of pK54N on the Δ*papX* Δ*focX* mutant strain of *E. coli* CFT073. Transcriptional data (microarray and qPCR) were both gathered at $37^{o}C$, but if PapX activity is influenced by temperature, as indicated by the preliminary swarming data, then transcriptional analysis should be repeated, varying conditions that are known to influence swarming motility. These might include temperature at which samples are collected, the type of broth in which samples are prepared (LB or tryptone broth), the salt concentration under which samples are collected, and potential samples should be taken directly from swarming plates (center and edge portions of a swarm). These types of transcriptional profiling experiments could be performed using wild-type PapX as well as the $Lys^{54}Asn$ mutant. Other mutants also had an influence on PapX activity (**FIGURE 2-4C**), and these, too, might be illuminating when scrutinized under variable temperatures and conditions in swarming assays.

Though the expression of *papX* did not appear to vary under conditions of Type 1 fimbrial L-ON and L-OFF state, further work should be done to formally characterize the expression levels of *papX* under various environmental conditions. Using a *focX* knockout background of *E. coli* strain CFT073, transcript levels of *papX* should be scrutinized under variable temperatures and conditions known to affect both swimming and swarming motility, as above. It will be important to establish to our satisfaction the

precise mechanism and role for *papX* expression before considering and experimentally investigating alternative mechanisms of *papX* regulation. The work presented here provides strong initial data upon which to base continued studies of PapX mechanism. Many questions of the action of PapX have been resolved, and investigation of the remaining unknowns is likely to bear scientifically valuable dividends.

APPENDIX A

Material and Methods

Materials and methods

Bacterial strains and plasmid

For bacterial strains and plasmids used in this study, refer to **Table 2-2.** *E. coli* CFT073, a fully-sequenced [32] strain of UPEC, was cultured from the blood and urine of a hospitalized patient with acute pyelonephritis and urosepsis [219]. *E. coli* BL21(DE3) was used for over-expression of PapX-His_6. All *E. coli* strains were cultured overnight at 37°C in Luria broth with aeration or on Luria agar plates containing either ampicillin (100 μg/ml), chloramphenicol (20 μg/ml), kanamycin (25 μg/ml or no antibiotic. For western blots, PapX was expressed from the isopropyl β-D-1-thiogalactopyranoside (IPTG) -inducible plasmid pPxWT in *E. coli* K12 strain MG1655 in the presence or absence of 300 μM IPTG. Growth curves were generated in triplicate using a Microbiology Reader Bioscreen C (Oy Growth Curves AB, Ltd.) in 0.2 ml volumes; OD_{600} was recorded every 15 min for 24 h.

Motility assay

Flagellum-mediated swimming motility was measured as described previously [169]. Briefly, soft Luria agar [0.25% (w/v)] plates containing ampicillin (100 μg/ml) were stabbed in the center of the plate using an inoculating needle with either Luria broth cultures normalized to OD_{600}=0.9-1.0 or from colonies struck onto plates the day prior. Care was taken not to touch the bottom of the plate during inoculation to ensure only swimming motility was assessed. Plates were incubated at 30°C for 16 h and swimming diameter was measured. Results were analyzed using a paired *t*-test.

Swarming assays were performed on Eiken agar as described previously [220] with minor modiciation. Briefly, Eiken agar was composed of 3.5 g yeast extract, 7 g tryptone, 3.5 g NaCl, 10.5 g Eiken agar, autoclaved, and 0.5% glucose added per liter. Plates were allowed to cool for 1 day before use. 5 μl of overnight culture of *E. coli* normalized to the culture with the lowest OD_{600} was inoculated onto the center of the plate and allowed to fully absorb into the agar. Plates were then incubated at 23^{o}C, 30^{o}C or 37^{o}C for 48-72 hrs after which images of swarming were recorded.

Construction of promoter-*lacZ* fusions and β-galactosidase assays

Nested deletions of the *flhDC* promoter and the intergenic regions between *papG* and *papX* as well as between *papB* and *papI* were fused to *lacZ* using a promoterless *lacZ*-containing plasmid, pRS551, using previously described methodology [151]. Briefly, primers containing flanking BamHI or EcoRI overhangs were designed to amplify DNA fragments that were -130, -215, -360, -501, -597 and -718 -bp upstream of the translational start site of *flhD* of *E. coli* CFT073, the 262 bases between the end of

papG and the start of *papX*, and the 330 bases preceding *papB* which includes the *pap* operon's promoter (**Table 2-1**). Fragments and the pRS551 vector were digested with EcoRI and BamHI and ligated to form constructs $pRS551_{(a\text{-}f,\ GX\ and\ BA)}$ (**Table 2-2**). Electrocompetent *E. coli* DH5α was electroporated with $pRS551_{(a\text{-}f)}$, incubated for 1 h in 500-800 µl Super Optimal broth with Catabolite repression (SOC), and plated onto Luria agar plates containing kanamycin (25 µg/ml). Plasmids were extracted from Kan^R colonies using a Miniprep kit (Qiagen) and sequenced to confirm the proper fusion constructs. Plasmid constructs were transformed into *E. coli* CFT073 Δ*papX*Δ*lacZ* (Cam^R) or *E. coli* CFT073 Δ*lacZ* (Cam^R). β-galactosidase activities of the constructs were measured using the modified Miller assay as described [195]. Briefly, overnight cultures of each construct were diluted 1:100 in LB broth and incubated at 37°C with aeration. OD_{600} was measured after 2.5 h and samples (5 µl) were added to 95 µl 100 mM Na_2HPO_4, 20 mM KCl, 2mM $MgSO_4$, 0.8 mg/ml CTAB, 0.4mg/ml Na-deoxycholate, 5.4 µl/ml β-mercaptoethanol to lyse bacterial cells. Lysates (100 µl) were mixed with 600 µl 60 mM Na_2HPO_4, 40 mM NaH_2PO_4, 1 mg/ml ONPG, 2.7 µl/ml β-mercaptoethanol reaction buffer and incubated at 30°C. Reactions were stopped after 5-25 min with 700 µl 1 M sodium carbonate. Bacterial cells were pelleted by centrifugation (10,000 *x g*, 10 min, 23°C) and OD_{420} of the supernatant was measured. Miller units were computed using the formula: 1000 X $(Abs_{420})/(Abs_{600}*0.05*$time in min). Results were compared using a paired *t*-test.

Construction of E. coli mutants

Deletion mutants were constructed using the λ-red recombinase system [221]. Briefly, flanking homologous primers were designed (**Table 2-1**) and used to generate amplicons from plasmids pKD3 or pKD4 to generate kanamycin- or chloramphenicol-resistance cassettes. Amplicons were treated with DpnI for 3 h at 37°C to remove parental DNA and transformed into electrocompetent CFT073 bacteria expressing pKD46 induced in 10 mM (final concentration) arabinose. Resistance cassettes were removed by transforming resultant constructs with pCP20 and leaving cultures to incubate at 37°C overnight.

Site-directed mutants were constructed in pPxWT using the QuikChange II kit (Strategene) according to the manufacturer's instructions. Briefly, overlapping primers were designed to contain nucleotide substitutions and used for rolling PCR of pPxWT. PCR products were treated for 3 h with DpnI at 37°C to remove parental DNA. Amplicons were transformed into electrocompetent CFT073, recovered for 1 h in 500-800 μl SOC and then 200 μl plated onto LB agar containing ampicillin (100 μg/ml). Resistant colonies were cultured overnight in LB broth at 37°C and plasmid DNA was extracted and sequenced as described above.

PapX purification

Nucleotides encoding amino acids 13-183 of PapX were cloned into pMCSG7 in front of a His_6-TEV cleavage site by ligation-independent cloning. pMCSG7 was transformed into *E. coli* BL21 (DE3). Bacteria were cultured in 500 ml terrific broth in a 2 L flask at 250 rpm at 37°C to $OD_{600} \cong 1.0$, then cooled to room temperature for 1 h.

Bacteria were induced overnight with 300 μM IPTG. After induction, cells were centrifuged (13,000 x *g*, 20 min, 4°C) and pellets were frozen at -80°C. Frozen pellets were suspended in 20 mM Tris-HCl, pH 8.0, 10% glycerol, 1 mM DTT, 300 mM NaCl and ruptured by sonication using a blunt-tip probe for 7 cycles of 60% maximum energy for 30-sec bursts with 1 min intervals to cool (GENEQ, 500 watt Ultrasonic Processor). The lysate was centrifuged (112,000 x *g*, 1 h, 4°C) and the supernatant was passed through 0.22 μm pore size filter (MillexGP, Millipore) and incubated for 1 h at 4°C with 5-10 ml Ni-NTA resin (Invitrogen). Resin was washed 3 times in 20 mM Tris-HCl, pH 8.0, 10% glycerol, 1 mM DTT, 300 mM NaCl, 20 mM imidazole and protein was eluted with 20 mM Tris-HCl, pH 8.0, 10% glycerol, 1 mM DTT, 300 mM NaCl, 500 mM imidazole. Protein was concentrated to 2 mg/ml and 50 μl aliquots were frozen at -80^{o}C. The presence of PapX was confirmed as the predominant species on a 12% acrylamide gel. Bands were analyzed using tandem mass spectroscopy (LC-MS/MS) by the Michigan State University Proteomics Core Facility.

Co-precipitation

PapX-His_6 was purified as above, with minor modification. Starting culture volume in TB was set to 250 ml, and the culture was not induced. Recombinant PapX-His_6 was bound to Ni-NTA resin and incubated with 1 l wildtype *E. coli* strain CFT073 lysate for 30 minutes at 37^{o}C. Purification proceeded as above from that point onward, and eluate was analyzed using tandem mass spectroscopy (LC-MS/MS) as above.

Electrophoretic motility shift assays

Gel shift assays were performed as described [169] with the following modifications. Recombinant PapX-His_6, expressed from pMCGS7, was purified as described above and used for gel shift assays. The DNA fragment used was the -345 through -501 bp region upstream of the *flhD* translational start site in *E. coli* CFT073. Poly-d(I-C) and poly-L-lysine were substituted with TEN buffer in the binding reaction. Competition was performed using four unlabeled fragments Three fragments, TTACGGTGAGTTATTTTAACTGTGCGCAA, TTACGGTGAGTTATTTTAAC and GTTATTTTAAC, were derived from the predicted PapX binding site from SELEX (see Results). The fourth fragment, ACTCCACTCACGGCCGTTTCGACGGTACCG, was derived from the *gapA* promoter, which has previously been demonstrated not to shift in the presence of PapX [169].

