VALIDATION OF EFFECTS OF SNV FROM GWAS ON BIOFILM PHENOTYPE USING A CRISPR SYSTEM WITH UNIVERSAL GUIDE RNA SEQUENCE

A Thesis
Submitted to the Graduate Faculty
of the
North Dakota State University
of Agriculture and Applied Science

By

Allan Kelvin Gramillo

In Partial Fulfillment of the Requirements
for the Degree of
MASTER OF SCIENCE

Major Department:
Microbiological Sciences

June 2021

Fargo, North Dakota

North Dakota State University
Graduate School

Title

Validation of effects of SNV from GWAS on biofilm phenotype using a CRISPR system with universal guide RNA sequence

By

Allan Kelvin Gramillo

The Supervisory Committee certifies that this ***disquisition*** complies with North Dakota State University's regulations and meets the accepted standards for the degree of

MASTER OF SCIENCE

SUPERVISORY COMMITTEE:

Peter Bergholz
Chair

Sarah Signor

Penelope Gibbs

Approved:

July 7, 2021
Date

John McEvoy
Department Chair

ABSTRACT

Extraintestinal *E. coli* have adapted to survive in secondary environments outside of the intestines of a host. The genetically diverse phylogroup D was use to validate two previous GWAS analysis that identified single nucleotide variants to have a positive or negative effect on biofilm formation. We selected ten variants from TreeWAS based on current literature to identify patterns among a new set of *E. coli* isolates, to identify SNV that can be used as genetic markers for biofilm formation. DBGWAS was used to predict and validate the predicted value of a SNV, by conducting an allelic exchange by using a CRISPR/Cas9 system, pCAGO, that allowed for scarless genome editing. This allelic exchange and deletion of the gene led to no statistical significance when biofilm formation and growth rate were tested. This led to variation in biofilm formation but didn't affect the phenotype in a statistically significant manner.

ACKNOWLEDGMENTS

I would like to express my gratitude and appreciation to a Dr. Peter Bergholz, for giving me the opportunity to work in his laboratory. I would like to acknowledge Kaycie Schmidt and Madelyn Schwartz for their technical advice and words of encouragement. Additionally, I would like to thank Dr. Birgit Pruess and Shelley Horne for their technical advice and guidance throughout my time at NDSU. In addition, I would like to thank my supervisory committee, Dr. Sarah Signor and Dr. Penelope Gibbs. I'd also like to acknowledge my funding sources that gave me the opportunity to complete my research North Dakota State University: National Science Foundation award no. DEB-1453397 to P.W.B. , Department of Microbiological Science and ND-EPSCoR Preliminary Data Proposal (Award no:32091). Lastly, I would like to thank everyone in the Department of Microbiological Sciences for being so welcoming and available for assistance.

DEDICATION

I dedicate my work to my siblings, Jeffrey and Jenny Gramillo and my best friend Edwin Loarca for the support that was provided outside of school. I would also like to dedicate this to my parents, Efren and Hilda Gramillo, for their amazing support that allowed me to reach my goals.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: LITERATURE REVIEW

Escherichia coli life cycles increase the Pangenome

Escherichia coli is a gram-negative species that is commonly found as part of the lower intestinal microbiota of humans and warm-blooded mammals (1–5). Fecal contamination containing *E. coli* continues to be a problem in our water systems, soil, and food (1, 5–8).This has led to the study of *E. col*i outside of a host intestine and how they have adapted to a variety of environments (9–12). To understand the life cycle of *E. coli*, we need to understand how they can adapt to different environments from the intestinal tract, urinary tract, and outside of a host (11–15).

E. coli reproduce asexually and the requirement to adapt to new ecological niches leads to a faster evolution as this provides a competitive advantage over residing microbes within the host or outside of a host (11, 13–15). When studying *E. coli* host colonization in a mouse gut, it was shown that invading *E. coli* can successfully colonize the gut by undergoing phage-mediated horizontal gene transfer followed by mutations that improve the metabolism of specific gut carbons, mannose and gluconate. In this same study, the resident ancestor strain initially had a higher growth rate and yield when compared to the invading strain, but phage-driven horizontal gene transferred led to the evolved invading *E. coli* having a competitive advantage in growth rate and yield (16).

The evolutionary potential of a species depends on pre-existing genetic diversity as well as de novo mutations which both may give rise to populations that have an advantage in a novel environment (17, 18). Rapid changes in an environment can lead to population decline and extinction if the change is lethal to the cells. It has been demonstrated in *Pseudomonas fluorescens* SBW25, that initial genetic diversity and population size play a role in the evolution rescue of the species when exposed to a lethal environment, streptomycin (19). A larger population will contain a larger standing variation then small populations, which can improve survival when exposed to a lethal environment. This was shown in *P. fluorescens,* as the 0.2 mL population had a very low rescue under streptomycin stress when compared to the 1.5 mL population. Additionally, when population density decreased to very low levels before population rescue occurred, it is possible that a de novo mutation occurred giving rise to streptomycin resistant mutants or there was a significant lag in exponential growth from cells that conferred resistance (19).

The differences between the host environment and secondary environments may allow for a quicker response as nutrient available may promote growth of populations that weren't dominant in a previous environment (15, 17). For example, when *E. coli* is deposited into the environment, they may face predation from soil dwelling amoeba that may cause death to the cells unless certain mutations improve survival. Macrophages and amoeba share similar characteristics such as phagocytosis and autophagy so the ability to evade phagocytosis may be beneficial within a host and outside of a host (21). In a study to identify how commensal *E. coli* strains became more pathogenic within the host, constant macrophage pressure on commensal *E. coli* led to rapidly adapted isolates (22). By acquiring mutations that lead to changes in the transcriptome they had an improved intracellular survival and ability to escape macrophages (22, 23). Pathogens have been shown to have an increased mutation and recombination rate which has led to distinct metabolic and pathogenic capabilities (11, 24)

Neutral Theory states that most genetic variability can be attributed to neutral or near neutral mutations opposed to all mutations being adaptive (25–28). The constant environmental pressures lead to genetic changes that may confer an advantage, deleterious, or neutral but often they are neutral (27). This has been previously demonstrated in *fimH*, in which there is an accumulation of mutations that eventually lead to increase binding to mannose residues within the urinary tract. Most of these mutations are neutral mutations until *E. coli* disperse from the intestine into the urinary tract. *E. coli* that infects the urinary tract must be able to retain the ability to adhere to monomannose and survive shear stress if they will colonize the urinary tract (23). The functional trade-off has been demonstrated when a mutation within *fimH* led to unusually strong binding to monomannose and increased colonization of the mouse bladder by 20-fold. The also led to the adhesion being inhibited under shear stress (29). Mutations can improve the functionality of a protein but that can also be detrimental under different conditions.

E. coli also can survive outside of the intestine of a host such as surviving in and on plants, soil, and cause extraintestinal infections such as neonatal meningitis and urinary tract infections (8, 30–35). Uropathogenic *E. coli (UPEC)* are able to utilize D-serine as a carbon and nitrogen source as is commonly found within the urinary tract and it. Deletion of the *dsdA* gene led to growth defects leading to a long lag phase in the urinary tract of mice (15, 31, 32) . The acquisition of genes leads to adaptation to a new environment (9, 33).

E. coli gene content diversity is due to the acquisition or loss of genes and pathogenicity islands that cause variability in genome size. In a comparison of genome size of 3 different strains, it was shown that *E. coli* K-12, intestinal pathogen O157:H7 and extraintestinal pathogen *E. coli* CFT073 shared a core genome of 2,996 protein-coding genes. The differences among these three isolates is their ability to cause disease and phylogroup; K-12 is B1- like, CFT073 is B2 and EDL933 is F. The two pathogens contained larger genome when compared to the size of K-12 and this is likely due to pathogenicity islands (34). The *E. coli* K-12 genome is 4,401 gene , CFT073 contains 5,379 genes and EDL933 contains 5,449 which presents the genome variability within the *E. coli* species (34). The genome variability of *E. coli* has been displayed by its pangenome, with the accessory genes counting for a large portion of the content. Depending on the study, the core genome is predicted to be about 1,700-2,500 gene families and pangenome (9, 35, 36). This wide distribution in genome size can be attributed to the eight phylogroups that are associated with numerous niches leading acquisition of new genes (37, 38). There are eight phylogenetic groups A, B1, B2, C, D, E, F and G, with the majority of extraintestinal infections *E. coli* being caused by phylogroup B2 and D (39, 40).

Extraintestinal *E. coli* tend to have a larger genome, when compared to commensal isolates as they are able to occupy numerous niches. *E. coli* that are outside of a host, require different genes to survive when compared to *E. coli* residing within a host (36). The ability to survive in water, soil, and grasslands shows the wide distribution of *E. coli* (1, 41–43). Soil is a reservoir for *E. coli* thus allowing for genomic variations to increase based on heterogeneity of environmental characteristics within soil (1, 6, 41, 44). The ability to survive in different environments is influenced by the ability to endure harsh conditions by forming microbial communities called biofilm (4, 45–47).

Biofilms: a developing survival mechanism

A survival mechanism that is commonly used is the ability to form a microbial community called biofilm (4, 48). Within a biofilm, microbes have a higher survivability rate against external threats as it is a protective barrier against antibiotics, desiccation and also acts as a nutrient reservoir with increased genetic exchange which drives natural selection (37, 49–51). 40-80% of microbial life is in biofilms (52, 53). Microbes can thrive within a biofilm because they can feed off one another's metabolic by-products and increase structural integrity as biofilms are composed of a variety of bacteria, fungi, viruses, and protists (49, 54–58).

The environment leads to different biofilm structures as they adapt to new external factors. This has been shown in *Pseudomonas aeruginosa* biofilm as the cells retain their rod shape and form a mushroom like matrix during oxic conditions (73). When *P. aeruginosa* is placed under denitrifying conditions the majority of the cells become filamentous and creates a mesh-like structure (77). The environmental effect on biofilm formation has also been demonstrated in *Clostridium perfringens.* The biofilm formation at 37˚C was densely packed but in ambient temperatures, 25˚C, the cells had less attachment activity and formed elastic thick pellicle-like biofilms (62). The ability for the environment to affect biofilm structure is also affected by the surface that the biofilm is being formed. In *E. coli,* biofilm production has been shown to be dependent on the surfaces hydrophobic or hydrophilic properties. During growth in hydrophobic surfaces, *ompA* was shown to increase biofilm formation but decreased biofilm formation on hydrophilic surfaces due to the production of cellulose (78). The environment alters the biofilm structure as well as its ability to form biofilms, indicating the various genes involved with biofilm formation (62, 67, 79–81).

A biofilm can be constructed with dominant microorganisms such as *E. coli* in urinary tract infections or a multitude of microbes as in dental plaque (27, 59). A single species biofilm would contain a variety of cells in a motile or sessile phenotype(60) Sessile and planktonic cultures of *E. coli* released metabolites which improve growth of single species biofilm but also dual species biofilm (61).

Environmental factors such as low temperature, osmotic stress, host immune response, and nutrient availability are known to affect the expression of a variety of genes in a microbe leading to biofilm formation (47). When *E .coli* K-12 was grown at low temperatures there was an upregulation of genes for biofilm, cold-shock and *rpoS* dependent genes (47). For example, an increase in temperature leads to a different biofilm matrix (62). Additionally, biofilm formation is a survival mechanism in which the metabolic burden should be decreased to increase the chance of survival. This was shown in a study in which 11 of the 100 least ATP consuming proteins in *E. coli* K-12 were extracellular proteins, including: curli, major subunit of flagella and type 1 pili all of which are critical during biofilm formation (63).

Biofilm formation consist of five step life cycle which consist of the initial reversible attachment, irreversible attachment, microcolony formation, maturation and dispersion, with the initial attachment and irreversible attachment overlapping (64) . During biofilm formation, each cell may be in a different stage of the life cycle.

During the initial stage of biofilm formation, cells must use energy to swim to a biotic or abiotic surface using their flagella for the initial adherence to surface. The flagellum master regulator, *flhDC*, changes expression levels when exposed to different pH or osmolarity levels (65) . In a gene knock out study, it was found that cells lacking the genes to form flagella (*fliCD)* or for motility, *motA*B*,* had a decrease ability in biofilm formation. In the same study, deletion of the type 1 fimbrial protein gene, *fimH,* led to a nearly no visible cells adhering the abiotic surface, polyvinyl chloride (PVC) under a microscope (66).

High levels of c-di-GMP have been shown to increase biofilm formation and low levels keep the cells motile (67, 68). As mention previously, initial reversible/irreversible attachment, overlap with each other as the conversion from planktonic to sessile is regulated by the intracellular levels of cyclic diguanylate monophosphate (c-di-GMP), a second messenger molecule (68, 69). Another regulator during this initial adhesion process, is the catabolite repressor protein (CRP) - cyclic AMP (cAMP) complex that can directly affect curli fiber synthesis by activating the master curli regulator, *csgD*, positively regulate the master motility regulator *flhDC* and inhibition of *rpoS* during this initial stages of biofilm formation (70). The transition to maturation stage continues to have transient cells that are affected by environmental conditions as the biofilm matrix is a continuously evolving matrix (68, 71–73).

During this maturation stage, cells begin to produce extracellular polymeric substance (EPS) that is composed of water and extracellular polysaccharides, proteins, lipids and extracellular DNA (eDNA) (73–75). This EPS matrix is a protective barrier allowing for the cells to remain in a hydrated state, retain nutrients within the matrix and protects against environmental stress such antibiotic treatment, pH changes and salinity (73, 76). The maturation stage of biofilm formation is when cells undergo exponential growth (75). In a recent study, *Streptococcus mutans* biofilm formation consists of two different cells types, cells actively replicating forming microcolonies and cells that remain in a steady state. The actively growing cells contributed to the increasing cell density with EPS production and eventually intertwining the various microcolonies to form the matured biofilm (60, 75).

The final stage of biofilm formation is the dispersal of the attached cells which is done by active or passive dispersal with three different modes of actions: erosion, sloughing, and seeding (82). All three mechanisms can be considered active dispersal, but erosion and sloughing can be either passive or active. Sloughing has been demonstrated to be active in *P. aeruginosa* as nutrients are increased (89). In *S. marcescens*, it has been demonstrated that sloughing in controlled by quorum sensing (90). The degradation of c-di-GMP by phosphodiesterases (PDEs) due to environmental signals such as nitrious oxide are also a factor in biofilm dispersal (67, 69, 91).