Genomic SELEX

A library of 80-100-bp sequences was generated from CFT073 genomic DNA and paired-end ligated as described [196] at the University of Michigan DNA Sequencing Core. PapX-His_6 was purified as described above; however, before eluting bound protein, 200 µl samples of the resin were incubated with gentle agitation for 30 min at 37°C in the presence of 150 ng of library DNA. The resin was washed with 20 mM Tris-HCl, pH 8.0, 10% glycerol, 1 mM DTT, 300 mM NaCl, 20 mM imidazole to remove unbound library fragments, and bound DNA-PapX complexes were eluted using the same buffer but containing 500 mM imidazole. DNA was isolated from the eluted complexes

by two phenol:chloroform extractions. DNA was subjected to 12-15 rounds of PCR using primers specific to the ligated paired-ends to ensure no contaminating DNA was amplified. Libraries were quantified using quantitative PCR. The resultant library (150 ng) was run again over a fresh 200 μl sample of PapX-His_6 bound to Ni-NTA resin and the process was repeated four times. The unselected library, as well as enriched rounds two, three and four, were sequenced using Illumina high-throughput sequencing. A local BLAST comparison to the CFT073 genome was done for each short read per run and the number of reads at each chromosomal position in CFT073 was recorded. Matlab (Student version 7.12.0) was used to identify peaks that had enriched competitively throughout the cycles of amplification to identify relevant peaks. 60-bp regions, centered on each relevant peak, were analyzed using the MEME suite [192] to generate most likely motifs.

Protein structural predictions

PHYRE (version 0.2) [172] and I-TASSER (version 1.1) [173, 174] were employed to predict the most likely structure of the PapX protein. The model predicted by I-TASSER was rendered using PyMOL Molecular Graphics System (version 1.3, Schrödinger, LLC). Recombinant purified PapX-His_6 was also analyzed by size-exclusion chromatography on a Sephacryl S-200 size-exclusion column (GE Healthcare) and semi-native gel electrophoresis in 12% SDS-polyacrylamide gel.

Antiserum production

Antiserum to PapX was commercially generated in rabbits using a polyclonal antibody fast production protocol (Rockland Immunochemicals Custom Antibody Services) from a single-protein peak eluted from a Sephacryl S-200 size-exclusion column (GE Healthcare) of nickel affinity purified recombinant PapX-His_6.

Transcriptional profiling

Quantitative real-time PCR was performed as described [169]. Briefly, total RNA, extracted from 200 μl of bacterial culture, was reverse-transcribed to generate cDNA, and analyzed using Sybr Green and primers for 100-bp fragments of genes of interest (**Table 2-1**). *gapA* was used as a normalizer for all conditions.

DNase footprinting

Primers, end-labeled with 6-carboxyfluorescein, were used to generate fluorescent probes from the first 130-, 215-, 360-, and 500-bp-fragments of the *flhDC* promoter as well as a 300-bp fragment of the *gapA* promoter. Fragments were incubated with purified recombinant PapX-His_6 and exposed to low-levels of DNase I for 1-10 min. Sheared fragments were processed by electrophoresis (as for DNA sequencing) to produce chromatograms representing all sheared fragments and examined for stretches of troughs between peaks.

Multiplex PCR

Multiplex PCR was performed using methods and a 294-member subset of a 315-member strain collection as described [194, 222]. Primers were designed to unique flanking sequences of *papX*, *focX* and common regions of both (**Table 2-1**).

Semi-native gel electrophoresis

Non-reducing/semi-native gel electrophoresis was performed as described previously [31]. Briefly, protein was isolated as described above. SDS-PAGE loading buffer lacking dithiothreitol was added to samples, which were electrophoresed in the presence of 0.1% SDS on a 12% acrylamide gel (3.75% stacking gel) at 200V for 1 h at room temperature. Gels were stained using Coomassie Brilliant Blue R-250 dissolved in a buffer composed of 10% acetic acid, 40% methanol, and 50% water. Gels were de-stained in a buffer composed of 10% acetic acid, 40% methanol, and 50% water.

APPENDIX B

TFAST Instructions for use

Overview of TFAST

TFAST was originally adapted from a series of algorithms used to analyze SELEX with high-throughput sequencing data. To make optimal use of TFAST, first the initial (unselected) library used for SELEX must be sequenced, as well as selected subsequent cycles. Once sequence data has been procured, they can be processed using TFAST. Alternatively, the user may process files as she or he wishes and enter completed files into the TFAST workflow, so long as files are properly formatted. Included here are instructions for the use of TFAST and the file format used at each stage, so that the user can integrate features as they see fit.

Dependencies: TFAST is written in Java. Java must be installed to execute the .jar files.

Terms

“Probe length” or “fragment length” refer to the average size of fragment submitted for sequencing in each round of SELEX sequenced.

"Upper limit of reads" refers to the number of reads aligned in the cycle of SELEX in which the greatest number of reads was aligned to the target sequence.

"N" and "#" are placeholders for integer values.

"Power" refers to the area under peaks and is a surrogate for representation of a peak in the final cycle of SELEX sequenced

"Weight" is an indicator of how well a given peak enriched. Weights per cycle of SELEX are determined by the user, and should reflect the bulk binding characteristics of the library (ie, should increase with stringency as cycles are iteratively enriched).

Quick use guide

1. Run fileconversion.jar on Illumina _fastq.txt formatted sequence folder

2. Run folder blast.jar on the resulting /fasta directory from step "1"

3. Run freq counter.jar on the resulting /blast directory from step "2"

3. Alternatively, run sam freq counter.jar on .SAM files generated by the user using another aligner.

4. Run peak finder.jar on the resulting freq_n.txt files generated in either step "3"

5. Run evaluation.jar on the resulting peak.txt or spoof.txt files generated in step "4"

Instructions for use

The first step in analyzing aptamer-free SELEX-seq data is to align the reads generated by next-generation sequencing (NGS) to a target sequence (eg, a chromosome). TFAST can accept alignment in the form of BLASTn output or .SAM files. BLASTn tends to generate a very complete alignment of NGS reads to a target sequence, but tends to be computationally heavy and may take some time. To align sequences using BLASTn follow the steps as outlined. TFAST can also accept .SAM alignment files, which can be generated very quickly using programs such as BOWTIE. To use .SAM files, enter the workflow at step "alternate step 3 sam freq counter".

Step 1 fileconversion.jar

BLASTn requires that sequences are in FASTA format, so the first program in the TFAST workflow, fileconversion.jar, converts Illumina high-throughput sequence data (typically in a form similar to "s_#_1_0#_qseq.txt", ie "s_5_1_0001_qseq.txt" or "s_1_1_0120_qseq.txt") into FASTA format. Fileconversion currently only supports Illumina's _qseq.txt format. Fileconversion.jar only coverts uninterrupted reads from _qseq.txt files into FASTA format; partial, degenerate or incomplete reads will not be converted. Fileconversion.jar does not take read quality (PHRED score) into account when converting sequences. For files in fastq format, existing open-source resources exist

for converting sequences to FASTA (such as the Galaxy suite of tools, http://usegalaxy.org)

To use fileconversion.jar, execute the application. This will open the command window and prompt the user to enter the directory that contains _qseq.txt files.

Input: Directory of _qseq.txt files, each line of the format:

HWUSI-EAS1707 1 6 1 1871 1088 0

1AAACAATTGACAATGATTATCATTTGCATTAAAAAGACTACGTAGTA TTTTTATTTCATGAGGAACTATACCCGCCAGCA

fffaWd`ff]ffd^_SaYa^fffdfcdRccacc\^[]`dccW^ccRb^bbccccccccfccacc_cc[WWd bb````BBB

1

Output: /fasta subdirectory of the original directory, with a file corresponding to each _qseq.txt file (naming in line with fasta_ s_#_1_000#_qseq.txt), with lines from the _qseq.txt files formatted as follows (corresponding to the above input line):

>18711088

AAACAATTGACAATGATTATCATTTGCATTAAAAAGACTACGTAGTAT TTTTATTTCATGAGGAACTATACCCGCCAGCA

Notes: If the raw sequence was not determinable at one point (indicated by one or more '.' characters instead of A/C/G/T), the read will be discarded.

Step 2 Folder blast.jar

Requirements: Download and install the appropriate version of BLASTn for your operating system. BLASTn is a utility provided by the NCBI at no cost at (http://www.ncbi.nlm.nih.gov/books/NBK1762).

Folder blast.jar uses FASTA formatted sequences, provided by the user or generated by fileconversion.jar in Step 1. The user will also have to provide a sequence file of the genome of the organism being investigated, in .fna format. .fna files can be found on the NCBI website (http://www.ncbi.nlm.nih.gov/pubmed/).

Folder blast.jar uses a shell script to run BLASTn on the FASTA formatted sequences against the appropriate .fna file of interest. To use folder blast.jar, execute the application. This will open the command window and prompt the user to enter the directory containing the FASTA formatted sequence files as well as the .fna genome sequence file. The result of folder blast.jar are BLAST-formatted .txt files (naming in line with blast_fasta_ s_#_1_000#_qseq.txt). Check the NCBI website for tips on streamlining very large BLAST sets (http://www.ncbi.nlm.nih.gov/books/NBK1762).

Input: Directory of FASTA formatted sequences, as described above. Each .txt file in the directory will be run with BLAST - be sure to include only FASTA sequence files.

Output: /blast subdirectory of the /fasta, with a file corresponding to each FASTA_ s_#_1_000#_qseq.txt file. Output is a series of BLAST files corresponding to each FASTA sequence read. Also requires a plain sequence .fna file.

Step 3 freq counter.jar

Freq counter.jar uses results of local BLAST files to generate a position-frequency table, provided by the user or by running folder blast.jar. Frequency tables can be generated by the user using other utilities or program packages (see description of _freq.txt format, below). Freq counter.jar takes the aligned sequences from folder blast.jar and counts the number of sequence reads that overlap at each nucleotide within the chromosome of interest. This results in a frequency table (displayed freq_#.txt), which is a tab-delimited single column table the length of the target sequence (sequence aligned to, ie chromosome or genome) that indicates frequencies at each position.

To use freq counter.jar, execute the application. This will open the command window and prompt the user to enter the directory containing the files produced in the local BLAST of sequences, the genome's .fna file, the estimated upper bound on the number of aligned reads and the cycle number. Cycle number is "1" for the control cycle (the unselected cycle, i.e. the starting library of sequences), "2" for the second cycle to be analyzed, etc. The estimated upper bound for the number of aligned reads allows freq counter.jar to normalize the number of reads processed. This ensures that freq counter.jar will not bias frequency files due to variation in raw number of reads between or within lanes during sequencing. Blast sample counter.jar can estimate the number of aligned

reads for BLASTn aligned files. To use it, execute the program, select the directory containing the blast_ files and run. The fields will populate automatically.

Input: Directory of BLAST results, as described above. Each .txt file in the directory will be included in the frequency table generation so be sure to include only relevant BLAST results. Also requires a plain sequence .fna file.

Output: /freq subdirectory of the /blast directory, with a freq_n.txt file which is a list of integers, with the nth integer corresponding to the frequency of the nth location in the genome.

Alternate step 3 sam freq counter.jar

A user may choose to use a different method of aligning their NGS data to a target sequence. The easiest way to incorporate alternative (non-BLASTn) alignments into the TFAST workflow is to generate .SAM files (such as by using BOWTIE). There are several good, freely available NGS alignment tools. Many tools of this type require FASTQ formatted sequences (instead of Illumina's _qseq.txt format). An easy way to convert _qseq.txt to FASTQ is to use Bash (the command line in Cygwin and Unix-based systems, *i.e.*, Linux, OSX, etc).

Once the user has generated .SAM files using their alignment tool of choice, execute sam freq counter.jar. This will open the command window and prompt the user to

enter the directory containing .SAM files for a particular cycle, the genome .fna file, the upper bound on number of aligned reads (usually given in aligner reports), the cycle number, the fragment length subjected to sequencing, and the read length (number of bases sequenced/read per fragment). Sam freq counter.jar will then generate freq_n.txt files as described in step 3.

Step 4 – peak finder.jar

Peak finder.jar uses the frequency tables of the control library and the final round of SELEX (that is, the most enriched round that was sequenced). Frequency tables can be generated using TFAST or generated by the user. Peak finder.jar determines local maxima, and rejects peaks that are less than two standard deviations above the mean of an n-bp window (set by the user) around that peak position in the control library. Peak finder.jar also rejects peaks that were two standard deviations above the mean in the control library, to reduce bias from aberrant variation in the control cycle being counted as peaks. Peak finder.jar examines the sequence within a fragment length (probe length) of each peak to determine whether or not a given peak is an aberration or represents a true maxima for that region.

To use peak finder.jar, execute the application. This will open the command window and prompt the user to enter the frequency file of the most enriched SELEX cycle, the frequency file of the control library, the .fna file of the genome of interest, a manually set threshold for peaks in the final cycle (optional), the threshold window (we suggest a value equal to or greater than probe length) and the estimated fragment length

(probe length). Probe length can be determined empirically by running the library on a high-resolution gel, or when the initial library is selected for SELEX analysis.

Peak finder.jar produces two files – Peak.txt and Spoof.txt. Peak.txt is a .txt file containing a single column of chromosomal positions corresponding to predicted peaks. The user may elect to use his/her own peak finding algorithms, in which case, format the output similarly to proceed to Step 5. Spoof.txt is a similarly formatted file that contains chromosomal positions generated randomly. Spoof.txt can be used as a control file to validate results and improve downstream data analysis when using motif finding software and to generate background models of genome sequence behavior and nucleotide representation.

TFAST can also find minima instead of maxima if the user desires. In most SELEX experiments, this will result in a vast number of minima as SELEX tends to produce zero frequency between peaks. This option is useful for knockout experiments, such as transposon mutagenesis studies coupled with high-throughput sequencing. When selecting minima, TFAST subtracts the final cycle from the control cycle and counts regions that are two standard deviations of the control cycle (within the set window) above zero as peaks.

Input: freq_n.txt (where n is the number of cycles), freq_1.txt (the control library), a plain sequence .fna file, a set threshold if necessary, threshold window and the average probe (fragment) length.

Output: /peak subdirectory, with a peak.txt file and a spoof.txt file.

Step 5 – Evaluaton.jar

Evaluaton.jar uses peak positions determined in Step 4 or provided by the user to grade peaks. Before using evaluaton.jar, place all freq_#.txt files of interest (including the control) into a single directory. To use evaluator.jar, execute the application. This will open the command window and prompt the user to enter the directory containing all freq_#.txt files, the .fna file of the genome of interest, the peak.txt file, weights for every round, average probe length (fragment length) and the length of sequence around each peak to be included in the output.

Weights for every round should be delimited with commas. Every cycle of SELEX selects for the most strongly binding species in a pool. Thus, as cycles continue, the requirements for binding increase. Therefore, we advise the user to use a scheme that weighs enrichment in later rounds higher than enrichment in earlier rounds. For example, for a SELEX with three rounds of enrichment (control library, enriched1, enriched2 and enriched3), one scoring schema could be “1, 1.5, 2”. “1,1,1” will result in equal weights in every round, which does not reflect the increasing stringency of competitive binding in the DNA fragment libraries. Ideally, a scoring scheme for weights between rounds will be based on the fraction of input species that are recovered after each round in the SELEX experiment, reflecting the bulk affinity of the library for the protein of interest. These considerations are less of a concern if the user performs “fixed stringency SELEX”.

Once completed, evaluaton.jar will produce a tab-delimited table called "output.txt", including data on each peak.

-The first (left-most) column describes the peak number, which for most users can be disregarded (useful for testing the program internally).

-The second column describes the location of the peak within the chromosome, based on the .fna file provided.

-The third through nth columns describe the frequency at that chromosomal position for each SELEX cycle amplified, where "n" is the cycle number (1 = the control library, 2 = first enrichment entered, etc).

-The (n+1)th column describes the weight of a peak. Weight is determined by adding the "weight" values when the frequency at a position increased between cycles, and by adding 0 when the frequency at a position remained unchanged or declined. High weights correspond to strongly enriching sequences.

-The (n+2)th column describes the power of a peak. Power represents the area under a peak as compared to the total area under peaks for the final cycle sequence. Power is another way to grade how strongly a fragment has enriched relative to every other fragment.

-the (n+3)th column describes an estimated binding site length. This is determined by finding the bounds under a peak that includes 99% of the peak area, and subtracting that distance from twice the probe length. Because probe length will vary slightly for technical reasons, it is best to use an average of several peaks to assess the theoretical

binding length for a site. This can be useful for determining cases of multiple adjacent binding sites as well as for discarding large repeated regions.

-The (n+4)th column describes The chromosomal sequence that falls under a given peak, with the amount of sequence included determined by the user. The length included should conform to the standards of any downstream analysis planned.

Evaluation.jar can also analyze the quality of minima (valleys) instead of peaks. This utility is useful for knockout experiments that incorporate NGS, but is unlikely to yield results in SELEX data.

Input: A folder containing only the freq_1.txt through freq_n.txt files, with every file from 1 to n included. Also, a plain sequence .fna file, average probe length, length of flanking sequence to be printed, and a list of weights.

Output: /output subdirectory, with a output.txt file as referenced above.

A dataset of SELEX data using high-throughput sequencing is provided in supplemental data. Feel free to make use of them as a practice set. The files are from an actual experiment that used the algorithms found in TFAST to determine and

experimentally validate a novel, unique binding site for a transcription factor (PapX) in E. coli.

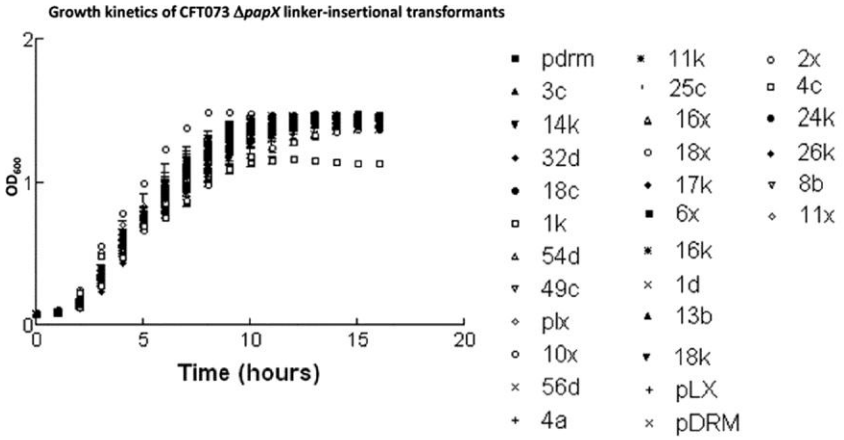
APPENDIX C

TFAST availability

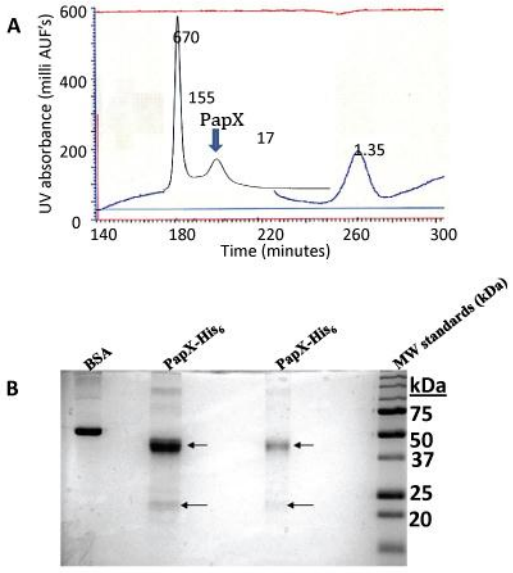
TFAST is implemented in Java and supported on MS Windows and Mac OSX. TFAST was designed in compliance with the GNU GPL. TFAST, documentation, source code, instructions for use, sample data and example output are freely available for download at http://www-personal.umich.edu/~hmobley/ or http://sourceforge.net/projects/tfast/files/.