Lastly, the duration of biofilm can be highly dependent on the microbes involved in the biofilm formation. The wild-type *E. coli* K-12 MC1000 strain is highly motile and has biofilm degradation after two weeks on polystyrene surface under static conditions in TB media. When grown under similar condition in a mixture of wild-type MC1000 and non-motile MC1000 mutants, biofilm mass increased after three weeks when compared to the non-mixed biofilm (60). Mixed biofilms had increased cell viability and an increase in biofilm mass when compared to the single MC1000 biofilm. This indicates that cell a combination of motile and nonmotile cells improve biofilm formation (60).

Genetic adaptation is crucial for the survival of any microorganism as this allows them to endure external environments where the fluctuation in temperature and oxygen may play a critical role in biofilm architecture (11, 62, 67–69, 81). Biofilm formation is an adaptive phenotype to the environmental factors that alter gene expression (64, 65, 69, 78, 79, 92–94). Biofilm formation is an epistatic phenotype as it is not controlled by a single gene, rather it is the result of various genes being regulated by one another and environmental factors(45, 62, 65, 81). This leads to the importance of statistical models, that can identify associations between genetic variants and a phenotype.

Genome-wide associated studies

Genome-Wide Associated Studies (GWAS) have been of use to identify genes or single nucleotide polymorphisms (SNP) associated with disease. Various statistical tools are currently being used to identify an association between genes or single nucleotide polymorphisms and a phenotype (95–99). Statistical tools all contain an advantage and disadvantage. Such as TreeWAS, it is a phylogenetic method for GWAS in which it uses a reference genome, accounts for epistasis by allowing each SNP to partially contribute to the phenotype and corrects for homologous recombination but the use of a reference genome may not account for variants that occur rarely (100, 101). Another tool that has been used is DBGWAS, which doesn’t use a reference genome, extracts novel variants, and identifies local polymorphisms and mobile genetic elements but this tool may lead to identification of various hypothetical proteins or genes associated with bacteriophages (101, 102). One of the major disadvantages of DBGWAS is that it doesn't completely account for phylogeny in contrast to TreeWAS does. (101, 102). These two separate tools can identify large amounts of variants as do many models, so the need to validate these studies continues to be of great importance (101, 103).

The increase use of bacterial GWAS has led to the identification of association between genes and a phenotype in antibiotic resistance, cancer and also virulence factors (95, 99, 104). To date, there has been few papers in which GWAS results were validated on naturally occurring variants that leads to a specific phenotype. GWAS has been used to identify genes associated with antibiotic resistance as well as identifying genetic variations within *Helicobacter pylori* associated with gastric cancer (95, 105). A variety of results can come from a single GWAS, as was the case for *Mycobacterium tuberculosis*, where they identification of one gene, *ponA1*, was associated with antibiotic resistance. In addition, isolates with unknown sources of antibiotic resistance were not associated to any specific gene (104). In another study, an association between virulence and iron capture systems within a high pathogenicity island containing the siderophore yersiniabactin was associated with a high death rate in mice and growth in the presence of numerous antibiotics (99). There have been various studies that find association between gene and phenotype but not many studies on naturally occurring isolates. A study on *Campylobacter jejuni*, identified various genes associated with disease and survival were identified, and deletion mutants of the identified genes were validated the association (106).

Genome editing tools for allelic exchanges

To identify how a specific SNP may affect a phenotype, the simplest way is to swap alleles containing the desired SNP. There are a variety of genetic tools that can be used to create allelic exchanges but a variety of problems have been associated with these systems.

Recombineering has been used extensively as a method to disrupt gene functionality via insertion, deletions and point mutations (107–109). A genetic tool that has taken advantage of bacteriophage proteins is the lambda Red recombination system that was taken from the bacteriophage lambda which requires three proteins Gamma, Beta and Exo proteins that normally would help the lambda phage integrate itself into *E. coli* chromosome. These three proteins have been cloned on to a plasmid which are usually under the control of a *lac*-repressor (110). Recombineering has been useful tool as the requirements for the insertion of the editing cassette requires as little as 35bp homologous arms flanking an antibiotic marker as a method to indicate that the desired insertion has been obtained (109). Recombineering has been a useful tool in the metabolic engineering field as researchers have enhanced insulin production by removing genes that interfere or divert resources from insulin production (111). Its major advantage is that it is not limited to restriction enzymes which allows the insertion to be made into plasmid, chromosome and improve bacterial artificial chromosomes (BAC) engineering. Lambda Red recombination was originally introduced as a method to replace genes with antibiotic markers and remove the antibiotic marker with the, FLP recombinase target (FRT), flanking the target which eventually left a scar at the site which can be as short as 36bp and as long as 85bp (109). This method has been crucial for our understanding of functionality of these genes that were knocked out leading to the Keio collection (112). The Keio collection, consist of single gene deletion of all non-essential genes in *E. coli* K-12, which led to the collection of

3,985 *E. coli* mutants (112). Leaving behind a scar site can lead to unexpected results that may affect cell growth causing false positive results. When studying genetic variability among wild type isolates, we want to only introduce our desired mutation as undesired scar sites have shown to have a negative impact on the growth rate of *Salmonella enterica* when a FRT scar was left near the open reading frame of *rpsT* (113).

Based on the review by Gaj *et al*, we considered the following genome editing techniques for use in validating GWAS results: Zinc Finger Nucleases (ZFN), Transcription activator-like effector nucleases (TALEN) and Clustered, Regularly Interspaced, Short, Palindromic Repeats (CRISPR)/ CRISPR associated nuclease(Cas) (114). These three tools have the benefits of genome editing but the ability to be cost saving and precise were key consideration for the present research. Zinc Finger Endonucleases and TALEN are considered for gene manipulation as they are considered artificial restriction enzymes, with a cleavage domain that can be designed to have high affinity to double stranded sites (115, 116). ZFN and TALEN both use DNA binding modules that can be customized for each desired target by purchasing modular assembly kits available at Addgene (117, 118). The kit availability of ZFN and TALEN increases the accessibility of genome editing tools but can be time consuming to properly design and construct (114, 118–120). The reason zinc fingers weren't chosen is that they aren't used in bacteria, and zinc finger design requires validation to improve affinity to the desired region due to interactions among zinc fingers (119). The second genome editing tool TALEN, is simplified using customized module plasmids which uses golden gate cloning to complete the final construct. The ability to recognize one single nucleotide within the desired target region based on the protein-DNA interaction (114). The reason TALEN proteins weren't selected was due to the high cost and it is time consuming.

The genome editing tool that I chose to use was a CRISPR/Cas9 system that combines the usage of lambda Red recombineering to improve intrachromosomal recombination leaving a scarless edit (121). CRISPR/Cas9 system relies on the endonuclease Cas9, which recognizes a 20 base pair target sequence along the chromosome followed by a protospacer adjacent motif (PAM) that is approximately 2-6 base pairs upstream which allows for sequence recognition (122). The PAM sequence limitation in CRISPR genome editing was avoided with the usage of a universal N20PAM sequence(123–125). In addition to PAM dependent limitation, to target various genomic sites, a new guide RNA (gRNA) needs to be constructed for every site desired to create a mutation (114). A new system that eliminates the PAM restriction was recently designed in combination with lambda Red recombination, which are regulated by a single plasmid (121). This novel system uses a universal target sequence(N20PAM), that is inserted via homologous recombination and a universal gRNA encoded on the plasmid which recognizing the N20PAM sequence leading to the double stranded break by Cas9 endonuclease. This new CRISPR/Cas9 system consist of a plasmid, pCAGO, which eliminates the need to generate new gRNA for every target site, as the target N20PAM sequence is integrated into the desired region via homologous recombination (121). This system can be used in genomic regions that don't contain a recognized PAM sequence or CRISPR intolerant regions. This genome editing also does not require multiple plasmids or multiple transformations, so this simplistic approach eventually allowed us to genetically modify an isolate with the initial transformation and induction of lambda Red proteins and Cas9 endonuclease (114, 121, 124, 126, 127).

Rationale of research

Biofilm formation is a complex phenotype that requires a combination of genes or genetic variations. *E. coli* is known to form biofilms and within the Peter Bergholz lab the biofilm phenotype has been shown to be variable across phylogroup D environmental isolates with some of these isolates being closely related to the uropathogenenic *E. coli* strain UM026. Previous GWAS analysis has identified single nucleotide variants associated with biofilm formation in Minnesota and North Dakota *E. coli* soil isolates. One of the goals of this research was to test the generalizability of two different GWAS results to a separate *E. coli* population from New York soil. This generalization would allow us to identify genetic variations patterns that can later be used as genetic markers for biofilm formation in newly sequenced *E. coli* genomes. We used ten genes that are part of the core genome with known roles in biofilm formation so the presence or absence of the SNV within the gene may alter the genes functionality and alter the biofilm phenotype. Additionally, validation of GWAS was completed on a SNV that was associated with increased biofilm density. The SNV was identified by DBGWAS analysis, so an allelic exchange between two Minnesota/ North Dakota isolates was conducted to validate the true effect of this genetic variant. To complete the allelic exchange, pCAGO a CRISPR/Cas9 system was used because it doesn't require the design of multiple guide RNA for each target as it contains a universal N20PAM sequence. The pCAGO system allowed us to complete an allelic exchange and gene deletion with just a change in primer design. Once an isolate contains the pCAGO plasmid, it can be used for gene deletion, addition, and allelic exchange.

CHAPTER 2: VALIDATION OF EFFECTS OF SNV FROM GWAS ON BIOIFLM PHENOTYPE USING A CRISPR SYSTEM WITH UNIVERSAL GUIDE RNA SEQUENCE

Introduction

The dual lifestyle of *Escherichia coli* allows for survival in diverse environments within a host and outside of a host such as freshwater, plants and soil (1, 6, 45, 128). The open pangenome of *E. coli* has undergone a complex process in which the environment has selected for genetic traits that increase survival (9, 70, 129). With the large genomic diversity of *E. coli*, it is important to focus on a single phylogroup to remove additional variable during the research. In this research, phylogroup D is the focus because its genetic background has shown to be very diverse in its ability to adapt to different environments which isolates are commonly associated with urinary tract infections, neonatal meningitis but can also be found in soil and water (9, 28, 44, 130, 131). Biofilm formation is a survival mechanism that improves survival in these different environments and increases genetic variability (132). Biofilm formation is a sessile phenotype that is a complex network of genes that interact in an epistatic fashion that contributes to the difficulty in identifying the roles of genes or SNPs. The importance of identifying how a single gene or point mutation can alter the phenotype continues to be studies due to its clinical importance and to improve our understanding of the microbial evolution.

A method still in use, is transposon mutagenesis as it identifies the association between gene disruption and a phenotype. Transposon mutagenesis has identified intergenic regions importance in antimicrobial tolerance, pathoadaptive traits in *E. coli*-macrophage interactions and the cause of a single point mutation in the intergenic region of *csgDEFG* and *csgBAC* led to biofilm formation and invasive capability in E. coli 0157:H7 (20, 133, 134). This has led to the identification of genes essential for biofilm formation, growth and invasion of epithelial cells (76, 135). An alternative method is the use of Genome-Wide-Associated Studies to identify naturally occurring variants and their associations with a phenotype. (97, 99). The ability to understand natural variation associations with survival and niche-specific genes has been limited to *Campylobacter jejuni,* avian *E. coli* and non-clinical *Aspergillus fumigatus* as of most recently (105, 106, 136).

The first objective of the research was to test the generalizability of GWAS on a new set of *E. coli* isolates that are from New York soil. If successful, genetic variants can be used in clinical settings where the genomes sequence is enough information to know the isolates biofilm capabilities. The goal was to identify a pattern among the genetic variants that are associated with positively or negatively impacting biofilm formation This would also allow for quick identification of good biofilm formers and allow for better treatment for patients.

The second goal was to validate GWAS, by creating an allelic exchange between isolates that contain a genetic variant that has been associated with a positive effect on biofilm formation. The allelic swap will be conducted using a CRISPR/Cas9 editing tool to create a mutation without a scar site (121). The target gene produces a hypothetical protein, which contains a conserved domain of *ddrB*-like *ParB* superfamily domain (e-value =1.62e-47) and was also annotated as a D-serine transporter during the DBGWAS analysis (e-value 6). The allelic exchange between to isolates would show the impact that this specific nucleotide has on biofilm formation. The isolate obtaining the nucleotide from the better biofilm producer should see an improvement in biofilm formation. In addition, deletion of the whole gene was be conducted as various nucleotides may be required to see a phenotypic effect.

Results

Validation of TreeWAS and DBGWAS analysis using New York isolate ability to form biofilm

Biofilm formation was measured using the crystal violet assay due to the ease and ability to screen a large quantity of isolates. The crystal violet assay is a method that allows us to quantify the biomass produced during biofilm formation by measuring the optical density. Some of the best techniques to measure biofilm formation is confocal scanning microscopy as it can give a 3D view of the biofilm and allows for the least disruption of the natural biofilm.

From a collection of 131 Phylogroup D *E. coli* soil isolates from the Cornell University Food Safety Laboratory collection, we used 40 isolates to test the generalizability of the TreeWAS results from Minnesota/North Dakota *E. coli* isolates (130, 137) . In addition, a subset of 21 was used from the original 40 isolates because not all the isolates contained the variants identified by DBGWAS.

Seven replicates of the crystal violet assay were completed to validate the consistency of results and to identify a pattern among the isolates that contain the genetic variants. Ten single nucleotide variants were selected from the TreeWAS analysis to assess the predictability of the biofilm phenotype. A positive score, such as 53.6 in the *kpsS* variant (figure1), indicates that the presence of the variant has a predicted positive effect on biofilm formation and a negative score indicates a negative effect on biofilm formation. If a pattern is present, the top or bottom biofilm formers would contain the same SNP within the gene. For example, if the alternative variant within the *kpsS* gene increased biofilm formation, the top biofilm formers would contain the alternative variant more often while the poor biofilm formers would have the reference variant regardless of the other genes.