APPENDIX D

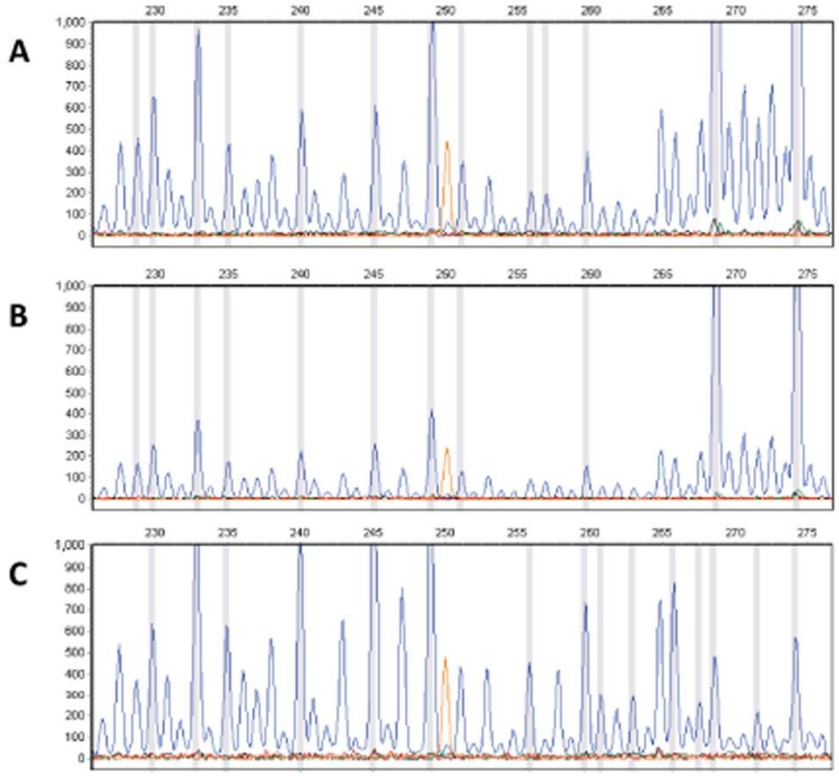
Supplemental data



Supplementary Figure 2-1. *papX* linker-insertional mutants do not have impaired growth rate. Thirty 15-bp insertional mutants of *papX* in pDRM001, as well as WT *papX* in pDRM001, were electroporated into strain CFT073 Δ*papX*. Strains were diluted 1:100 in LB broth containing 100 μg/ml ampicillin and not IPTG and OD_{600} was automatically recorded every 15 min by a bioanalyzer. No significant difference in growth rate was detected among the constructs. Name of transforming plasmid is shown on right.



Supplementary Figure 2-2. PapX is a dimer. (A) Recombinant purified PapX-His_6 and molecular-weight standards were run through a Sephacryl S-200 size-exclusion column at 1 ml/min and OD_{280} was measured in mAU over a 300 ml volume. Traces of each were superimposed by aligning the void-volumes. PapX-His_6 peak appears at approximately twice the predicted molecular weight of the monomer. **(B)** SDS-PAGE was performed on non-denatured recombinant PapX-His6. A major band appears at twice the molecular weight of the monomer (the minor band).



Supplementary Figure 2-3. Non-radiolabeled fluorescent DNase protect assay **(A)** 0.5 μg of recombinant purified PapX-His_6 was incubated with a 6-carboxyfluorescein end-labeled fragment derived from the first 360 bases of the *flhD* promoter and exposed to low-levels of DNase I for 1-20 minutes. Sheared fragments were processed by electrophoresis as for DNA sequencing to produce chromatograms representing all sheared fragments. Fluorescence of the 6-carbosyfluorescein is shown on the y-axis. Position of the fragment is shown on the x-asis. **(A)** Representative chromatogram of a fragment exposed to DNase I for 1 minute in the absence of PapX protein **(B)** Representative chromatogram of a fragment exposed to DNase I for 1 minute in the presence of 0.5 μg of PapX. **(C)** Representative chromatogram of a fragment exposed to DNase I for 20 minutes in the presence of 0.5 μg of PapX. Ratios of peak heights remained the same between conditions.

REFERENCES

1. Moore, K.L., A.F. Dalley, and A.M.R. Agur, *Clinically oriented anatomy*. 6th ed2010, Philadelphia: Wolters Kluwer/Lippincott Williams & Wilkins. xxix, 1134 p.
2. Czaja, C.A., et al., *Prospective cohort study of microbial and inflammatory events immediately preceding Escherichia coli recurrent urinary tract infection in women.* J Infect Dis, 2009. **200**(4): p. 528-36.
3. Schilling, J.D., R.G. Lorenz, and S.J. Hultgren, *Effect of trimethoprim-sulfamethoxazole on recurrent bacteriuria and bacterial persistence in mice infected with uropathogenic Escherichia coli.* Infection and Immunity, 2002. **70**(12): p. 7042-9.
4. Hooton, T.M., *Recurrent urinary tract infection in women.* Int J Antimicrob Agents, 2001. **17**(4): p. 259-68.
5. Bacheller, C.D. and J.M. Bernstein, *Urinary tract infections.* Med Clin North Am, 1997. **81**(3): p. 719-30.
6. Gupta, K., et al., *International clinical practice guidelines for the treatment of acute uncomplicated cystitis and pyelonephritis in women: A 2010 update by the Infectious Diseases Society of America and the European Society for Microbiology and Infectious Diseases.* Clin Infect Dis, 2011. **52**(5): p. e103-20.
7. Faro, S. and D.E. Fenner, *Urinary tract infections.* Clin Obstet Gynecol, 1998. **41**(3): p. 744-54.
8. Ikaheimo, R., et al., *Community-acquired pyelonephritis in adults: characteristics of E. coli isolates in bacteremic and non-bacteremic patients.* Scand J Infect Dis, 1994. **26**(3): p. 289-96.
9. Litwin, M., C.S. Saigal, and E.M. Beerbohm, *The burden of urologic diseases in America.* Journal of Urology, 2005. **173**(4): p. 1065-1066.
10. DeFrances, C.J., et al., *2006 National Hospital Discharge Survey.* Natl Health Stat Report, 2008(5): p. 1-20.
11. Bent, S., et al., *Does this woman have an acute uncomplicated urinary tract infection?* JAMA, 2002. **287**(20): p. 2701-10.
12. Warren, J.W., et al., *Guidelines for antimicrobial treatment of uncomplicated acute bacterial cystitis and acute pyelonephritis in women. Infectious Diseases Society of America (IDSA).* Clin Infect Dis, 1999. **29**(4): p. 745-58.
13. Fairley, K.F., et al., *Site of infection in acute urinary-tract infection in general practice.* Lancet, 1971. **2**(7725): p. 615-8.
14. Stamm, W.E., *Measurement of pyuria and its relation to bacteriuria.* Am J Med, 1983. **75**(1B): p. 53-8.
15. Kunin, C.M., *Urinary tract infections : detection, prevention, and management*. 5th ed1997, Baltimore: Williams & Wilkins. ix, 419 p.
16. Heptinstall, R.H., *Pathology of the kidney*. 3rd ed1983, Boston: Little, Brown. 3 v. (xi, 1695, 42 p.).

17. Svanborg, C. and G. Godaly, *Bacterial virulence in urinary tract infection.* Infectious Disease Clinics of North America, 1997. **11**(3): p. 513-&.
18. Raz, R., R. Colodner, and C.M. Kunin, *Who are you--Staphylococcus saprophyticus?* Clin Infect Dis, 2005. **40**(6): p. 896-8.
19. Wallmark, G., I. Arremark, and B. Telander, *Staphylococcus saprophyticus: a frequent cause of acute urinary tract infection among female outpatients.* J Infect Dis, 1978. **138**(6): p. 791-7.
20. Gupta, K., et al., *The prevalence of antimicrobial resistance among uropathogens causing acute uncomplicated cystitis in young women.* Int J Antimicrob Agents, 1999. **11**(3-4): p. 305-8.
21. Gillespie, W.A., et al., *Urinary tract infection in young women, with special reference to Staphylococcus saprophyticus.* J Clin Pathol, 1978. **31**(4): p. 348-50.
22. Jordan, P.A., et al., *Urinary tract infection caused by Staphylococcus saprophyticus.* J Infect Dis, 1980. **142**(4): p. 510-5.
23. Czaja, C.A., et al., *Population-based epidemiologic analysis of acute pyelonephritis.* Clin Infect Dis, 2007. **45**(3): p. 273-80.
24. Echols, R.M., et al., *Demographic, clinical, and treatment parameters influencing the outcome of acute cystitis.* Clin Infect Dis, 1999. **29**(1): p. 113-9.
25. Podschun, R. and U. Ullmann, *Klebsiella spp. as nosocomial pathogens: epidemiology, taxonomy, typing methods, and pathogenicity factors.* Clin Microbiol Rev, 1998. **11**(4): p. 589-603.
26. Lin, W.H., et al., *Clinical and microbiological characteristics of Klebsiella pneumoniae isolates causing community-acquired urinary tract infections.* Infection, 2010. **38**(6): p. 459-64.
27. Ronald, A., *The etiology of urinary tract infection: traditional and emerging pathogens.* Dis Mon, 2003. **49**(2): p. 71-82.
28. Johnson, D.E. and R.G. Russell, *Animal Models of urinary tract infection*, in *Urinary Tract Infections: Molecular Pathogenesis and Clinical Management*, H.L. Mobley and J.W. Warren, Editors. 1996, ASM Press: Washington, DC. p. 377-403.
29. Dlugaszek, M., M. Kaszczuk, and M. Mularczyk-Oliwa, *Magnesium, calcium, and trace elements excretion in 24-h urine.* Biol Trace Elem Res, 2011. **142**(1): p. 1-10.
30. Wandersman, C. and P. Delepelaire, *Bacterial iron sources: from siderophores to hemophores.* Annu Rev Microbiol, 2004. **58**: p. 611-47.
31. Hagan, E.C. and H.L. Mobley, *Haem acquisition is facilitated by a novel receptor Hma and required by uropathogenic Escherichia coli for kidney infection.* Molecular Microbiology, 2009. **71**(1): p. 79-91.
32. Welch, R.A., et al., *Extensive mosaic structure revealed by the complete genome sequence of uropathogenic Escherichia coli.* Proc Natl Acad Sci U S A, 2002. **99**(26): p. 17020-4.
33. Horwitz, M.A. and S.C. Silverstein, *Influence of the Escherichia coli capsule on complement fixation and on phagocytosis and killing by human phagocytes.* J Clin Invest, 1980. **65**(1): p. 82-94.

34. Burns, S.M. and S.I. Hull, *Loss of resistance to ingestion and phagocytic killing by O(-) and K(-) mutants of a uropathogenic Escherichia coli O75:K5 strain.* Infection and Immunity, 1999. **67**(8): p. 3757-62.
35. Russo, T., et al., *The O4 specific antigen moiety of lipopolysaccharide but not the K54 group 2 capsule is important for urovirulence of an extraintestinal isolate of Escherichia coli.* Infection and Immunity, 1996. **64**(6): p. 2343-8.
36. Russo, T.A., et al., *The construction and characterization of colanic acid deficient mutants in an extraintestinal isolate of Escherichia coli (O4/K54/H5).* Microb Pathog, 1995. **18**(4): p. 269-78.
37. Smith, S.N., et al., *Dissemination and systemic colonization of uropathogenic Escherichia coli in a murine model of bacteremia.* MBio, 2010. **1**(5).
38. Snyder, J.A., et al., *Transcriptome of uropathogenic Escherichia coli during urinary tract infection.* Infection and Immunity, 2004. **72**(11): p. 6373-81.
39. Welch, R.A., et al., *Haemolysin contributes to virulence of extra-intestinal E. coli infections.* Nature, 1981. **294**(5842): p. 665-7.
40. Welch, R.A., R. Hull, and S. Falkow, *Molecular cloning and physical characterization of a chromosomal hemolysin from Escherichia coli.* Infect Immun, 1983. **42**(1): p. 178-86.
41. Smith, Y.C., et al., *Hemolysin of uropathogenic Escherichia coli evokes extensive shedding of the uroepithelium and hemorrhage in bladder tissue within the first 24 hours after intraurethral inoculation of mice.* Infect Immun, 2008. **76**(7): p. 2978-90.
42. Vigil, P.D., C.J. Alteri, and H.L. Mobley, *Identification of in vivo-induced antigens including an RTX family exoprotein required for uropathogenic Escherichia coli virulence.* Infect Immun, 2011. **79**(6): p. 2335-44.
43. Vigil, P.D., et al., *Presence of putative repeat-in-toxin gene tosA in Escherichia coli predicts successful colonization of the urinary tract.* MBio, 2011. **2**(3): p. e00066-11.
44. Hofman, P., et al., *Escherichia coli cytotoxic necrotizing factor-1 (CNF-1) increases the adherence to epithelia and the oxidative burst of human polymorphonuclear leukocytes but decreases bacteria phagocytosis.* J Leukoc Biol, 2000. **68**(4): p. 522-8.
45. Davis, J.M., et al., *Cytotoxic necrotizing factor type 1 delivered by outer membrane vesicles of uropathogenic Escherichia coli attenuates polymorphonuclear leukocyte antimicrobial activity and chemotaxis.* Infect Immun, 2006. **74**(8): p. 4401-8.
46. Blum, G., et al., *Gene clusters encoding the cytotoxic necrotizing factor type 1, Prs-fimbriae and alpha-hemolysin form the pathogenicity island II of the uropathogenic Escherichia coli strain J96.* Fems Microbiology Letters, 1995. **126**(2): p. 189-95.
47. Lloyd, A.L., D.A. Rasko, and H.L. Mobley, *Defining genomic islands and uropathogen-specific genes in uropathogenic Escherichia coli.* Journal of Bacteriology, 2007. **189**(9): p. 3532-46.
48. Smith, T.G. and T.R. Hoover, *Deciphering bacterial flagellar gene regulatory networks in the genomic era.* Adv Appl Microbiol, 2009. **67**: p. 257-95.

49. Lane, M.C., et al., *Role of motility in the colonization of uropathogenic Escherichia coli in the urinary tract.* Infection and Immunity, 2005. **73**(11): p. 7644-56.
50. Wright, K.J., P.C. Seed, and S.J. Hultgren, *Uropathogenic Escherichia coli flagella aid in efficient urinary tract colonization.* Infection and Immunity, 2005. **73**(11): p. 7657-68.
51. Bahrani-Mougeot, F.K., et al., *Type 1 fimbriae and extracellular polysaccharides are preeminent uropathogenic Escherichia coli virulence determinants in the murine urinary tract.* Molecular Microbiology, 2002. **45**(4): p. 1079-93.
52. Connell, I., et al., *Type 1 fimbrial expression enhances Escherichia coli virulence for the urinary tract.* Proc Natl Acad Sci U S A, 1996. **93**(18): p. 9827-32.
53. Gunther, N.W.t., et al., *Assessment of virulence of uropathogenic Escherichia coli type 1 fimbrial mutants in which the invertible element is phase-locked on or off.* Infection and Immunity, 2002. **70**(7): p. 3344-54.
54. Snyder, J.A., et al., *Role of phase variation of type 1 fimbriae in a uropathogenic Escherichia coli cystitis isolate during urinary tract infection.* Infection and Immunity, 2006. **74**(2): p. 1387-93.
55. Eden, C.S. and H.A. Hansson, *Escherichia coli pili as possible mediators of attachment to human urinary tract epithelial cells.* Infection and Immunity, 1978. **21**(1): p. 229-37.
56. Eden, C.S., et al., *Variable adherence to normal human urinary-tract epithelial cells of Escherichia coli strains associated with various forms of urinary-tract infection.* Lancet, 1976. **1**(7984): p. 490-2.
57. Kallenius, G., et al., *Identification of a carbohydrate receptor recognized by uropathogenic Escherichia coli.* Infection, 1980. **8 Suppl 3**: p. 288-93.
58. Kallenius, G., et al., *Structure of carbohydrate part of receptor on human uroepithelial cells for pyelonephritogenic Escherichia coli.* Lancet, 1981. **2**(8247): p. 604-6.
59. Lane, M.C., et al., *Expression of flagella is coincident with uropathogenic Escherichia coli ascension to the upper urinary tract.* Proc Natl Acad Sci U S A, 2007. **104**(42): p. 16669-74.
60. Simms, A.N. and H.L. Mobley, *Multiple genes repress motility in uropathogenic Escherichia coli constitutively expressing type 1 fimbriae.* Journal of Bacteriology, 2008. **190**(10): p. 3747-56.
61. Spurbeck, R.R., et al., *Fimbrial profiles predict virulence of uropathogenic Escherichia coli strains: contribution of Ygi and Yad Fimbriae.* Infection and Immunity, 2011. **79**(12): p. 4753-63.
62. Langermann, S., et al., *Prevention of mucosal Escherichia coli infection by FimH-adhesin-based systemic vaccination.* Science, 1997. **276**(5312): p. 607-11.
63. Langermann, S., et al., *Vaccination with FimH adhesin protects cynomolgus monkeys from colonization and infection by uropathogenic Escherichia coli.* J Infect Dis, 2000. **181**(2): p. 774-8.
64. Brinton, C.C., Jr., *The structure, function, synthesis and genetic control of bacterial pili and a molecular model for DNA and RNA transport in gram negative bacteria.* Trans N Y Acad Sci, 1965. **27**(8): p. 1003-54.

65. Brinton, C.C., Jr., A. Buzzell, and M.A. Lauffer, *Electrophoresis and phage susceptibility studies on a filament-producing variant of the E. coli B bacterium.* Biochimica et biophysica acta, 1954. **15**(4): p. 533-42.
66. Norgren, M., et al., *Mutations in E coli cistrons affecting adhesion to human cells do not abolish Pap pili fiber formation.* EMBO J, 1984. **3**(5): p. 1159-65.
67. O'Hanley, P., et al., *Gal-Gal binding and hemolysin phenotypes and genotypes associated with uropathogenic Escherichia coli.* N Engl J Med, 1985. **313**(7): p. 414-20.
68. Hultgren, S.J., et al., *Regulation of production of type 1 pili among urinary tract isolates of Escherichia coli.* Infection and Immunity, 1986. **54**(3): p. 613-20.
69. Lawn, A.M., *Morphological features of the pili associated with Escherichia coli K 12 carrying R factors or the F factor.* Journal of general microbiology, 1966. **45**(2): p. 377-83.
70. Brinton, C.C., Jr., *Non-flagellar appendages of bacteria.* Nature, 1959. **183**(4664): p. 782-6.
71. Korhonen, T.K., R. Virkola, and H. Holthofer, *Localization of binding sites for purified Escherichia coli P fimbriae in the human kidney.* Infection and Immunity, 1986. **54**(2): p. 328-32.
72. Vaisanen, V., et al., *Mannose-resistant haemagglutination and P antigen recognition are characteristic of Escherichia coli causing primary pyelonephritis.* Lancet, 1981. **2**(8260-61): p. 1366-9.
73. Korhonen, T.K., et al., *P-antigen-recognizing fimbriae from human uropathogenic Escherichia coli strains.* Infection and Immunity, 1982. **37**(1): p. 286-91.
74. Uhlin, B.E., et al., *Adhesion to human cells by Escherichia coli lacking the major subunit of a digalactoside-specific pilus-adhesin.* Proc Natl Acad Sci U S A, 1985. **82**(6): p. 1800-4.
75. Mulholland, S.G., M. Mooreville, and C.L. Parsons, *Urinary tract infections and P blood group antigens.* Urology, 1984. **24**(3): p. 232-5.
76. Lomberg, H., et al., *Correlation of P blood group, vesicoureteral reflux, and bacterial attachment in patients with recurrent pyelonephritis.* The New England journal of medicine, 1983. **308**(20): p. 1189-92.
77. Vaisanen-Rhen, V., et al., *P-fimbriated clones among uropathogenic Escherichia coli strains.* Infection and Immunity, 1984. **43**(1): p. 149-55.
78. Dowling, K.J., J.A. Roberts, and M.B. Kaack, *P-fimbriated Escherichia coli urinary tract infection: a clinical correlation.* South Med J, 1987. **80**(12): p. 1533-6.
79. Hagberg, L., et al., *Adhesion, hemagglutination, and virulence of Escherichia coli causing urinary tract infections.* Infection and Immunity, 1981. **31**(2): p. 564-70.
80. Jacobsen, S.H., et al., *P fimbriated Escherichia coli in adults with acute pyelonephritis.* J Infect Dis, 1985. **152**(2): p. 426-7.
81. Johnson, J.R., et al., *Association of carboxylesterase B electrophoretic pattern with presence and expression of urovirulence factor determinants and antimicrobial resistance among strains of Escherichia coli that cause urosepsis.* Infection and Immunity, 1991. **59**(7): p. 2311-5.