Among our isolates the wide diversity within these isolates is shown as the difference in optical density (OD) 580 nm between the top biofilm former and poorest biofilm former is an OD_{580} 1.64 (figure 1) . The uropathogen *E. coli* UMN026 was used as a control as it is a good biofilm former and was used as a reference genome during the TreeWAS that was completed previously (130, 138). There is five different isolates that had a OD_{580} similar or higher than UMN026. This indicates that these five isolates have the ability to have a biomass equivalent to or greater than UMN026 but it does not mean that this would translate to pathogenicity or ability to form biofilms within the urinary tract. The overall average OD_{580} of the top five biofilm formers is 1.47 when compared to UMN026 which had an OD_{580}= 1.24.

To assess the predictability of the TreeWAS variants selected, we examined SNPs present in genes known to cause an effect on biofilm formation (figure1). If the reference or alternative SNP were more prevalent in the top or worse biofilm formers it would allow us to state that there is a pattern that is consistent with good/poor biofilm formation. Predictability was based on the presence/absence of the 10 variants from the TreeWAS analysis. The ten variants within the ten different genes did not have any pattern to confirm any single variant influencing biofilm formation. The top biofilm former (Figure 2), FSL-B7-2805 and the 36th rank biofilm former FSL-B7-2129 contained the same profile within the 10 selected variants indicating that there may be a variety of different single nucleotide variants affecting biofilm formation between the isolates. TreeWAS analysis had previously identified 1,089 single nucleotide variants in core functional genes therefore it is possible that there is a single SNV or combination of SNV that can be used as a biofilm predictor.

	kpsS	dsdA	pduT	iucD	yfaL	entF	fimA	evgS	dcp.	flgG.
Isolate	(53.61)	(45.92)	(46.21)	(49.75)	(-44.57)	(-47.80)	(-48.11)	(-61.56)	(-44.37)	(-45.77)
FSL-B7-2805	C	C	G	A	T	G	A	G	A	G
FSL-B7-1564	▨	A	A	T	T	G	C	G	T	G
FSL-B7-1676	C	C	G	A	T	G	A	G	A	G
FSL-B7-2676	▨	▨	G	▨	C	G	A	G	T	G
FSL-B7-2728	C	C	G	A	C	G	A	G	T	G
UMN026	C	C	G	A	T	A	C	A	T	G
FSL-B7-1508	C	▨	G	A	C	G	A	G	T	G
FSL-B7-2774	T	A	A	T	T	G	C	A	T	G
FSL-B7-2788	▨	A	A	T	T	G	C	G	T	G
FSL-B7-1620	▨	A	A	T	C	G	C	G	T	G
FSL-B7-0711	▨	C	G	A	C	G	C	A	A	G
FSL-B7-1107	C	C	G	A	C	G	A	G	A	G
FSL-B7-2775	T	A	A	T	C	G	C	A	T	G
FSL-B7-1617	▨	A	A	T	T	G	C	G	T	G
FSL-B7-1108	C	A	A	T	C	G	C	G	A	G
FSL-B7-0360	▨	C	G	A	T	G	A	G	A	G
FSL-B7-0589	▨	A	A	T	T	G	C	G	T	G
FSL-B7-0709	C	▨	G	A	C	G	A	A	A	G
FSL-B7-1063	C	C	G	A	T	G	A	G	A	G
FSL-B7-2800	▨	A	G	A	C	G	A	G	A	G
FSL-B7-1109	▨	▨	G	A	C	G	A	A	A	G
FSL-B7-1110	▨	▨	G	A	C	G	A	G	A	G
FSL-B7-2543	▨	A	A	T	T	G	C	G	T	G
FSL-B7-0688	▨	C	G	A	T	A	C	A	T	G
FSL-B7-2793	▨	▨	G	A	C	G	A	G	A	G
FSL-B7-1085	▨	C	G	A	C	A	C	G	A	G
FSL-B7-2807	C	C	G	A	T	G	C	A	A	G
FSL-B7-0624	C	▨	G	A	C	G	C	G	A	G
FSL-B7-0590	▨	▨	G	A	C	G	A	G	A	G
FSL-B7-0392	▨	A	A	T	T	G	C	G	T	G
FSL-B7-0613	▨	A	A	T	T	G	C	G	T	G
FSL-B7-2629	C	C	G	A	T	G	A	G	A	G
FSL-B7-2677	T	A	A	T	C	G	C	A	T	G
FSL-B7-2131	T	A	A	T	C	G	C	A	A	G
FSL-B7-2129	C	C	G	A	T	G	A	G	A	G
FSL-B7-1674	T	A	A	T	T	G	C	A	T	G
FSL-B7-2156	▨	A	G	A	C	G	C	G	T	G
FSL-B7-0720	▨	▨	G	A	C	G	A	G	A	G
FSL-B7-0718	▨	C	G	A	T	G	A	G	A	G
FSL-B7-1067	C	▨	G	A	C	G	C	G	A	G
FSL-B7-2755	▨	A	G	A	T	G	C	A	A	G

Figure 1: Isolates used for biofilm formation based on variant presence. Isolates are ordered from best biofilm former to poor biofilm former. Green shading indicates references variant TreeWAS analysis and red indicates alternative variant and grey indicates variant not present in isolate. Predicted effect on biofilm rank under each SNV. Positive predicted value indicates a predicted positive impact on biofilm formation and negative value indicates a predicted negative impact.

In addition to finding a pattern among the SNP, I tried to identify a pattern indicating epistasis occurring as the alteration of a SNP may lead to a cascade of effects to different genes. This would be shown in a similar fashion where the top biofilm formers contain a combination of variants causing an improvement in biofilm formation, and vice versa. This was not presented as well among the ten variants.

When trying to identify variants associated with biofilm formation from the DBGWAS analysis, the variants showed no clear pattern among the isolates. The top ranking biofilm former, FSL-B7-2805 contained the same variants as the 21st rank biofilm former, FSL-B7-2755, except for three hypothetical proteins. The first hypothetical protein had an expected positive effect on biofilm formation but was present in seven different isolates ranked 1st, 4th, 6th, 10th, 13th, 14th and 20th. The second hypothetical was expected to have a positive effect on biofilm formation as well and was only present in isolates ranked 1st, 6th, 10th, 13th and 14th. The third hypothetical protein was present in the 1st, 6th, 10th, 13th and 14th ranked isolates. These 3 hypothetical proteins were the only difference between the top ranked and lowest ranked biofilm former. This may indicate that the presence of these SNP may play a small role in biofilm formation but not enough to alter the phenotype. It is possible that individually, the SNPs influence on the isolate may not be profound and it may require a combination of variants to alter the biofilm phenotype if present. It also possible that the SNPs that had a positive or negative impact on the North Dakota/Minnesota isolates wont translate due to different mutations giving the same effect.

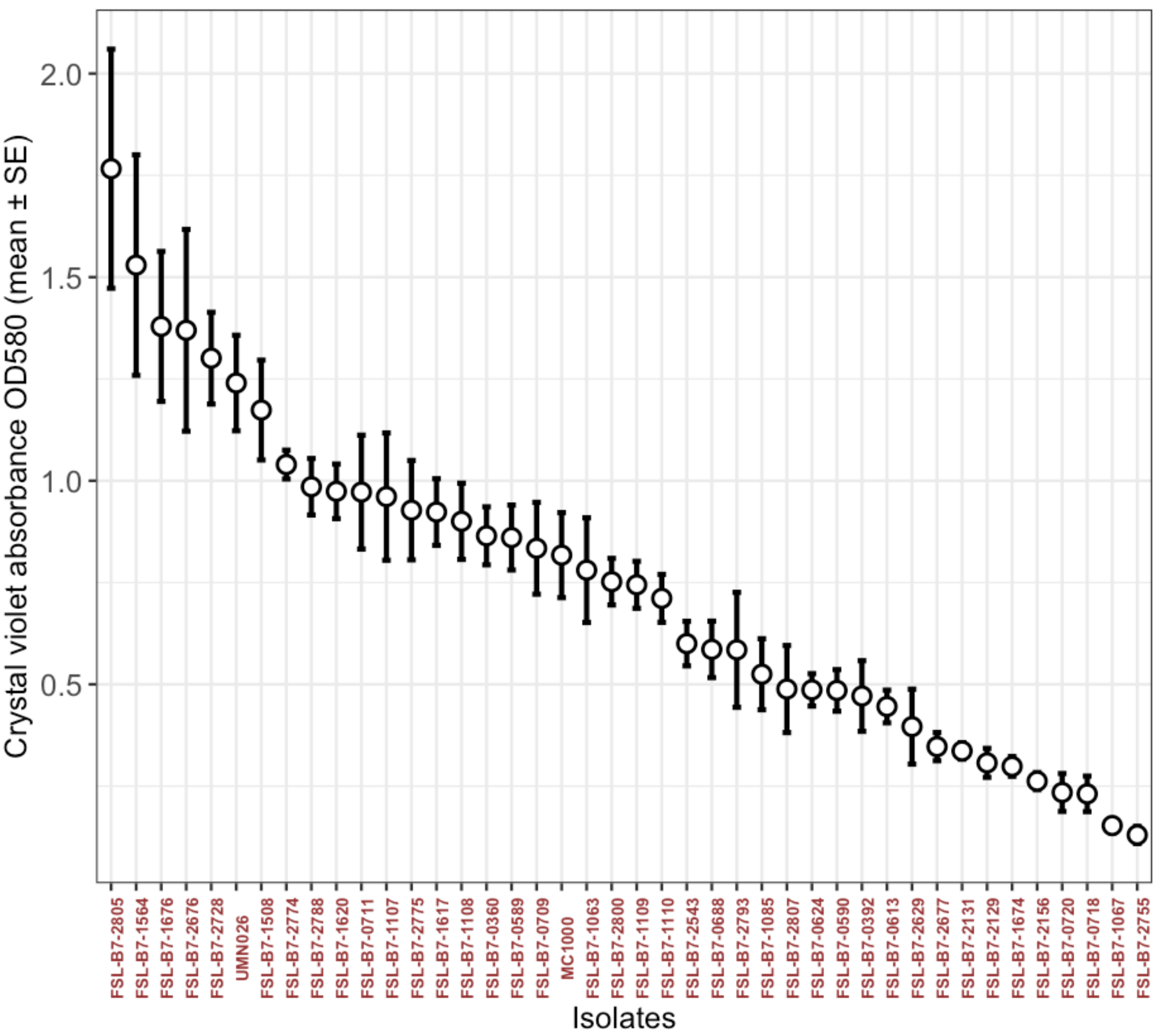
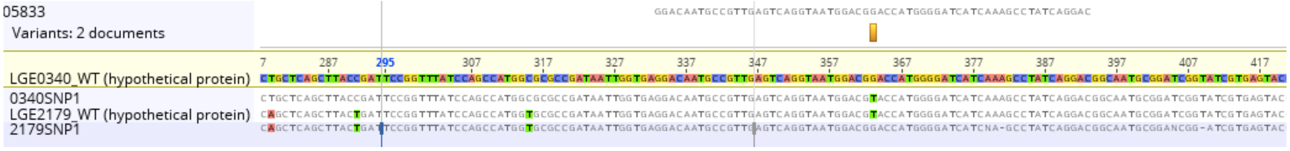


Figure 2: New York soil isolates crystal violet results, ranked from best biofilm formers to least.

Allelic exchange effect on biofilm formation and growth

To validate GWAS results, we used two isolates with different nucleotides in the same region creating a synonymous mutation. Allelic swaps were done using a CRISPR/Cas9 system, pCAGO(121) and primers for allelic exchange (table 4). The target we focused on had a positive predicted value on biofilm formation and a hypothetical protein. Results were confirmed via Sanger sequencing, with LGE0340 containing a guanine and LGE2179 containing a thymine in the wild type isolates. The target region was used as a template to create the swap, figure 3.



```
05833                                                                   GGACAATGCCGTTGAGTCAGGTAATGGACGGACCATGGGGATCATCAAAGCCTATCAGGAC
 Variants: 2 documents
                                   7    287  295   307   317   327   337   347   357   367   377   387   397   407   417
LGE0340_WT (hypothetical protein)  CTGCTCAGCTTACCGATTCCGGTTTATCCAGCCATGGCGCGCCGATAATTGGTGAGGACAATGCCGTTGAGTCAGGTAATGGACGGACCATGGGGATCATCAAAGCCTATCAGGACGGCAATGCGGATCGGTATCGTGAGTAC
0340SNP1                           CTGCTCAGCTTACCGATTCCGGTTTATCCAGCCATGGCGCGCCGATAATTGGTGAGGACAATGCCGTTGAGTCAGGTAATGGACGTACCATGGGGATCATCAAAGCCTATCAGGACGGCAATGCGGATCGGTATCGTGAGTAC
LGE2179_WT (hypothetical protein)  CAGCTCAGCTTACTGATTCCGGTTTATCCAGCCATGGTGCGCCGATAATTGGTGAGGACAATGCCGTTGAGTCAGGTAATGGACGTACCATGGGGATCATCAAAGCCTATCAGGACGGCAATGCGGATCGGTATCGTGAGTAC
2179SNP1                           CAGCTCAGCTTACTGATTCCGGTTTATCCAGCCATGGTGCGCCGATAATTGGTGAGGACAATGCCGTTGAGTCAGGTAATGGACGGACCATGGGGATCATCNA-GCCTATCAGGACGGCAATGCGGANCGG-ATCGTGAGTAC
```

Figure 3: Confirmation of allelic exchange via Sanger Sequencing. Desired point mutation is found presented at the yellow mark with no additional edits in the surrounding area. Generated on Geneious 11.1.5 (https://www.geneious.com).