82. Johnson, J.R., P.L. Roberts, and W.E. Stamm, *P fimbriae and other virulence factors in Escherichia coli urosepsis: association with patients' characteristics.* J Infect Dis, 1987. **156**(1): p. 225-9.
83. Kallenius, G., et al., *Occurrence of P-fimbriated Escherichia coli in urinary tract infections.* Lancet, 1981. **2**(8260-61): p. 1369-72.
84. Latham, R.H. and W.E. Stamm, *Role of fimbriated Escherichia coli in urinary tract infections in adult women: correlation with localization studies.* J Infect Dis, 1984. **149**(6): p. 835-40.
85. Sandberg, T., et al., *Virulence of Escherichia coli in relation to host factors in women with symptomatic urinary tract infection.* J Clin Microbiol, 1988. **26**(8): p. 1471-6.
86. Ulleryd, P., et al., *Virulence characteristics of Escherichia coli in relation to host response in men with symptomatic urinary tract infection.* Clin Infect Dis, 1994. **18**(4): p. 579-84.
87. Kisielius, P.V., et al., *In vivo expression and variation of Escherichia coli type 1 and P pili in the urine of adults with acute urinary tract infections.* Infection and Immunity, 1989. **57**(6): p. 1656-62.
88. de Ree, J.M. and J.F. van den Bosch, *Serological response to the P fimbriae of uropathogenic Escherichia coli in pyelonephritis.* Infection and Immunity, 1987. **55**(9): p. 2204-7.
89. Hull, R.A., et al., *Construction and expression of recombinant plasmids encoding type 1 or D-mannose-resistant pili from a urinary tract infection Escherichia coli isolate.* Infection and Immunity, 1981. **33**(3): p. 933-8.
90. Dodson, K.W., et al., *Outer-membrane PapC molecular usher discriminately recognizes periplasmic chaperone-pilus subunit complexes.* Proc Natl Acad Sci U S A, 1993. **90**(8): p. 3670-4.
91. Hultgren, S.J., et al., *Pilus and nonpilus bacterial adhesins: assembly and function in cell recognition.* Cell, 1993. **73**(5): p. 887-901.
92. Kuehn, M.J., et al., *P pili in uropathogenic E. coli are composite fibres with distinct fibrillar adhesive tips.* Nature, 1992. **356**(6366): p. 252-5.
93. Kuehn, M.J., S. Normark, and S.J. Hultgren, *Immunoglobulin-like PapD chaperone caps and uncaps interactive surfaces of nascently translocated pilus subunits.* Proc Natl Acad Sci U S A, 1991. **88**(23): p. 10586-90.
94. Lund, B., et al., *The PapG protein is the alpha-D-galactopyranosyl-(1----4)-beta-D-galactopyranose-binding adhesin of uropathogenic Escherichia coli.* Proc Natl Acad Sci U S A, 1987. **84**(16): p. 5898-902.
95. Normark, S., et al., *Genetics of digalactoside-binding adhesin from a uropathogenic Escherichia coli strain.* Infection and Immunity, 1983. **41**(3): p. 942-9.
96. Remaut, H., et al., *Fiber formation across the bacterial outer membrane by the chaperone/usher pathway.* Cell, 2008. **133**(4): p. 640-52.
97. Kline, K.A., et al., *A tale of two pili: assembly and function of pili in bacteria.* Trends in microbiology, 2010. **18**(5): p. 224-32.
98. Larsson, A., et al., *Quantitative studies of the binding of the class II PapG adhesin from uropathogenic Escherichia coli to oligosaccharides.* Bioorg Med Chem, 2003. **11**(10): p. 2255-61.

99. Ohlsson, J., et al., *Discovery of potent inhibitors of PapG adhesins from uropathogenic Escherichia coli through synthesis and evaluation of galabiose derivatives.* Chembiochem, 2002. **3**(8): p. 772-9.
100. Dodson, K.W., et al., *Structural basis of the interaction of the pyelonephritic E. coli adhesin to its human kidney receptor.* Cell, 2001. **105**(6): p. 733-43.
101. Lund, B., et al., *Uropathogenic Escherichia coli can express serologically identical pili of different receptor binding specificities.* Molecular Microbiology, 1988. **2**(2): p. 255-63.
102. Baga, M., et al., *Transcriptional activation of a pap pilus virulence operon from uropathogenic Escherichia coli.* EMBO J, 1985. **4**(13B): p. 3887-93.
103. Braaten, B.A., et al., *Methylation patterns in pap regulatory DNA control pyelonephritis-associated pili phase variation in E. coli.* Cell, 1994. **76**(3): p. 577-88.
104. Peterson, S.N. and N.O. Reich, *Competitive Lrp and Dam assembly at the pap regulatory region: implications for mechanisms of epigenetic regulation.* Journal of molecular biology, 2008. **383**(1): p. 92-105.
105. Hintsche-Kilger, B.B., et al., *[Benign liver tumors and oral contraceptives--difficulties in diagnosis]*. Zeitschrift fur arztliche Fortbildung, 1991. **85**(11): p. 513-5.
106. Hagan, E.C., et al., *Escherichia coli global gene expression in urine from women with urinary tract infection.* PLoS Pathog, 2010. **6**(11): p. e1001187.
107. Holden, N., et al., *Regulation of P-fimbrial phase variation frequencies in Escherichia coli CFT073.* Infection and Immunity, 2007. **75**(7): p. 3325-34.
108. Mobley, H.L., et al., *Isogenic P-fimbrial deletion mutants of pyelonephritogenic Escherichia coli: the role of alpha Gal(1-4) beta Gal binding in virulence of a wild-type strain.* Molecular Microbiology, 1993. **10**(1): p. 143-55.
109. Roberts, J.A., et al., *The Gal(alpha 1-4)Gal-specific tip adhesin of Escherichia coli P-fimbriae is needed for pyelonephritis to occur in the normal urinary tract.* Proc Natl Acad Sci U S A, 1994. **91**(25): p. 11889-93.
110. Roberts, J.A., et al., *Receptors for pyelonephritogenic Escherichia coli in primates.* J Urol, 1984. **131**(1): p. 163-8.
111. Melican, K., et al., *Uropathogenic Escherichia coli P and Type 1 fimbriae act in synergy in a living host to facilitate renal colonization leading to nephron obstruction.* PLoS pathogens, 2011. **7**(2): p. e1001298.
112. Melican, K., et al., *Uropathogenic Escherichia coli P and Type 1 fimbriae act in synergy in a living host to facilitate renal colonization leading to nephron obstruction.* PLoS Pathog, 2011. **7**(2): p. e1001298.
113. Snyder, J.A., et al., *Coordinate expression of fimbriae in uropathogenic Escherichia coli.* Infection and Immunity, 2005. **73**(11): p. 7588-96.
114. Duguid, J.P., S. Clegg, and M.I. Wilson, *The fimbrial and non-fimbrial haemagglutinins of Escherichia coli.* J Med Microbiol, 1979. **12**(2): p. 213-27.
115. Orndorff, P.E. and C.A. Bloch, *The role of type 1 pili in the pathogenesis of Escherichia coli infections: a short review and some new ideas.* Microb Pathog, 1990. **9**(2): p. 75-9.

116. Hultgren, S.J., S. Normark, and S.N. Abraham, *Chaperone-assisted assembly and molecular architecture of adhesive pili.* Annu Rev Microbiol, 1991. **45**: p. 383-415.
117. Schilling, J.D., M.A. Mulvey, and S.J. Hultgren, *Structure and function of Escherichia coli type 1 pili: new insight into the pathogenesis of urinary tract infections.* J Infect Dis, 2001. **183 Suppl 1**: p. S36-40.
118. Johnson, J.R., *Virulence factors in Escherichia coli urinary tract infection.* Clin Microbiol Rev, 1991. **4**(1): p. 80-128.
119. Jones, C.H., et al., *FimH adhesin of type 1 pili is assembled into a fibrillar tip structure in the Enterobacteriaceae.* Proc Natl Acad Sci U S A, 1995. **92**(6): p. 2081-5.
120. Abraham, S.N. and E.H. Beachey, *Assembly of a chemically synthesized peptide of Escherichia coli type 1 fimbriae into fimbria-like antigenic structures.* Journal of Bacteriology, 1987. **169**(6): p. 2460-5.
121. Abraham, S.N., J.D. Goguen, and E.H. Beachey, *Hyperadhesive mutant of type 1-fimbriated Escherichia coli associated with formation of FimH organelles (fimbriosomes).* Infection and Immunity, 1988. **56**(5): p. 1023-9.
122. Hacker, J., et al., *Deletions of chromosomal regions coding for fimbriae and hemolysins occur in vitro and in vivo in various extraintestinal Escherichia coli isolates.* Microb Pathog, 1990. **8**(3): p. 213-25.
123. Hung, C.S., et al., *Structural basis of tropism of Escherichia coli to the bladder during urinary tract infection.* Molecular Microbiology, 2002. **44**(4): p. 903-15.
124. Krogfelt, K.A., H. Bergmans, and P. Klemm, *Direct evidence that the FimH protein is the mannose-specific adhesin of Escherichia coli type 1 fimbriae.* Infection and Immunity, 1990. **58**(6): p. 1995-8.
125. Wu, X.R., T.T. Sun, and J.J. Medina, *In vitro binding of type 1-fimbriated Escherichia coli to uroplakins Ia and Ib: relation to urinary tract infections.* Proc Natl Acad Sci U S A, 1996. **93**(18): p. 9630-5.
126. Min, G., et al., *Localization of uroplakin Ia, the urothelial receptor for bacterial adhesin FimH, on the six inner domains of the 16 nm urothelial plaque particle.* J Mol Biol, 2002. **317**(5): p. 697-706.
127. Min, G., et al., *Structural basis of urothelial permeability barrier function as revealed by Cryo-EM studies of the 16 nm uroplakin particle.* J Cell Sci, 2003. **116**(Pt 20): p. 4087-94.
128. Zhou, G., et al., *Uroplakin Ia is the urothelial receptor for uropathogenic Escherichia coli: evidence from in vitro FimH binding.* J Cell Sci, 2001. **114**(Pt 22): p. 4095-103.
129. Sokurenko, E.V., et al., *Pathogenic adaptation of Escherichia coli by natural variation of the FimH adhesin.* Proc Natl Acad Sci U S A, 1998. **95**(15): p. 8922-6.
130. Donnenberg, M.S. and R.A. Welch, *Virulence determinants of uropathogenic Escherichia coli.*, in *Urinary Tract Infections: Molecular Pathogenesis and Clinical Management*, H.L. Mobley and J.W. Warren, Editors. 1996, ASM Press: Washington DC.

131. Pere, A., et al., *Occurrence of type-1C fimbriae on Escherichia coli strains isolated from human extraintestinal infections.* J Gen Microbiol, 1985. **131**(7): p. 1705-11.
132. Zingler, G., et al., *Clonal differentiation of uropathogenic Escherichia coli isolates of serotype O6:K5 by fimbrial antigen typing and DNA long-range mapping techniques.* Med Microbiol Immunol, 1993. **182**(1): p. 13-24.
133. Zingler, G., et al., *Clonal analysis of Escherichia coli serotype O6 strains from urinary tract infections.* Microb Pathog, 1992. **12**(4): p. 299-310.
134. Lindberg, S., et al., *Regulatory Interactions among adhesin gene systems of uropathogenic Escherichia coli.* Infection and Immunity, 2008. **76**(2): p. 771-80.
135. Korhonen, T.K., et al., *Tissue tropism of Escherichia coli adhesins in human extraintestinal infections.* Curr Top Microbiol Immunol, 1990. **151**: p. 115-27.
136. Pere, A., et al., *Expression of P, type-1, and type-1C fimbriae of Escherichia coli in the urine of patients with acute urinary tract infection.* J Infect Dis, 1987. **156**(4): p. 567-74.
137. Castelain, M., et al., *Fast uncoiling kinetics of F1C pili expressed by uropathogenic Escherichia coli are revealed on a single pilus level using force-measuring optical tweezers.* Eur Biophys J, 2011. **40**(3): p. 305-16.
138. Backhed, F., et al., *Identification of target tissue glycosphingolipid receptors for uropathogenic, F1C-fimbriated Escherichia coli and its role in mucosal inflammation.* J Biol Chem, 2002. **277**(20): p. 18198-205.
139. Autar, R., et al., *Adhesion inhibition of F1C-fimbriated Escherichia coli and Pseudomonas aeruginosa PAK and PAO by multivalent carbohydrate ligands.* Chembiochem, 2003. **4**(12): p. 1317-25.
140. Harshey, R.M. and A. Toguchi, *Spinning tails: homologies among bacterial flagellar systems.* Trends Microbiol, 1996. **4**(6): p. 226-31.
141. Chilcott, G.S. and K.T. Hughes, *Coupling of flagellar gene expression to flagellar assembly in Salmonella enterica serovar typhimurium and Escherichia coli.* Microbiol Mol Biol Rev, 2000. **64**(4): p. 694-708.
142. Kalir, S., et al., *Ordering genes in a flagella pathway by analysis of expression kinetics from living bacteria.* Science, 2001. **292**(5524): p. 2080-3.
143. Komeda, Y., *Fusions of flagellar operons to lactose genes on a mu lac bacteriophage.* Journal of Bacteriology, 1982. **150**(1): p. 16-26.
144. Komeda, Y., *Transcriptional control of flagellar genes in Escherichia coli K-12.* Journal of Bacteriology, 1986. **168**(3): p. 1315-8.
145. Kutsukake, K., Y. Ohya, and T. Iino, *Transcriptional analysis of the flagellar regulon of Salmonella typhimurium.* Journal of Bacteriology, 1990. **172**(2): p. 741-7.
146. Macnab, R.M., *Flagella and motility*, in *Escherichia coli and Salmonella typhimurium: Cellular and Molecular Biology*, F.C. Neidhardt, et al., Editors. 1996, ASM Press: Washington, DC. p. 123-145.
147. Soutourina, O.A. and P.N. Bertin, *Regulation cascade of flagellar expression in Gram-negative bacteria.* FEMS Microbiol Rev, 2003. **27**(4): p. 505-23.
148. Mizushima, T., et al., *Decrease in expression of the master operon of flagellin synthesis in a dnaA46 mutant of Escherichia coli.* Biological & pharmaceutical bulletin, 1997. **20**(4): p. 327-31.

149. Soutourina, O., et al., *Multiple control of flagellum biosynthesis in Escherichia coli: role of H-NS protein and the cyclic AMP-catabolite activator protein complex in transcription of the flhDC master operon.* Journal of Bacteriology, 1999. **181**(24): p. 7500-8.
150. Shin, S. and C. Park, *Modulation of flagellar expression in Escherichia coli by acetyl phosphate and the osmoregulator OmpR.* Journal of Bacteriology, 1995. **177**(16): p. 4696-702.
151. Clarke, M.B. and V. Sperandio, *Transcriptional regulation of flhDC by QseBC and sigma (FliA) in enterohaemorrhagic Escherichia coli.* Molecular Microbiology, 2005. **57**(6): p. 1734-49.
152. Krin, E., A. Danchin, and O. Soutourina, *RcsB plays a central role in H-NS-dependent regulation of motility and acid stress resistance in Escherichia coli.* Research in microbiology, 2010. **161**(5): p. 363-71.
153. Shi, W., et al., *DnaK, DnaJ, and GrpE are required for flagellum synthesis in Escherichia coli.* Journal of Bacteriology, 1992. **174**(19): p. 6256-63.
154. Takaya, A., et al., *The DnaK chaperone machinery converts the native FlhD2C2 hetero-tetramer into a functional transcriptional regulator of flagellar regulon expression in Salmonella.* Molecular Microbiology, 2006. **59**(4): p. 1327-40.
155. Hughes, K.T., et al., *Sensing structural intermediates in bacterial flagellar assembly by export of a negative regulator.* Science, 1993. **262**(5137): p. 1277-80.
156. Kutsukake, K. and T. Iino, *Role of the FliA-FlgM regulatory system on the transcriptional control of the flagellar regulon and flagellar formation in Salmonella typhimurium.* Journal of Bacteriology, 1994. **176**(12): p. 3598-605.
157. Ohnishi, K., et al., *Gene fliA encodes an alternative sigma factor specific for flagellar operons in Salmonella typhimurium.* Mol Gen Genet, 1990. **221**(2): p. 139-47.
158. Donnelly, M.A. and T.S. Steiner, *Two nonadjacent regions in enteroaggregative Escherichia coli flagellin are required for activation of toll-like receptor 5.* J Biol Chem, 2002. **277**(43): p. 40456-61.
159. Bahrani, F.K., et al., *Proteus mirabilis flagella and MR/P fimbriae: isolation, purification, N-terminal analysis, and serum antibody response following experimental urinary tract infection.* Infection and Immunity, 1991. **59**(10): p. 3574-80.
160. Rocha, S.P., J.S. Pelayo, and W.P. Elias, *Fimbriae of uropathogenic Proteus mirabilis.* FEMS Immunol Med Microbiol, 2007. **51**(1): p. 1-7.
161. Li, X., et al., *Identification of MrpI as the sole recombinase that regulates the phase variation of MR/P fimbria, a bladder colonization factor of uropathogenic Proteus mirabilis.* Molecular Microbiology, 2002. **45**(3): p. 865-74.
162. Mobley, H.L. and G.R. Chippendale, *Hemagglutinin, urease, and hemolysin production by Proteus mirabilis from clinical sources.* J Infect Dis, 1990. **161**(3): p. 525-30.
163. Bahrani, F.K., et al., *Construction of an MR/P fimbrial mutant of Proteus mirabilis: role in virulence in a mouse model of ascending urinary tract infection.* Infection and Immunity, 1994. **62**(8): p. 3363-71.