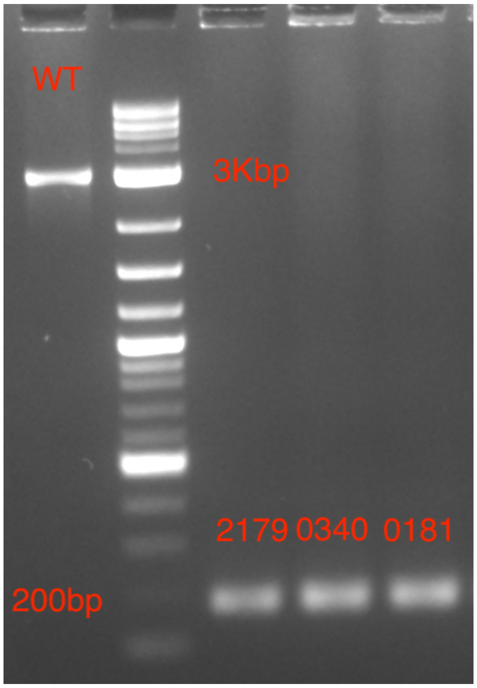


Figure 4: Deletion of hypothetical protein. The three isolates for gene deletion are LGE0181, LGE0340 and LGE2179 with amplification of WT LGE0340 as a control presented to the left of the DNA ladder.

Given the possibility that more than one single nucleotide effects the phenotype, the deletion of the hypothetical protein was completed using the CRISPR/Cas9 system (121). Primers used for genome deletion found in Table 4, with gene deletion in figure 4, with the left most band, 2,988bp, being the wild type from isolate LGE0340 and deletion leaving approximately 200bp behind which is the L-short homologous arm and right homologous arm. A two way ANOVA was completed, strain and mutation as factors, on R Version 1.2.1355. When comparing LGE0340 and SNP variant (p-value >0.05) there was no statistical significance. The

SNP exchange into LGE2179 had no statistically significant effect (p-value >0.05) on biofilm formation indicating that this variant had no statistically significance on these two isolates (table 1). There was not statistically significance between deletion mutants and wild type for biofilm formation, table 1.

Table 1: Crystal violet reading at OD580 for wild type mutants

Isolate	Average Crystal Violet OD_{580} reading	Standard Deviation	Coefficient of Variation
LGE0181	0.419	0.091	0.216
LGE0181Δn505833	0.530	0.188	0.355
LGE0340	1.141	0.232	0.203
LGE0340Δn505833	1.216	0.367	0.302
LGE0340SNP	0.933	0.089	0.096
LGE2179	0.580	0.180	0.311
LGE2179Δn505833	0.352	0.177	0.503
LGE2179SNP	0.382	0.065	0.170

Mutation effect on growth in LB broth and glucose minimal media

Due to the multitude of genes required for biofilm formation, the alteration of one gene can lead to a cascade of events that may affect growth or yield. The isolates were grown on LB broth or GDMM to identify any effects that may have occurred. A two-way ANOVA was conducted, with strain and mutant as factors. Two-way ANOVA and growth rate and yield were completed using R version 1.2.1355. Growth yields and growth rate had no statistical significance (p-value >0.05) indicating that this hypothetical protein might not play a role in growth rate or yield.

Table 2: Growth yield and growth rate of wild-type and mutants

	Growth on LB broth		Growth on GDMM	
Isolate	LOG10Nmax	μMax	LOG10Nmax	μMax
LGE0181	-0.218	1.847	-0.216	1.311
LGE0181Δ	-0.233	1.885	-0.214	1.276
LGE0340SNP	-0.273	1.985	-0.208	1.536
LGE0340Δ	-0.266	2.122	-0.282	1.481
LGE2179wt	-0.278	1.632	-0.240	0.957
LGE2179SNP	-0.301	1.787	-0.214	0.870
LGE2179Δ	-0.295	1.790	-0.265	0.937
UMN026	-0.254	2.019	-0.191	1.345
MC1000	-0.447	1.451	-0.376	1.010

Discussion

E. coli is known for being a versatile organism that can survive in various environments and occasionally lead to pathogenicity within the host(11, 14, 15). Within the GI tract, *E. coli* phylogroups A and B1 are the predominant phylogroups with phylogroup A in humans and B1 in animals (139). The ability to survive for prolonged periods outside of a host has led to the identification of environmental *E. coli* in soil which can act as a reservoir for *E. coli* and allow for commensal variations to be altered due to environmental factors leading to pathogenicity (9, 128, 140, 141). Phylogroup D are often found in soil and are one of the dominant phylogroups, pH during urinary tract infections and sepsis (15, 34, 42, 131, 139, 142). This study focused on *E. coli* phylogroup D for its genomic diversity, adaptation to numerous niches and ability to cause disease. Common virulence factors associated with extraintestinal infection such as adhesion (143, 144), motility (84, 145), and iron utilization (99, 146) were used as genetic markers containing the naturally occurring single nucleotide polymorphism that are associated with biofilm formation. This was an attempt to identify genetic patterns that may be consistent across different *E. coli* isolates. With the identification of a pattern, treatment for biofilm associated diseases can be used to identify known biofilm formers against poor biofilm formers.

In a study of non-domesticated *E. coli,* the biofilm capabilities were not dependent on strain origin rather growth medium. Biofilm formation was well distributed among pathogens and commensal *E. coli* (16). Biofilms are thought to play a crucial role is surviving in the environment, but recent report indicates the possibility of long-term persistence in soil without the ability of biofilm formation (147). The wide range of biofilm formers in the New York isolates, can indicate that our poor biofilm formers may have found a niche that doesn't require a robust biofilm rather have a different niche specific advantage that hasn't been depleted from soil dwelling isolates. On the other hand, the isolates that have the capability of producing a large amount of biomass have a niche that contributes to their survival. Therefore, the five isolates in this study that had similar or improved biofilm formation when compared to the uropathogen UMN026 may have improved survival in soil when compared to the pathogen but further investigation would be required. Additionally, it is possible that our poor biofilm forming *E. coli* may have improved biofilm formation when in multispecies biofilms. It has been observed that *E. coli* have improved biofilm formation in dual species biofilm when grown with *P. aeruginosa* a opportunist pathogen commonly found in water and soil (148).

Previous studies have used GWAS to identify genes associated with disease with great success (95, 99). At times, studies have identified an association between genotype and phenotype but may be limited by not further validating the results on a new set isolates, knockouts, or allelic exchanges. Further validation via genetically modifying the vast number of variants associated with a phenotype allows for an efficient way to study complex traits such as biofilm formation or antibiotic resistance. In a recent GWAS study on non-clinical *A. fumigatus*, one SNP was identified within *Afu2g02220*, which was knocked out and found to play a minor role in sensitivity to itraconazole (105). This identified the role of the gene in azole sensitivity, but the role of the SNP remained inconclusive as gene deletion eliminates the possibility of understanding the role that the SNP plays in non-clinical isolates. This also can aid in the identification of false positives. This limitation is not due to GWAS but rather the large number of genes and SNPs identified in GWAS. In a recent GWAS study, when cross referencing the results with previous studies linking pathogenicity, both identified *eae*, encodes intimin, and *ompT*, cleaves colocin, as associated with pathogenicity but when comparing to a different study minimal overlapping occurred (136). This coincides with my results during the validation of the GWAS generalization to a new set of isolates as new genetic variation may be present in the new set of isolates, that produce the same effect therefore the variants I used may not be as useful. In our study, the SNPs expected to influence the biofilm phenotype did not consistently correspond to the biofilm phenotype. Further computation analysis would be needed to be completed to identify any single nucleotide variant within the new set of *E .coli* and the previous GWAS data such as a random forest analysis. Improvements can be made in generalization of GWAS analysis to a new set of data, there is always the chance that the variants identified in one study won't be identified as significant in another.

A recent study was conducted to identify prediction of disease severity but led to no known genetic markers. This led to inconclusive results as Hendriks *et al*., couldn't identify a single gene or combination of genes statistically associated with symptoms or severity of disease (98). Within this same study, multiple genes couldn't be statistically associated with specific symptoms or severity and a multitude of associated k-mers that led to a 100bp sequence identified as significant was later considered a false positive upon further analysis (98). The results presented in this study demonstrate a problem inherent to bacterial GWAS as it may not always lead to positive results rather inconclusive results that need further investigation.

The difference from my research was that I focused on a SNPs present in a set of ten genes associated with biofilm formation or adaptation to nutrient depleted environments. Although the genes may be present on all isolates, the variants will not be present in all isolates. The difficulty of identifying a pattern within a different set of isolates is likely due to epistatic interactions among SNPs or genes cause a change in the phenotype. The possibility of one single nucleotide variant shifting how a single strain acts has previously been shown in the Long Term Evolution Experiment on *E. coli*. They identified nonsynonymous mutations occurring within *spoT* which increased fitness when compared against the ancestor strain and led to gene expression changes which were shown in 8 of the 12 evolved lines. This same mutation didn’t have any effect on fitness when it was introduced to a different evolved *E. coli* population, indicating that other mutations in the other population likely had the same effect as this single nucleotide polymorphism (11). It is possible that the selected set of variants to predict the biofilm phenotype shouldn’t be based on the literature rather machine learning should be implemented to identify patterns.

Previous research on *Campylobacter jejuni* has led to the verification of genetic markers based on the presence of specific accessory genes, and some of these can be used in combination to detect clinically significant isolates (149). Our results didn’t end up with results that would allow us to verify the association between genetic variants and biofilm formation, but it allows us to remove them from the list of variants that would be considered as significant independently in a different set of isolates.

A common theme among GWAS, is the identification of hypothetical proteins as having a statistically significant association with the phenotype being studied. In this research, a hypothetical protein was identified containing a conserved domain of ddrB-like ParB superfamily domain and was originally annotated as a hypothetical protein but also annotated as a D-serine transporter (e-value 6). Previous research on genes containing a p*arAB l*ike nuclease domain, were linked to the formation of mesh-like structures in *Actinomyces oris* K20 (150). In addition*, parAB* in *Pseudomonas aeruginosa*, are important for cell division and motility, so deletion of *parB* led to various changes in gene expression (150, 151). The importance of the *parB* in gene regulation, may have led to the increased coefficient of variation within our isolate that had the gene deleted but no major effects when an allelic exchange was made.

Overall, we can state that genetic variants causing increased biofilm formation in one set of *E. coli* may not be representative of a new set. The ability to predict the biofilm capabilities of a microbe based on their genome sequence improve the point-of-care for patients that are dealing with chronic infections. Additionally, CRISPR systems can be incorporated in the future to target these specific genetic variants and disrupt biofilm formation.

We were not able to identify a pattern from the ten genetic variants from the TreeWAS analysis but machine learning can be implemented to find patterns among the new set of E. coli isolates associated with biofilm formation. The deletion of the hypothetical protein didn't make a statistical significance in biofilm formation nor in growth. It did have an impact on the variability of biofilm formation indicating that the role of this gene may play some sort of role in biofilm formation, but the large network of genes required for biofilm formation may be upregulated/downregulated when the hypothetical protein was deleted. Further analysis on the mutants would be required to identify the exact effects that they are causing on the isolate overall. Moving forward, given the flexibility of pCAGO we can continue to add make allelic exchanges among similar isolates that contain statistically significant SNV. These additional edits would allow us to understand the role of a single nucleotide variant and multitude of SNV.

Material and methods

E. coli strains used in this study

To demonstrate the genomic diversity of *E. coli*, a collection of Phylogroup D from the Cornell University Food Safety Laboratory collection was used (137). Forty isolates were used to conduct the prediction study among the new set of *E. coli* isolates. This new set of *E. coli* isolates had not yet been tested for their ability to form a biofilm . DBGWAS(102) analysis was performed previously by Manoj Kumar (manuscript in preparation), and TreeWAS analysis was performed by (130). The DBGWAS variants are used for comparison on a subset of 21 isolates from the original 40 isolates. The present experiment was designed to compare the predictive value of DBGWAS and TreeWAS on a new set of isolates but occasionally variants were not present in the selected isolates. The isolates used for prediction of the biofilm phenotype are found in table 3. The uropathogen *E. coli* UMN026 was used as a positive control since it is part of phylogroup D and was used as the reference genome for TreeWAS analysis. *E. coli* K-12 MC1000 was used as a control due to it being a poor biofilm former since it contains an IS5 in the *flhD* promoter making it highly motile with reduced biofilm production (72, 152). For genome editing, pCAGO plasmid was purified using QIAprep Spin Miniprep kit and was transformed into desired cells for genome editing. Two isolates were used to produce an allelic exchange and three isolates were used for deletion of the entire gene. Isolates used for genome editing were LGE0181, LGE0340 and LGE2179.

Table 3: *E. coli* isolates for this study

Isolate	Phylogroup/Source/Characteristic
FSL-B7-2805*	Phylogroup D/New York Soil/ Extraintestinal *E. coli*
FSL-B7-1564*	Phylogroup D/New York Soil/ Extraintestinal *E. coli*
FSL-B7-1676*	Phylogroup D/New York Soil/ Extraintestinal *E. coli*
FSL-B7-2676*	Phylogroup D/New York Soil/ Extraintestinal *E. coli*
FSL-B7-2728*	Phylogroup D/New York Soil/ Extraintestinal *E. coli*
UMN026	Phylogroup D/Uncomplicated acute cystitis/Good biofilm former
FSL-B7-1508	Phylogroup D/New York Soil/ Extraintestinal *E. coli*
FSL-B7-2774*	Phylogroup D/New York Soil/ Extraintestinal *E. coli*
FSL-B7-2788*	Phylogroup D/New York Soil/ Extraintestinal *E. coli*
FSL-B7-1620*	Phylogroup D/New York Soil/ Extraintestinal *E. coli*
FSL-B7-0711	Phylogroup D/New York Soil/ Extraintestinal *E. coli*
FSL-B7-1107	Phylogroup D/New York Soil/ Extraintestinal *E. coli*
FSL-B7-2775*	Phylogroup D/New York Soil/ Extraintestinal *E. coli*
FSL-B7-1617*	Phylogroup D/New York Soil/ Extraintestinal *E. coli*
FSL-B7-1108	Phylogroup D/New York Soil/ Extraintestinal *E. coli*
FSL-B7-0360	Phylogroup D/New York Soil/ Extraintestinal *E. coli*
FSL-B7-0589	Phylogroup D/New York Soil/ Extraintestinal *E. coli*
FSL-B7-0709	Phylogroup D/New York Soil/ Extraintestinal *E. coli*
K-12 MC1000	Phylogroup A/Lab strain/Poor Biofilm former
FSL-B7-1063	Phylogroup D/New York Soil/ Extraintestinal *E. coli*
FSL-B7-2800*	Phylogroup D/New York Soil/ Extraintestinal *E. coli*
FSL-B7-1109	Phylogroup D/New York Soil/ Extraintestinal *E. coli*
FSL-B7-1110	Phylogroup D/New York Soil/ Extraintestinal *E. coli*
FSL-B7-2543*	Phylogroup D/New York Soil/ Extraintestinal *E. coli*
FSL-B7-0688	Phylogroup D/New York Soil/ Extraintestinal *E. coli*
FSL-B7-2793*	Phylogroup D/New York Soil/ Extraintestinal *E. coli*
FSL-B7-1085	Phylogroup D/New York Soil/ Extraintestinal *E. coli*
FSL-B7-2807*	Phylogroup D/New York Soil/ Extraintestinal *E. coli*
FSL-B7-0624	Phylogroup D/New York Soil/ Extraintestinal *E. coli*
FSL-B7-0590	Phylogroup D/New York Soil/ Extraintestinal *E. coli*
FSL-B7-0392	Phylogroup D/New York Soil/ Extraintestinal *E. coli*
FSL-B7-0613	Phylogroup D/New York Soil/ Extraintestinal *E. coli*
FSL-B7-2629*	Phylogroup D/New York Soil/ Extraintestinal *E. coli*
FSL-B7-2677*	Phylogroup D/New York Soil/ Extraintestinal *E. coli*

Table 3: *E. coli* isolates for this study (Continued)

Isolate	Phylogroup/Source/Characteristic
FSL-B7-2131*	Phylogroup D/New York Soil/ Extraintestinal *E. coli*
FSL-B7-2129*	Phylogroup D/New York Soil/ Extraintestinal *E. coli*
FSL-B7-1674*	Phylogroup D/New York Soil/ Extraintestinal *E. coli*
FSL-B7-2156*	Phylogroup D/New York Soil/ Extraintestinal *E. coli*
FSL-B7-0720	Phylogroup D/New York Soil/ Extraintestinal *E. coli*
FSL-B7-0718	Phylogroup D/New York Soil/ Extraintestinal *E. coli*
FSL-B7-1067	Phylogroup D/New York Soil/ Extraintestinal *E. coli*
FSL-B7-2755*	Phylogroup D/New York Soil/ Extraintestinal *E. coli*
LGE0181	Phylogroup D/Minnesota and North Dakota soil/Extraintestinal *E. coli*
LGE0181Δn505833	This study
LGE0340	Phylogroup D/Minnesota and North Dakota soil/Extraintestinal *E. coli*
LGE0340Δn505833	This study
LGE0340_SNP	This study
LGE2179	Phylogroup D/Minnesota and North Dakota soil/Extraintestinal *E. coli*
LGE2179Δn505833	This study
LGE2179_SNP	This study
E.coli K-12 MG1655	Phylogroup A/Lab Strain/pCAGO_Ampicillin resistance_CRISPR/Lambda Red system
E.coli DH5-alpha	Phylogroup A/ Lab strain/pBMT-3_Chloramphenicol resistance

Note: *E. coli* K-12 strain MG1655 was used to store pCAGO(121) plasmid. *E.coli* DH5-alpha pBMT-3 was a gift from Ryan Gill (Addgene plasmid # 22838 ; http://n2t.net/addgene:22838 ; RRID:Addgene_22838). Isolates containing "*" were used for prediction study using DBGWAS analysis. All FSL-B7-#### were used for prediction study using TreeWAS analysis. Isolates labeled with a "*" after the identification number were used for prediction analysis of DBGWAS. Isolates labeled with LGE#### were used for genome editing.

Prediction of biofilm phenotype based on GWAS analysis

To assess the predictive value of the GWAS results, a set of *E. coli* isolates from New York soil (Figure 2) were selected to validate the two GWAS tools used: TreeWAS and DBGWAS. TreeWAS analysis was completed by Morgan Petersen (130) and DBGWAS by

Manoj Kumar (manuscript in preparation), TreeWAS was used as the main prediction model due to its use of a reference genome, which identified 1,089 single nucleotide variants (130). We selected ten variants based on current literature and score from the analysis, *kpsS, dsdA, pduT, iucD, yfaL, entF, fimA, evgS, dcp,* and *flgG* (130). TreeWAS analysis found association between variants and biofilm formation which were found within these ten genes. The ability to use universal genetic markers to identify biofilm formation would allow for improved diagnostics. Using the predictive values from TreeWAS, we wanted to validate if we can use it as a prediction model on a new set of *E. coli*.

Genes associated with adaptative advantages were selected due to the importance in surviving within a host or in nutrient limited environments. D-serine utilization (*dsdA*) has shown to have a growth advantage in uropathogenic *E. coli* within the urinary tract and dsdA mutants had a prolonged lag phase (15, 35, 153). In a study to identify the importance of the *pdu* operon during anaerobic growth in *S. enterica*, the genes within the *pdu* operon are required to form polyhedral bodies involved with the degradation of 1,2- propanediol so *ECUMN_2350/PduT* was selected (154). Iron acquisition systems association with virulence has been determined previously as *iucD* mutants led to reduced colonization in internal organs within chickens (155). In *E. coli* with a reduced genome, enterobactin was required for biofilm development (156), and *entF* mutants showed increased sensitivity to oxidative stress response required for colony development (157). The *evgS,* sensor kinase*,* senses external and internal pH, so mutations within this gene can lead to a slow/quicker response in low pH and alkali conditions (158). *KpsS* encodes for capsule polysaccharide export, which are essential for evasion of host neutrophil attack and polysaccharide capsule promote intracellular biofilm structure (153, 159, 160). The ability to adhere or swim to a surface is critical for biofilm formation, but overproduction of such genes may lead to detrimental effects as well. Over expression of *flgG*, flagellar basal body rod protein, leads to no flagella production which would make the cells immobilized thus not allowing for colonization of new locations (161). Two different adhesion genes were selected *fimA* (major type 1 fimbrial subunit) being the first as previous a deletion study showed delay biofilm formation (162). The second adhesion gene selected *yfaL*, was selected as expression led to increased adhesion in *E. coli* K-12 MG1655 (163). Lastly, dipeptidyl carboxypeptidase II (*dcp)* a C-terminal zinc-dependent metallopeptidase, was selected as matrix metalloproteinases disruption can lead to inhibition of microbial cell growth and effect homeostasis (164–166).

DBGWAS analysis was used to identify potential trends in biofilm-affecting SNPs and for genome editing. DBGWAS doesn't use a reference genome, so it can identify novel variants that may not be present within the reference genome (102). A total of 56 variants were identified within the 21 isolates, with no presence being found in 19 of the 40. The variants identified as negative effect on biofilm formation were nine pangenome products of CPS-53 (KpLE1) prophage, six e14 prophage, two exonucleases VII, 37 hypothetical proteins Manoj Kumar (manuscript in preparation).

Genome editing was conducted on node 505833, a variably present k-mer within a gene that was annotated as a hypothetical protein on. It was additionally annotated as a D-serine transporter (e-value 6). The hypothetical protein contained a conserved domain of ddrB-like ParB superfamily domain (e-value =1.62e-47). Previous research on genes containing a *ParB* like nuclease domain, were linked to the formation of mesh-like structures in *Actinomyces oris* K20 (150). In addition*, parAB* in *Pseudomonas aeruginosa*, are important for cell division and motility, so deletion of *parB* led to various changes in gene expression (150, 151). The identification of this target region was focused on the possibility as a D-serine transporter, but conserved domain may have a small role in the proteins structure which led to this selection in addition to the positive estimated effect on biofilm formation Manoj Kumar (manuscript in preparation).

Preparation of electrocompetent cells for pCAGO insertion or edit cassette insertion

Electrocompetent cells were prepared following the protocol rom Datsenk and Wanner, with some modification to adjust for volume needed (109). When cells harbored the pCAGO plasmid a final concentration of 100ug/mL ampicillin was added to the media. If cells contain the pCAGO plasmid, a 5% inoculation was done instead of a 1% inoculum, with a final concentration of 100ug/mL of ampicillin was added to the media and induction of lambda Red recombination proteins was done from the start of this incubation with a final concentration of 0.1mM of Isopropyl β- d-1-thiogalactopyranoside (IPTG) (121). From this step forward, the cells were kept on ice as much as possible as centrifugation will make the cells very fragile and lead to cell lysis if not kept chilled. After the 30 minutes, 50mL of the cells were aliquoted into a 50mL conical tube as this made it convenient for the lab. All the centrifugation steps were completed on a Allegra X-22R benchtop centrifuge, Rotor SX4250 Swinging Bucket was used, at 3,901 RCF x G for 15 minutes at 4˚C. The cells can be used for electroporation immediately or placed in the -80˚C freezer (109, 121).

Electroporation of pCAGO into desired cells

Electroporation was completed using a Biorad MicroPulser, with setting following Ec1 protocol which indicated 1.8kV for 5.8ms with a 1mm gapped electroporation cuvette. Before electroporation of pCAGO plasmid (121),1 liter of Super Optimal Broth was made or Sigma Aldrich SOB powder was used. Super Optimal broth with Catabolite repression (SOC) was necessary for the recovery stage of transformation, so 20mM glucose was added. The SOC media was prepared the day it was going to be used, and chilled for at least an hour in a ice bath to keep cool with electroporation cuvettes in the ice bath to chill for at least an hour. After the SOC and cuvettes were chilled, the cells were thawed out for 5-10 minutes in an ice bath. When transforming the pCAGO plasmid into cells, at least 50ng was required for efficient transformation which usually was about 2ul of plasmid DNA. The 50 ng of plasmid were resuspended into the 50 µl thawed cells and 52 µl were transferred into the center of the chilled electroporation cuvette. Immediately after electroporation 1 mL of chilled SOC media is added directly into the electroporation cuvette to help with recovery. The cells are resuspended gently in the SOC media and 1mL of this culture is transferred into a glass culture tube, incubated at 30˚C at 230rpm for 2 hours. After the 2 hour recovery of the cells, 200ul of culture are spread onto five different LB agar plates containing 100ug/mL of ampicillin as electroporation has a high cell death rate with best transformation occurring when 40-50% of cells survive the pulse (167). The cells are grown overnight in the LB + ampicillin agar plates and isolated colony is grown in LB broth containing 15% glycerol plus 100ug/mL of ampicillin to retain the cells under antibiotic selection. (109, 121, 167, 168)

Constructing editing cassette for genome editing

All genome editing was done on isolates LGE0181, LGE0340 and LGE2179 which were originally isolated from soil across North Dakota and Minnesota (41). The genomes of these three isolates have previously been sequenced and we selected them based on unpublished work by Manoj Kumar, in which a statistically significant variant was identified. This variant is found in node 505833 which is a hypothetical protein but an e-value of 6 for D-serine transporter. Primers were designed on Geneious using Primer3 function (169). All primers had a similar melting temperature, so all Polymerase Chain Reaction (PCR) were completed with the following conditions: 1 cycle at 95˚C for 5 minutes, 30 cycles for the PCR reaction 95˚C for 45 seconds, 55.8˚C for 1 minute and 30 seconds and 72˚C for 1 min and lastly a 5 minute extension at 72˚C.

The editing cassette was constructed using golden gate cloning to have a four part construct similar to figure 5a (121, 170). The editing cassette consisted of three homologous regions during the allelic exchange which are present on, figure 5b for reference to regions amplified.

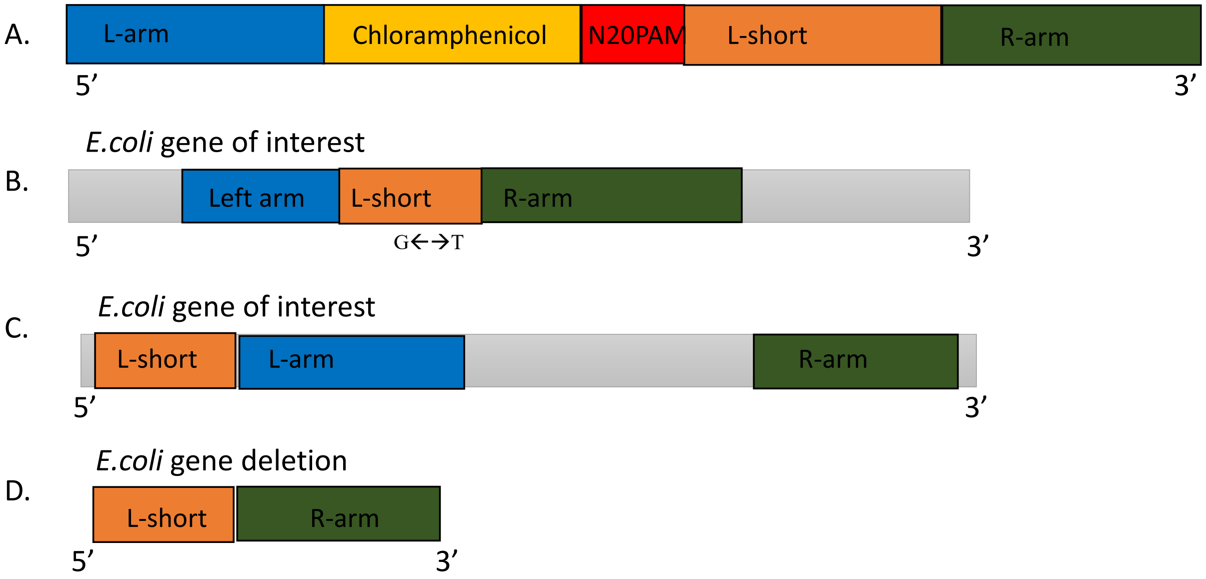


Figure 5: Editing cassette constructs. a: Editing cassette after golden gate cloning. Figure 5b. Homologous arms used for amplification withing desired target region for allelic swap with grey regions indicating the rest of the gene. The 3' end of L-short is also the end of the L-arm with R-arm beginning immediately afterwards. The only region not homologous is L-short which contains the desired mutation to be swapped. Figure 5c is homologous arms design for editing cassette when genomic deletion is desired, with the L-short flanking the L-arm. Figure 5d. Final product from genome deletion. Modified from source (121).