164. Zunino, P., et al., *Mannose-resistant Proteus-like and P. mirabilis fimbriae have specific and additive roles in P. mirabilis urinary tract infections.* FEMS Immunol Med Microbiol, 2007. **51**(1): p. 125-33.
165. Li, X., et al., *Use of translational fusion of the MrpH fimbrial adhesin-binding domain with the cholera toxin A2 domain, coexpressed with the cholera toxin B subunit, as an intranasal vaccine to prevent experimental urinary tract infection by Proteus mirabilis.* Infection and Immunity, 2004. **72**(12): p. 7306-10.
166. Scavone, P., et al., *Nasal immunization with attenuated Salmonella Typhimurium expressing an MrpA-TetC fusion protein significantly reduces Proteus mirabilis colonization in the mouse urinary tract.* J Med Microbiol, 2011. **60**(Pt 7): p. 899-904.
167. Scavone, P., et al., *Effects of the administration of cholera toxin as a mucosal adjuvant on the immune and protective response induced by Proteus mirabilis MrpA fimbrial protein in the urinary tract.* Microbiol Immunol, 2009. **53**(4): p. 233-40.
168. Pearson, M.M. and H.L. Mobley, *Repression of motility during fimbrial expression: identification of 14 mrpJ gene paralogues in Proteus mirabilis.* Molecular Microbiology, 2008. **69**(2): p. 548-58.
169. Simms, A.N. and H.L. Mobley, *PapX, a P fimbrial operon-encoded inhibitor of motility in uropathogenic Escherichia coli.* Infection and Immunity, 2008. **76**(11): p. 4833-41.
170. Li, X., et al., *Repression of bacterial motility by a novel fimbrial gene product.* EMBO J, 2001. **20**(17): p. 4854-62.
171. Mobley, H.L.T. and A.N. Simms, *Multiple genes repress motility in uropathogenic Escherichia coli constitutively expressing type 1 fimbriae.* Journal of Bacteriology, 2008. **190**(10): p. 3747-3756.
172. Kelley, L.A. and M.J.E. Sternberg, *Protein structure prediction on the Web: a case study using the Phyre server.* Nature Protocols, 2009. **4**(3): p. 363-371.
173. Roy, A., A. Kucukural, and Y. Zhang, *I-TASSER: a unified platform for automated protein structure and function prediction.* Nat Protoc, 2010. **5**(4): p. 725-38.
174. Zhang, Y., *Template-based modeling and free modeling by I-TASSER in CASP7.* Proteins, 2007. **69 Suppl 8**: p. 108-17.
175. Owens, R.J., et al., *The structure of NMB1585, a MarR-family regulator from Neisseria meningitidis.* Acta Crystallographica Section F-Structural Biology and Crystallization Communications, 2009. **65**: p. 204-209.
176. Alekshun, M.N., et al., *The crystal structure of MarR, a regulator of multiple antibiotic resistance, at 2.3 angstrom resolution.* Nature Structural Biology, 2001. **8**(8): p. 710-714.
177. Hong, M., et al., *Structure of an OhrR-ohrA operator complex reveals the DNA binding mechanism of the MarR family.* Molecular Cell, 2005. **20**(1): p. 131-141.
178. Kumarevel, T., et al., *Crystal structure of the MarR family regulatory protein, ST 1710, from Sulfolobus tokodaii strain 7.* Journal of Structural Biology, 2008. **161**(1): p. 9-17.

179. Saito, K., et al., *Mutations affecting DNA-binding activity of the MexR repressor of mexR-mexA-mexB-oprM operon expression.* Journal of Bacteriology, 2003. **185**(20): p. 6195-8.
180. Wilkinson, S.P. and A. Grove, *Ligand-responsive transcriptional regulation by members of the MarR family of winged helix proteins.* Curr Issues Mol Biol, 2006. **8**(1): p. 51-62.
181. Park, P.J., *ChIP-seq: advantages and challenges of a maturing technology.* Nat Rev Genet, 2009. **10**(10): p. 669-80.
182. MacBeath, G. and S.L. Schreiber, *Printing proteins as microarrays for high-throughput function determination.* Science, 2000. **289**(5485): p. 1760-3.
183. Tuerk, C. and L. Gold, *Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase.* Science, 1990. **249**(4968): p. 505-10.
184. Ellington, A.D. and J.W. Szostak, *In vitro selection of RNA molecules that bind specific ligands.* Nature, 1990. **346**(6287): p. 818-22.
185. Djordjevic, M., *SELEX experiments: new prospects, applications and data analysis in inferring regulatory pathways.* Biomol Eng, 2007. **24**(2): p. 179-89.
186. Zhao, Y., D. Granas, and G.D. Stormo, *Inferring binding energies from selected binding sites.* PLoS Comput Biol, 2009. **5**(12): p. e1000590.
187. Jolma, A., et al., *Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities.* Genome Res, 2010. **20**(6): p. 861-73.
188. Reiss, D.J. and H.L. Mobley, *Determination of the target sequence bound by PapX, a repressor of bacterial motility, in the flhD promoter using SELEX and high-throughput sequencing.* J Biol Chem, 2011.
189. Roulet, E., et al., *High-throughput SELEX SAGE method for quantitative modeling of transcription-factor binding sites.* Nature Biotechnology, 2002. **20**(8): p. 831-5.
190. Bulyk, M.L., *Discovering DNA regulatory elements with bacteria.* Nature Biotechnology, 2005. **23**(8): p. 942-4.
191. Zhang, Y., et al., *Model-based analysis of ChIP-Seq (MACS).* Genome Biol, 2008. **9**(9): p. R137.
192. Bailey, T.L. and C. Elkan, *Fitting a mixture model by expectation maximization to discover motifs in biopolymers.* Proc Int Conf Intell Syst Mol Biol, 1994. **2**: p. 28-36.
193. Ott, M., et al., *Analysis of the genetic determinants coding for the S-fimbrial adhesin (sfa) in different Escherichia coli strains causing meningitis or urinary tract infections.* Infection and Immunity, 1986. **54**(3): p. 646-53.
194. Spurbeck, R.R., et al., *Fimbrial Profiles Predict Virulence of Uropathogenic E. coli Strains: Contribution of Ygi and Yad Fimbriae.* Infection and Immunity, 2011.
195. Zhang, X. and H. Bremer, *Control of the Escherichia coli rrnB P1 promoter strength by ppGpp.* J Biol Chem, 1995. **270**(19): p. 11181-9.
196. Kozarewa, I., et al., *Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G+C)-biased genomes.* Nat Methods, 2009. **6**(4): p. 291-5.

197. Imai, M., et al., *Uncoupling of the Cytochrome P-450cam Monooxygenase Reaction by a Single Mutation, Threonine-252 to Alanine or Valine - a Possible Role of the Hydroxy Amino-Acid in Oxygen Activation.* Proceedings of the National Academy of Sciences of the United States of America, 1989. **86**(20): p. 7823-7827.
198. McCarter, L.L., *MotY, a component of the sodium-type flagellar motor.* Journal of Bacteriology, 1994. **176**(14): p. 4219-25.
199. Rowbury, R.J., *Regulatory components, including integration host factor, CysB and H-NS, that influence pH responses in Escherichia coli.* Letters in applied microbiology, 1997. **24**(5): p. 319-28.
200. Slattery, M., et al., *Cofactor binding evokes latent differences in DNA binding specificity between Hox proteins.* Cell, 2011. **147**(6): p. 1270-82.
201. Wang, L., et al., *Selection of DNA aptamers that bind to four organophosphorus pesticides.* Biotechnology letters, 2012.
202. Feng, H., et al., *A SELEX-screened aptamer of human hepatitis B virus RNA encapsidation signal suppresses viral replication.* PLoS One, 2011. **6**(11): p. e27862.
203. Zhang, K., et al., *A novel aptamer developed for breast cancer cell internalization.* ChemMedChem, 2012. **7**(1): p. 79-84.
204. Schutze, T., et al., *Probing the SELEX process with next-generation sequencing.* PLoS One. **6**(12): p. e29604.
205. Altschul, S.F., et al., *Basic local alignment search tool.* J Mol Biol, 1990. **215**(3): p. 403-10.
206. Langmead, B., et al., *Ultrafast and memory-efficient alignment of short DNA sequences to the human genome.* Genome biology, 2009. **10**(3): p. R25.
207. Wilbanks, E.G. and M.T. Facciotti, *Evaluation of algorithm performance in ChIP-seq peak detection.* PLoS One, 2010. **5**(7): p. e11471.
208. Djordjevic, M. and A.M. Sengupta, *Quantitative modeling and data analysis of SELEX experiments.* Phys Biol, 2006. **3**(1): p. 13-28.
209. Zimmermann, B., et al., *Genomic SELEX: a discovery tool for genomic aptamers.* Methods, 2010. **52**(2): p. 125-32.
210. Sampson, T., *Aptamers and SELEX: the technology.* World Patent Information, 2003. **25**(2): p. 123.
211. McClure, W.R., *Mechanism and control of transcription initiation in prokaryotes.* Annu Rev Biochem, 1985. **54**: p. 171-204.
212. Gabanyi, M.J., et al., *The Structural Biology Knowledgebase: a portal to protein structures, sequences, functions, and methods.* J Struct Funct Genomics, 2011. **12**(2): p. 45-54.
213. Sjostrom, A.E., et al., *The SfaXII protein from newborn meningitis E. coli is involved in regulation of motility and type 1 fimbriae expression.* Microbial pathogenesis, 2009. **46**(5): p. 243-52.
214. Hultdin, U.W., et al., *Structure of FocB--a member of a family of transcription factors regulating fimbrial adhesin expression in uropathogenic Escherichia coli.* The FEBS journal, 2010. **277**(16): p. 3368-81.

215. Svensson, A., et al., *Design and evaluation of pilicides: potential novel antibacterial agents directed against uropathogenic Escherichia coli.* Chembiochem : a European journal of chemical biology, 2001. **2**(12): p. 915-8.
216. Nicholson, T.F., K.M. Watts, and D.A. Hunstad, *OmpA of uropathogenic Escherichia coli promotes postinvasion pathogenesis of cystitis.* Infection and Immunity, 2009. **77**(12): p. 5245-51.
217. Bonacorsi, S., V. Houdoin, and E. Bingen, *[Virulence factors associated with E. coli neonatal meningitis]*. Archives de pediatrie : organe officiel de la Societe francaise de pediatrie, 2001. **8 Suppl 4**: p. 726s-731s.
218. Mobley, H.L.T. and M.M. Pearson, *Repression of motility during fimbrial expression: identification of 14 mrpJ gene paralogues in Proteus mirabilis.* Molecular Microbiology, 2008. **69**(2): p. 548-558.
219. Mobley, H.L., et al., *Pyelonephritogenic Escherichia coli and killing of cultured human renal proximal tubular epithelial cells: role of hemolysin in some strains.* Infection and Immunity, 1990. **58**(5): p. 1281-9.
220. Harshey, R.M. and T. Matsuyama, *Dimorphic transition in Escherichia coli and Salmonella typhimurium: surface-induced differentiation into hyperflagellate swarmer cells.* Proceedings of the National Academy of Sciences of the United States of America, 1994. **91**(18): p. 8631-5.
221. Datsenko, K.A. and B.L. Wanner, *One-step inactivation of chromosomal genes in Escherichia coli K-12 using PCR products.* Proceedings of the National Academy of Sciences of the United States of America, 2000. **97**(12): p. 6640-6645.
222. Vigil, P.D., et al., *Presence of Putative Repeat-in-Toxin Gene tosA in Escherichia coli Predicts Successful Colonization of the Urinary Tract.* MBio, 2011. **2**(3).