A left homologous arm (260bp) contained the desired target towards the 3' region, a small homologous arm, L-short (76bp) at the 3' end of the left homologous which contained the desired target region and a right homologous arm (269bp), immediately flanking the end of the

3’ end of the left homologous arm. Each module contained a recognition region ‘GAAGAC’ for the type IIS restriction enzyme BbsI (171). The fourth modular part consisted of the antibiotic selection marker, chloramphenicol(873bp). The selection marker was amplified from the pBMT-3 plasmid. The pBMT-3 plasmid was a gift from Ryan Gill (Addgene plasmid # 22838(172). The reverse primer for the chloramphenicol cassette contained the universal N20PAM sequence 5′-TAGTCCATCGAACCGAAGTA-3′ in addition to the BbsI extensions (121). The final construct, figure 5a, was designed to increase intrachromosomal recombination once Cas9 was induced. All modules were amplified via PCR from the original isolate to keep 100 % homology with the exception of the L-short containing the desired single nucleotide variant that will be swapped, figure 5b indicating that our desired target region contains the desired variant .When the editing cassette is constructed for gene deletion Fig 5c, the L-short and R-arm were designed to not overlap with neighboring genes to ensure no accidental frameshifts were created. With the design, Fig 5c, our final product was truncated to about 200bp as seen on Fig 5d. All PCR products were purified using the AccuPrep® PCR/Gel Purification Kit and followed the protocol with DNA concentrations checked with NanoDrop and visualized on 2.5 % agarose gel (168). Following a similar protocol from the original pCAGO paper, in a 15 µl reaction, 100ng of chloramphenicol PCR product was added which is the equivalent of 0.174 pmol of DNA. Each modular part consisted of 0.174 pmol, in addition to 5 units of BbsI, 20units of T4 ligase, 1.5 µl of ligase buffer containing ATP, and 1.5 µl of cut smart buffer and the remaining amount was H_2O for a final volume of 15 µl (121, 170, 171). The golden gate reaction was extended to 60 cycles at 37˚C and 16˚C for 5 minutes intervals as shorter time periods (2-4 minutes) led to incomplete digestion/ligation. A final enzyme inactivation step was conducted at 65˚C for 20 minutes. Golden gate products were then PCR amplified in two separate 50 µl reactions to increase DNA concentration with the left homologous arm forward primer and right homologous arm reverse primer used for PCR. Products were visualized on 2.5 % agar to verify desired product size, with the remaining PCR products pooled, purified using the AccuPrep PCR/Gel purification kit and eluted into a 1.5mL microcentrifuge tube with 100 µL of purified H_2O (168). The 1.5mL microcentrifuge tube was placed into the SpeedVac at 45˚C for approximately 20 minutes to increase the concentration of the DNA. The DNA editing cassette was lastly checked on the NanoDrop to verify DNA concentration, and purity of samples (121, 167–172).

Table 4: Primers used for allelic swap between LGE0340 and LGE2179

Name	Sequence
Cmr_F	CACCACAGAAGACGACGCATGTTTTCTGGACGATGGCGA
L-arm_F	CGATGCCTCAGCTCTTTTGG
R-arm_F	CACCACAGAAGACGACGGCGTCTTCCGACCAACACGGCAACTTG
L-short_F	CACCACAGAAGACGAACTAATTGGTGAGGACAATGCCGT
CmrR_N20	CACCACAGAAGACGATAGTCCATCGAACCGAAGTAAGGCTTTTGACTTGAGGGGG
L-arm_R	CACCACAGAAGACGATGCGGCTAAGCAGATCTCC
R-arm_R	AAAAGACCGCCGGATATCCC
L-short_R	CACCACAGAAGACGAGCCGTCCTGATAGGCTTTGA

Note: Forward primers consist of a F at the end of the name and reverse primers include an R at the end of the primer. Product size: Left arm= 260bp, Chloramphenicol insert with N20PAM = 873bp, L-short(TR)= 76bp, Right arm= 269bp

Table 5: Primers used for deletion of n505833

Name	Sequence
R-arm_F	CACCACAGAAGACGATCGAGACTATTTCCAAAACGTCAGCA
L-short_F	CACCACAGAAGACGAACTATCGTATGTGGTGACAGCGAA
L-arm_R	CACCACAGAAGACGATGCGCCTGAAGAAGGCTGGATGCG
L-short_R	CACCACAGAAGACGATCGACGATATCAAACGCTGT
L-arm_F	TGCCTCAGCTCTTTTGGTATCC
R-arm_R	GTTCCAGCAATCCGTCACCA

Note: Primers used for deletion Forward primers are marked with an F at the end of the name and reverse primers with an R at the end of the name. Antibiotic marker primers were the same as the allelic swap as the overhang left were used as a template for the deletion. Product size: Left arm= 107bp, L-short= 60bp, R-arm= 114bp

Transformation of editing cassette into desired isolate

Electrocompetent cells expressing lambda Red proteins were thawed for 5-10 minutes in an ice bath and electroporation procedure was followed as stated above. A concentration of approximately 450 ng-500ng of double stranded(ds) DNA or 0.54 pmol was resuspended in *E. coli* electrocompetent cells expressing lambda Red proteins and electroporated. The cells recovered in ice cold SOC media for two hours at 30˚C with 230 rpm. After the incubation period, 200 µl of culture were spread onto different LB agar plates containing 30ug/mL chloramphenicol, 100ug/mL ampicillin, and 1% glucose. Plates were then incubated at 30˚C for 16-20 hours. Colony formation were often too small, so colonies were propagated onto a new LB plate containing ampicillin, chloramphenicol and 1% glucose and incubated at 30˚C for 12-16 hours.

Colony PCR verification of correct insert orientation

Colony PCR was performed with the forward primer binding to the chloramphenicol cassette (CmrF) and reverse primer binding to the 3' end of the right homologous arm with PCR. Following PCR amplification, products were visualized on 2.5 % agarose gel. PCR conditions were as followed: Initial denature step at 95˚C for 5 minutes, 30 cycles at 95˚C for 45 seconds, 55.8˚C for 2 minutes 72˚C for 1 min and a final extension step for five minutes at 72˚C.

CRISPR/Lambda Red induction

Once an isolate with the desired insert was identified, the cells are placed in conditions to promote intrachromosomal recombination. The positive colony is inoculated onto 10mL of LB media containing 100ug/mL of ampicillin and incubated at 30˚C with 230rpm. Culture was incubated until the cells reached an OD600 reading of 0.2-0.4. Once the cells reached the desired optical density, a final concentration of 0.1mM (IPTG) to induce lambda Red proteins and a final concentration of 0.2% arabinose to induce Cas9 expression (121). Cells were thoroughly resuspended before incubation for at least 8 hours and no longer then 10 hours. After the incubation period, cells were diluted to a final concentration of 10^{-4},10^{-5}, 10^{-6} in 1x Phosphate Buffered Saline(PBS). The cells were spread onto plates containing LB agar and 100ug/mL ampicillin where positive mutants should grow after editing cassette removal. Additionally, diluted cells were also spread onto LB agar containing 100ug/mL ampicillin, 30ug/mL chloramphenicol and 1 % glucose to verify cassette has been removed. Cells were then incubated overnight at 30˚C. When cassette removal wasn't successful, >50 colonies were found on plates containing both antibiotics so a new colony from the previous step was selected to induce the following day. If cassette removal was successful, 0 colonies should be found on plates containing both antibiotics but if there was a presence of 1-20 abnormally small *E. coli* colonies after overnight incubation, additional colonies were screened to verify absence of chloramphenicol cassette insertion. Colony PCR was completed to verify removal using the L-arm forward primer and R-arm reverse primer. The pCAGO plasmid contains a temperature sensitive replicon which allows for removal of the plasmid when grown at 42˚C, which removed the plasmid makes the isolate sensitive to ampicillin once again (121). Upon verification of removal of cassette, a positive colony was grown on LB media containing 15% glycerol and incubated overnight at 42˚C for pCAGO plasmid curing. If pCAGO plasmid was retained, cells were grown on LB media containing 15 % glycerol and 100µg/mL ampicillin and can be used for additional editing. Upon curing of pCAGO plasmid, PCR amplification of target region was conducted to verify allelic swap via Sanger sequencing at MCLabs in San Francisco, California. Sequencing results were checked for quality and desired allelic exchange. If deletion was completed, gel electrophoresis was conducted to verify deletion of desired target. Confirmed mutants were stored in -80˚C (121).

Crystal violet assay for measuring biofilm formation

To analyze the capabilities of an isolates form biofilm formation, the crystal violet assay was conducted. Glucose defined minimal media (GDMM) containing 0.1 % glucose from Teknova (173) was used and supplemented with 0.5% casamino acids(CAA) to produce visible biofilms. A study showed that without the addition of CAA to the media, there was no visibly stained biofilm formation (66). This was further verified by Morgan Petersen in her thesis work in the P. Bergholz lab (130) so the same protocol will be used. The crystal violet assay was conducted six times to verify consistency as it is known to error prone (174).

Isolates used were stored in the -80˚C freezer, so they were plated on LB agar the day prior to the start of the experiment. All incubation steps were done at 37˚C. In the morning of day 1, isolates were inoculated into 200 µL of LB broth into the designated well in a 96-well plate. Culture were grown to stationary phase for 8 hours or 16 hours(overnight). Once the cells reached stationary phase, 2 µL of culture were transferred into 198 µL of GDMM + 0.5 % CAA and incubated overnight. The cells were then transferred into fresh GDMM + 0.5 % CAA on day 2 in the morning and evening, and on the third day the cultures were incubated for 48 hours to allow for biofilm maturation. The various transfers from the initial nutrient rich media to nutrient limited media ensures that there is no residual LB broth to be metabolized during the final 48 hours and cells are acclimated to the GDMM, so they aren't shocked into a starvation state. This allows for cultures to acclimate to the limited nutrients. Additionally, the various transfers eliminate isolates that are not able to make it past the third or fourth transfer, thus making them not ideal for studying biofilm formation (77, 130).

On day 5 after 48 hours of incubation, the cultures have formed their biofilm and can be stained. Excess media is removed from the plate using a multi-channel pipette. The wells are carefully washed three times with 200 µL of 1X PBS to avoid disrupting biofilm. The three washes ensure the removal of non-surface attached cells and any residual media to allow for proper staining. The wells were air dried for an hour to ensure proper drying. Once the wells were completely dried, the cells were stained with 200 µL of 0.1% crystal violet and incubated for 15 minutes at room temperature. The 200 µL of crystal violet solution is removed from wells and three washes with 200 µL of 1X PBS to removal excess crystal violet is done. The wells are air dried for an hour. Once the stained cells have been air dried, a 200 µL of fresh 80:20 ethanol solution is aliquoted into the wells and incubated for 15 minutes. 150 µL of the solution is transferred to a new 96-well plate and optical density at 580 nm was measured using the BioTek® Synergy H1 Hybrid Reader spectrophotometer. Empty wells were used as blanks with 150 µL of pure 80:20 ethanol: acetone solution used (130, 145, 175).

REFERENCES

1. Ishii S, Ksoll WB, Hicks RE, Sadowsky MJ. 2006. Presence and Growth of Naturalized &lt;em&gt;Escherichia coli&lt;/em&gt; in Temperate Soils from Lake Superior Watersheds. Applied and Environmental Microbiology 72:612 LP – 621.

2. Ishii S, Sadowsky MJ. 2008. Escherichia coli in the environment: Implications for water quality and human health. Microbes and Environments.

3. Duriez P, Clermont O, Bonacorsi S, Bingen E, Chaventré A, Elion J, Picard B, Denamur E. 2001. Commensal Escherichia coli isolates are phylogenetically distributed among geographically distinct human populations. Microbiology 147:1671–1676.

4. Moreira S, Brown A, Ha R, Iserhoff K, Yim M, Yang J, Liao B, Pszczolko E, Qin W, Leung KT. 2012. Persistence of Escherichia coli in freshwater periphyton: Biofilm-forming capacity as a selective advantage. FEMS Microbiology Ecology 79:608–618.

5. Pachepsky YA, Shelton DR. 2011. Escherichia coli and fecal coliforms in freshwater and estuarine sediments. Critical Reviews in Environmental Science and Technology 41:1067–1110.

6. Brennan FP, O'Flaherty V, Kramers G, Grant J, Richards KG. 2010. Long-term persistence and leaching of escherichia coli in temperate maritime soils. Applied and Environmental Microbiology 76:1449–1455.

7. Lyautey E, Lu Z, Lapen DR, Wilkes G, Scott A, Berkers T, Edge TA, Topp E. 2010. Distribution and diversity of escherichia coli populations in the South Nation River Drainage basin, eastern Ontario Canada. Applied and Environmental Microbiology 76:1486–1496.

8. Habteselassie MY, Bischoff M, Applegate B, Reuhs B, Turco RF. 2010. Understanding the role of agricultural practices in the potential colonization and contamination by Escherichia coli in the rhizospheres of fresh produce. Journal of Food Protection 73:2001–2009.

9. Touchon M, Perrin A, de Sousa JAM, Vangchhia B, Burn S, O'Brien CL, Denamur E, Gordon D, Rocha EPC. 2020. Phylogenetic background and habitat drive the genetic diversification of Escherichia coli. PLoS Genetics https://doi.org/10.1371/journal.pgen.1008866.

10. Goldstone RJ, Popat R, Schuberth H-J, Sandra O, Sheldon I, Smith DG. 2014. Genomic characterisation of an endometrial pathogenic Escherichia coli strain reveals the acquisition of genetic elements associated with extra-intestinal pathogenicity. BMC Genomics 15:1075.

11. Brzuszkiewicz E, Gottschalk G, Ron E, Hacker J, Dobrindt U. 2009. Adaptation of pathogenic E. coli to various niches: Genome flexibility is the key. Genome Dynamics. S. Karger AG.

12. Vidovic S, Korber DR. 2016. Escherichia coli O157: Insights into the adaptive stress physiology and the influence of stressors on epidemiology and ecology of this human pathogen. Critical Reviews in Microbiology 42:83–93.

13. Lennon JT, Aanderud ZT, Lehmkuhl BK, Schoolmaster DR. 2012. Mapping the niche space of soil microorganisms using taxonomy and traits. Ecology https://doi.org/10.1890/11-1745.1.

14. Ingle DJ, Clermont O, Skurnik D, Denamur E, Walk ST, Gordon DM. 2011. Biofilm formation by and thermal niche and virulence characteristics of Escherichia spp. Applied and Environmental Microbiology 77:2695–2700.

15. Connolly JPR, Goldstone RJ, Burgess K, Cogdell RJ, Beatson SA, Vollmer W, Smith DGE, Roe AJ. 2015. The host metabolite D-serine contributes to bacterial niche specificity through gene selection. The ISME Journal 9:1039–1051.

16. Reisner A, Krogfelt KA, Klein BM, Zechner EL, Molin S. 2006. In vitro biofilm formation of commensal and pathogenic Escherichia coli strains: Impact of environmental and genetic factors. Journal of Bacteriology 188:3572–3581.

17. Lai YT, Yeung CKL, Omland KE, Pang EL, Hao Y, Liao BY, Cao HF, Zhang BW, Yeh CF, Hung CM, Hung HY, Yang MY, Liang W, Hsu YC, Yao C te, Dong L, Lin K, Li SH. 2019. Standing genetic variation as the predominant source for adaptation of a songbird. Proceedings of the National Academy of Sciences of the United States of America 116:2152–2157.

18. Swings T, van den Bergh B, Wuyts S, Oeyen E, Voordeckers K, Verstrepen KJ, Fauvart M, Verstraeten N, Michiels J. 2017. Adaptive tuning of mutation rates allows fast response to lethal stress in escherichia coli. eLife 6.

19. Cosson P, Lima WC. 2014. Intracellular killing of bacteria: Is Dictyostelium a model macrophage or an alien? Cellular Microbiology. Blackwell Publishing Ltd.

20. Proença JT, Barral DC, Gordo I. 2017. Commensal-to-pathogen transition: One-single transposon insertion results in two pathoadaptive traits in Escherichia coli-macrophage interaction. Scientific Reports 7:1–12.

21. Sokurenko E v., Chesnokova V, Dykhuizen DE, Ofek I, Wu XR, Krogfelt KA, Struve G, Schembri MA, Hasty DL. 1998. Pathogenic adaptation of Escherichia coli by natural variation of the FimH adhesin. Proceedings of the National Academy of Sciences of the United States of America 95:8922–8926.

22. Wirth T, Falush D, Lan R, Colles F, Mensa P, Wieler LH, Karch H, Reeves PR, Maiden MCJ, Ochman H, Achtman M. 2006. Sex and virulence in Escherichia coli: An evolutionary perspective. Molecular Microbiology 60:1136–1151.

23. Rocha EPC. 2018. Neutral theory, microbial practice: Challenges in bacterial population genetics. Molecular Biology and Evolution 35:1338–1347.

24. Jensen JD, Payseur BA, Stephan W, Aquadro CF, Lynch M, Charlesworth D, Charlesworth B. 2019. The importance of the Neutral Theory in 1968 and 50 years on: A response to Kern and Hahn 2018. Evolution. Society for the Study of Evolution.

25. Eyre-Walker A, Keightley PD. 2007. The distribution of fitness effects of new mutations. Nature Reviews Genetics.

26. The Neutral Theory of Molecular Evolution - Motoo Kimura - Google Books.

27. Weissman SJ, Beskhlebnaya V, Chesnokova V, Chattopadhyay S, Stamm WE, Hooton TM, Sokurenko E v. 2007. Differential stability and trade-off effects of pathoadaptive mutations in the Escherichia coli FimH adhesin. Infection and Immunity 75:3548–3555.

28. Bonacorsi S, Bingen E. 2005. Molecular epidemiology of Escherichia coli causing neonatal meningitis. International Journal of Medical Microbiology. Elsevier GmbH.

29. Johnson JR. 1991. Virulence factors in Escherichia coli urinary tract infection. Clinical Microbiology Reviews 4:80–128.

30. Delcaru C, Alexandru I, Podgoreanu P, Grosu M, Stavropoulos E, Chifiriuc MC, Lazar V. 2016. Microbial biofilms in urinary tract infections and prostatitis: Etiology, pathogenicity, and combating strategies. Pathogens. MDPI AG.

31. Kumon H, Hashimoto H, Nishimura M, Monden K, Ono N. 2001. Catheter-associated urinary tract infections: Impact of catheter materials on their management, p. 311–316. *In* International Journal of Antimicrobial Agents. Elsevier.

32. Yaron S, Römling U. 2014. Biofilm formation by enteric pathogens and its role in plant colonization and persistence. Microbial Biotechnology. John Wiley and Sons Ltd.

33. Johnson JR, Kuskowski MA, O'Bryan TT, Colodner R, Raz R. 2005. Virulence genotype and phylogenetic origin in relation to antibiotic resistance profile among Escherichia coli urine sample isolates from Israeli women with acute uncomplicated cystitis. Antimicrobial Agents and Chemotherapy. American Society for Microbiology Journals.

34. Connolly JPR, Gabrielsen M, Goldstone RJ, Grinter R, Wang D, Cogdell RJ, Walker D, Smith DGE, Roe AJ. 2016. A Highly Conserved Bacterial D-Serine Uptake System Links Host Metabolism and Virulence. PLoS Pathogens 12:e1005359.

35. Roesch PL, Redford P, Batchelet S, Moritz RL, Pellett S, Haugen BJ, Blattner FR, Welch RA. 2003. Uropathogenic Escherichia coli use D-serine deaminase to modulate infection of the murine urinary tract. Molecular Microbiology 49:55–67.

36. Goldstone RJ, Popat R, Schuberth H-JJ, Sandra O, Sheldon IM, Smith DGEG. 2014. Genomic characterisation of an endometrial pathogenic Escherichia coli strain reveals the acquisition of genetic elements associated with extra-intestinal pathogenicity. BMC Genomics 15:1075.

37. Welch RA, Burland V, Plunkett G, III, Redford P, Roesch P, Rasko D, Buckles EL, Liou S-R, Boutin A, Hackett J, Stroud D, Mayhew GF, Rose DJ, Zhou S, Schwartz DC, Perna NT, Mobley HLT, Donnenberg MS, Blattner FR. 2002. Extensive mosaic structure revealed by the complete genome sequence of uropathogenic Escherichia coli. Proceedings of the National Academy of Sciences of the United States of America 99:17020.

38. Bohlin J, Brynildsrud OB, Sekse C, Snipen L. 2014. An evolutionary analysis of genome expansion and pathogenicity in Escherichia coli. BMC Genomics 15:1–13.

39. Rasko DA, Rosovitz MJ, Myers GSA, Mongodin EF, Fricke WF, Gajer P, Crabtree J, Sebaihia M, Thomson NR, Chaudhuri R, Henderson IR, Sperandio V, Ravel J. 2008. The pangenome structure of Escherichia coli: comparative genomic analysis of E. coli commensal and pathogenic isolates. Journal of bacteriology 190:6881–6893.

40. Abe K, Nomura N, Suzuki S. 2021. Biofilms: Hot spots of horizontal gene transfer (HGT) in aquatic environments, with a focus on a new HGT mechanism. FEMS Microbiology Ecology. Oxford University Press.

41. Brown EW, LeClerc JE, Li B, Payne WL, Cebula TA. 2001. Phylogenetic evidence for horizontal transfer of mutS alleles among naturally occurring Escherichia coli strains. Journal of Bacteriology 183:1631–1644.

42. Dadi BR, Abebe T, Zhang L, Mihret A, Abebe W, Amogne W. 2020. Distribution of virulence genes and phylogenetics of uropathogenic Escherichia coli among urinary tract infection patients in Addis Ababa, Ethiopia. BMC Infectious Diseases 20:1–12.

43. Clermont O, Dixit OVA, Vangchhia B, Condamine B, Dion S, Bridier-Nahmias A, Denamur E, Gordon D. 2019. Characterization and rapid identification of phylogroup G in Escherichia coli, a lineage with high virulence and antibiotic resistance potential. Environmental Microbiology 21:3107–3117.

44. Dusek N, Hewitt AJ, Schmidt KN, Bergholz PW. 2018. Landscape-scale factors affecting the prevalence of Escherichia coli in surface soil include land cover type, edge interactions, and soil pH. Applied and Environmental Microbiology 84.

45. Ishii S, Yan T, Hansen DL, Hicks RE, Sadowsky MJ. 2010. Factors Controlling Long-Term Survival and Growth of Naturalized Escherichia coli Populations in Temperate Field Soils. Microbes Environ 25:8–14.

46. Liu X, Gao C, Ji D, Walker SL, Huang Q, Cai P. 2017. Survival of Escherichia coli O157:H7 in various soil particles: importance of the attached bacterial phenotype. Biology and Fertility of Soils 53:209–219.

47. NandaKafle G, Christie AA, Vilain S, Brözel VS. 2018. Growth and Extended Survival of Escherichia coli O157:H7 in Soil Organic Matter. Frontiers in Microbiology 9:762.

48. Wu Y, Cai P, Jing X, Niu X, Ji D, Ashry NM, Gao C, Huang Q. 2019. Soil biofilm formation enhances microbial community diversity and metabolic activity. Environment International 132:105116.

49. Watnick P, Kolter R. 2000. Biofilm, City of Microbes. Journal of Bacteriology 182:2675 LP – 2679.

50. Costerton JW, Lewandowski Z, DeBeer D, Caldwell D, Korber D, James G. 1994. Biofilms, the customized microniche. Journal of Bacteriology. American Society for Microbiology.

51. White-Ziegler CA, Um S, Pérez NM, Berns AL, Malhowski AJ, Young S. 2008. Low temperature (23 °C) increases expression of biofilm-, cold-shock- and RpoS-dependent genes in Escherichia coli K-12. Microbiology 154:148–166.

52. Larsen P, Nielsen JL, Dueholm MS, Wetzel R, Otzen D, Nielsen PH. 2007. Amyloid adhesins are abundant in natural biofilms. Environmental Microbiology 9:3077–3090.

53. Hausner M, Wuertz S. 1999. High Rates of Conjugation in Bacterial Biofilms as Determined by Quantitative In Situ Analysis Downloaded fromAPPLIED AND ENVIRONMENTAL MICROBIOLOGY.

54. Riva F, Riva V, Eckert EM, Colinas N, di Cesare A, Borin S, Mapelli F, Crotti E. 2020. An Environmental Escherichia coli Strain Is Naturally Competent to Acquire Exogenous DNA. Frontiers in Microbiology 11.

55. Toyofuku M, Inaba T, Kiyokawa T, Obana N, Yawata Y, Nomura N. 2016. Environmental factors that shape biofilm formation. Bioscience, Biotechnology and Biochemistry. Japan Society for Bioscience Biotechnology and Agrochemistry.

56. Flemming HC, Wuertz S. 2019. Bacteria and archaea on Earth and their abundance in biofilms. Nature Reviews Microbiology 17:247–260.

57. Rijavec M, Müller-Premru M, Zakotnik B, Žgur-Bertok D. 2008. Virulence factors and biofilm production among Escherichia coli strains causing bacteraemia of urinary tract origin. Journal of Medical Microbiology 57:1329–1334.

58. Böhme A, Risse-Buhl U, Küsel K. 2009. Protists with different feeding modes change biofilm morphology. FEMS Microbiology Ecology 69:158–169.

59. Uppuluri P, Dinakaran H, Thomas DP, Chaturvedi AK, Lopez-Ribot JL. 2009. Characteristics of Candida albicans biofilms grown in a synthetic urine medium. Journal of Clinical Microbiology 47:4078–4083.

60. Secor PR, Sweere JM, Michaels LA, Malkovskiy A v., Lazzareschi D, Katznelson E, Rajadas J, Birnbaum ME, Arrigoni A, Braun KR, Evanko SP, Stevens DA, Kaminsky W, Singh PK, Parks WC, Bollyky PL. 2015. Filamentous bacteriophage promote biofilm assembly and function. Cell Host and Microbe 18:549–559.

61. Costerton JW, Stewart PS, Greenberg EP. 1999. Bacterial biofilms: A common cause of persistent infections. Science.

62. Harrison JJ, Turner RJ, Marques LLR, Ceri H. 2005. Biofilms: A new understanding of these microbial communities is driving a revolution that may transform the science of microbiology. American Scientist 93:508–515.

63. Toyofuku M, Obana N, Yawata Y. 2015. Environmental factors that shape biofilm formation. Article in Bioscience Biotechnology and Biochemistry https://doi.org/10.1080/09168451.2015.1058701.

64. Yawata Y, Nomura N, Uchiyama H. 2008. Development of a novel biofilm continuous culture method for simultaneous assessment of architecture and gaseous metabolite production. Applied and Environmental Microbiology 74:5429–5435.

65. Obana N, Nakamura K, Nomura N. 2020. Temperature-regulated heterogeneous extracellular matrix gene expression defines biofilm morphology in Clostridium perfringens. npj Biofilms and Microbiomes 6:1–11.

66. Ma Q, Wood TK. 2009. OmpA influences Escherichia coli biofilm formation by repressing cellulose production through the CpxRA two-component system. Environmental Microbiology 11:2735–2746.

67. Wolska KI, Grudniak AM, Rudnicka Z, Markowska K. 2016. Genetic control of bacterial biofilms. Journal of Applied Genetics. Springer Verlag.

68. Ren D, Bedzyk LA, Thomas SM, Ye RW, Wood TK. 2004. Gene expression in Escherichia coli biofilms. Applied Microbiology and Biotechnology 64:515–524.

69. Chua SL, Liu Y, Li Y, Ting HJ, Kohli GS, Cai Z, Suwanchaikasem P, Goh KKK, Ng SP, Tolker-Nielsen T, Yang L, Givskov M. 2017. Reduced intracellular c-di-GMP content increases expression of quorum sensing-regulated genes in Pseudomonas aeruginosa. Frontiers in Cellular and Infection Microbiology 7:451.

70. Feugeas J-P, Tourret J, Launay A, Bouvet O, Hoede C, Denamur E, Tenaillon O. 2016. Links between Transcription, Environmental Adaptation and Gene Variability in Escherichia coli: Correlations between Gene Expression and Gene Variability Reflect Growth Efficiencies. Molecular Biology and Evolution 33:2515–2529.

71. Elias S, Banin E. 2012. Multi-species biofilms: Living with friendly neighbors. FEMS Microbiology Reviews. Oxford Academic.

72. Horne SM, Sayler J, Scarberry N, Schroeder M, Lynnes T, Prüß BM. 2016. Spontaneous mutations in the flhD operon generate motility heterogeneity in Escherichia coli biofilm. BMC Microbiology 16:262.

73. Lopes SP, MacHado I, Pereira MO. 2011. Role of planktonic and sessile extracellular metabolic byproducts on Pseudomonas aeruginosa and Escherichia coli intra and

interspecies relationships, p. 133–140. *In* Journal of Industrial Microbiology and Biotechnology. Oxford Academic.

74. Smith DR, Chapman MR. 2010. Economical evolution: Microbes reduce the synthetic cost of extracellular proteins. mBio 1.

75. Vasudevan R. 2014. Biofilms: Microbial Cities of Scientific Significance. Journal of Microbiology & Experimentation 1.

76. Prüß BM, Besemann C, Denton A, Wolfe AJ. 2006. A complex transcription network controls the early stages of biofilm development by Escherichia coli. Journal of Bacteriology.

77. Pratt LA, Kolter R. 1998. Genetic analysis of Escherichia coli biofilm formation: Roles of flagella, motility, chemotaxis and type I pili. Molecular Microbiology 30:285–293.

78. Suchanek VM, Esteban-López M, Colin R, Besharova O, Fritz K, Sourjik V. 2020. Chemotaxis and cyclic-di-GMP signalling control surface attachment of Escherichia coli. Molecular Microbiology 113:728–739.

79. Chen Y, Chai Y, Guo J hua, Losick R. 2012. Evidence for cyclic Di-GMP-mediated signaling in Bacillus subtilis. Journal of Bacteriology 194:5080–5090.

80. Mobley HLT, Spurbeck RR, Tarrien RJ. 2012. Enzymatically active and inactive phosphodiesterases and diguanylate cyclases are involved in regulation of motility or sessility in escherichia coli CFT073. mBio 3.

81. Wang X, Lünsdorf H, Ehrén I, Brauner A, Römling U. 2010. Characteristics of biofilms from urinary tract catheters and presence of biofilm-related components in Escherichia coli. Current Microbiology 60:446–453.

82. Liu C, Sun D, Zhu J, Liu J, Liu W. 2020. The Regulation of Bacterial Biofilm Formation by cAMP-CRP: A Mini-Review. Frontiers in Microbiology. Frontiers Media S.A.

83. Reisner A, Haagensen JAJ, Schembri MA, Zechner EL, Molin S. 2003. Development and maturation of Escherichia coli K-12 biofilms. Molecular Microbiology 48:933–946.

84. Guttenplan SB, Kearns DB. 2013. Regulation of flagellar motility during biofilm formation. FEMS Microbiology Reviews. NIH Public Access.

85. Flemming HC, Neu TR, Wozniak DJ. 2007. The EPS matrix: The “House of Biofilm Cells.” Journal of Bacteriology. American Society for Microbiology (ASM).

86. Paula AJ, Hwang G, Koo H. 2020. Dynamics of bacterial population growth in biofilms resemble spatial and structural aspects of urbanization. Nature Communications 11:1–14.

87. Costa OYA, Raaijmakers JM, Kuramae EE. 2018. Microbial extracellular polymeric substances: Ecological function and impact on soil aggregation. Frontiers in Microbiology. Frontiers Media S.A.

88. Kaplan JB. 2010. Biofilm Dispersal: Mechanisms, Clinical Implications, and Potential Therapeutic Uses. Journal of Dental Research. International Association for Dental Research.

89. Sauer K, Cullen MC, Rickard AH, Zeef LAH, Davies DG, Gilbert P. 2004. Characterization of nutrient-induced dispersion in Pseudomonas aeruginosa PAO1 biofilm. Journal of Bacteriology 186:7312–7326.

90. Rice SA, Koh KS, Queck SY, Labbate M, Lam KW, Kjelleberg S. 2005. Biofilm formation and sloughing in Serratia marcescens are controlled by quorum sensing and nutrient cues. Journal of Bacteriology 187:3477–3485.

91. Ha D-G, O'Toole GA. 2015. c-di-GMP and its Effects on Biofilm Formation and Dispersion: a Pseudomonas Aeruginosa Review . Microbiology Spectrum 3.

92. Liu C, Sun D, Zhu J, Liu J, Liu W. 2020. The Regulation of Bacterial Biofilm Formation by cAMP-CRP: A Mini-Review. Frontiers in Microbiology. Frontiers Media S.A.

93. Zhang XS, García-Contreras R, Wood TK. 2007. YcfR (BhsA) influences Escherichia coli biofilm formation through stress response and surface hydrophobicity. Journal of Bacteriology 189:3051–3062.

94. Toyofuku M, Inaba T, Kiyokawa T, Obana N, Yawata Y, Nomura N. 2016. Environmental factors that shape biofilm formation. Bioscience, Biotechnology and Biochemistry. Japan Society for Bioscience Biotechnology and Agrochemistry.

95. Berthenet E, Yahara K, Thorell K, Pascoe B, Meric G, Mikhail JM, Engstrand L, Enroth H, Burette A, Megraud F, Varon C, Atherton JC, Smith S, Wilkinson TS, Hitchings MD, Falush D, Sheppard SK. 2018. A GWAS on Helicobacter pylori strains points to genetic variants associated with gastric cancer risk. BMC Biology 16:1–11.

96. Tak YG, Farnham PJ. 2015. Making sense of GWAS: Using epigenomics and genome engineering to understand the functional relevance of SNPs in non-coding regions of the human genome. Epigenetics and Chromatin. BioMed Central Ltd.

97. Niba ETE, Naka Y, Nagase M, Mori H, Kitakawa M. 2007. A genome-wide approach to identify the genes involved in biofilm formation in E. coli. DNA Research 14:237–246.

98. Hendriks ACA, Reubsaet FAG, Kooistra-Smid AMD, Rossen JWA, Dutilh BE, Zomer AL, van den Beld MJC, van den Beld MJC, Warmelink E, Kooistra-Smid AMD, Friedrich AW, Reubsaet FAG, Notermans DW, Petrignani MWF, Waegemaekers CHFM, Rossen JWA, van Dam AP, Svraka-Latifovic S, Verweij JJ, Bruijnesteijn Van Coppenraet LES, Waar K, Hermans M, Hess DLJ, van Mook LJM, Bergmans MC, Jansen RR, van de Bovenkamp JHB, Demeulemeester A, Reinders E, Linssen CFM. 2020. Genome-wide

association studies of Shigella spp. And Enteroinvasive Escherichia coli isolates demonstrate an absence of genetic markers for prediction of disease severity. BMC Genomics 21.

99. Galardini M, Clermont O, Baron A, Busby B, Dion S, Schubert S, Beltrao P, Denamur E. 2020. Major role of iron uptake systems in the intrinsic extra-intestinal virulence of the genus Escherichia revealed by a genome-wide association study. PLoS Genetics 16.

100. Collins C, Didelot X. 2018. A phylogenetic method to perform genome-wide association studies in microbes that accounts for population structure and recombination. PLoS Computational Biology 14:e1005958.

101. San JE, Baichoo S, Kanzi A, Moosa Y, Lessells R, Fonseca V, Mogaka J, Power R, de Oliveira T. 2020. Current Affairs of Microbial Genome-Wide Association Studies: Approaches, Bottlenecks and Analytical Pitfalls. Frontiers in Microbiology. Frontiers Media S.A.

102. Jaillard M, Lima L, Tournoud M, Mahé P, Belkum A van, Lacroix V, Jacob L. 2018. A fast and agnostic method for bacterial genome-wide association studies: bridging the gap between kmers and genetic events. PLoS Genetics 14:297754.

103. Power RA, Parkhill J, de Oliviera T. 2016. Microbial genome-wide association studies: lessons from human GWAS. Nature Publishing Group https://doi.org/10.1038/nrg.2016.132.

104. Farhat MR, Shapiro BJ, Kieser KJ, Sultana R, Jacobson KR, Victor TC, Warren RM, Streicher EM, Calver A, Sloutsky A, Kaur D, Posey JE, Plikaytis B, Oggioni MR, Gardy JL, Johnston JC, Rodrigues M, Tang PKC, Kato-Maeda M, Borowsky ML, Muddukrishna B, Kreiswirth BN, Kurepina N, Galagan J, Gagneux S, Birren B, Rubin EJ, Lander ES, Sabeti PC, Murray M. 2013. Genomic analysis identifies targets of convergent positive selection in drug-resistant Mycobacterium tuberculosis. Nature Genetics 45:1183–1189.

105. Zhao S, Ge W, Watanabe A, Fortwendel JR, Gibbons JG. 2021. Genome-Wide Association for Itraconazole Sensitivity in Non-resistant Clinical Isolates of Aspergillus fumigatus. Frontiers in Fungal Biology 1:617338.

106. Yahara K, Méric G, Taylor AJ, de Vries SPW, Murray S, Pascoe B, Mageiros L, Torralbo A, Vidal A, Ridley A, Komukai S, Wimalarathna H, Cody AJ, Colles FM, McCarthy N, Harris D, Bray JE, Jolley KA, Maiden MCJ, Bentley SD, Parkhill J, Bayliss CD, Grant A, Maskell D, Didelot X, Kelly DJ, Sheppard SK. 2017. Genome-wide association of functional traits linked with Campylobacter jejuni survival from farm to fork. Environmental Microbiology 19:361–380.

107. Heermann R, Zeppenfeld T, Jung K. 2008. Simple generation of site-directed point mutations in the Escherichia coli chromosome using Red®/ET® Recombination. Microbial Cell Factories 7:1–8.

108. Näsvall J, Knöppel A, Andersson DI. 2017. Duplication-Insertion Recombineering: A fast and scar-free method for efficient transfer of multiple mutations in bacteria. Nucleic Acids Research 45:e33.

109. Datsenko KA, Wanner BL. 2000. 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 97:6640–6645.

110. Murphy KC. 1998. Use of bacteriophage λ recombination functions to promote gene replacement in Escherichia coli. Journal of Bacteriology 180:2063–2071.

111. Juhas M, Ajioka JW. 2016. Lambda Red recombinase-mediated integration of the high molecular weight DNA into the Escherichia coli chromosome. Microbial Cell Factories 15:172.

112. Baba T, Ara T, Hasegawa M, Takai Y, Okumura Y, Baba M, Datsenko KA, Tomita M, Wanner BL, Mori H. 2006. Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: The Keio collection. Molecular Systems Biology 2.

113. Knöppel A, Näsvall J, Andersson DI. 2016. Compensating the Fitness Costs of Synonymous Mutations. Molecular Biology and Evolution 33:1461–1477.

114. Gaj T, Gersbach CA, Barbas CF. 2013. ZFN, TALEN, and CRISPR/Cas-based methods for genome engineering. Trends in Biotechnology. Elsevier.

115. Guo J, Gaj T, Barbas CF. 2010. Directed evolution of an enhanced and highly efficient FokI cleavage domain for zinc finger nucleases. Journal of Molecular Biology 400:96–107.

116. Kim Y-G, Cha J, Chandrasegaran S. 1996. Hybrid restriction enzymes: Zinc finger fusions to Fok I cleavage domain (Flavobacterium okeanokoites/chimeric restriction endonuclease/protein engineering/recognition and cleavage domains).

117. Addgene: Zinc Finger Consortium Reagents.

118. Miller JC, Tan S, Qiao G, Barlow KA, Wang J, Xia DF, Meng X, Paschon DE, Leung E, Hinkley SJ, Dulay GP, Hua KL, Ankoudinova I, Cost GJ, Urnov FD, Zhang HS, Holmes MC, Zhang L, Gregory PD, Rebar EJ. 2011. A TALE nuclease architecture for efficient genome editing. Nature Biotechnology 29:143–150.

119. Ramirez CL, Foley JE, Wright DA, Müller-Lerch F, Rahman SH, Cornu TI, Winfrey RJ, Sander JD, Fu F, Townsend JA, Cathomen T, Voytas DF, Joung JK. 2008. Unexpected failure rates for modular assembly of engineered zinc fingers. Nature Methods. NIH Public Access.

120. Hye JK, Lee HJ, Kim H, Cho SW, Kim JS. 2009. Targeted genome editing in human cells with zinc finger nucleases constructed via modular assembly. Genome Research 19:1279–1288.

121. Zhao D, Feng X, Zhu X, Wu T, Zhang X, Bi C. 2017. CRISPR/Cas9-assisted gRNA-free one-step genome editing with no sequence limitations and improved targeting efficiency. Scientific Reports 7:1–9.

122. Gleditzsch D, Pausch P, Müller-Esparza H, Özcan A, Guo X, Bange G, Randau L. 2019. PAM identification by CRISPR-Cas effector complexes: diversified mechanisms and structures. RNA Biology. Taylor and Francis Inc.

123. Walton RT, Christie KA, Whittaker MN, Kleinstiver BP. 2020. Unconstrained genome targeting with near-PAMless engineered CRISPR-Cas9 variants. Science 368:290–296.

124. Anders C, Niewoehner O, Duerst A, Jinek M. 2014. Structural basis of PAM-dependent target DNA recognition by the Cas9 endonuclease. Nature 513:569–573.

125. What is the PAM sequence for CRISPR and where is it?

126. Pyne ME, Moo-Young M, Chung DA, Chou CP. 2015. Coupling the CRISPR/Cas9 system with lambda red recombineering enables simplified chromosomal gene replacement in Escherichia coli. Applied and Environmental Microbiology 81:5103–5114.

127. Mojica FJM, Díez-Villaseñor C, García-Martínez J, Almendros C. 2009. Short motif sequences determine the targets of the prokaryotic CRISPR defence system. Microbiology 155:733–740.

128. Ratajczak M, Laroche E, Berthe T, Clermont O, Pawlak B, Denamur E, Petit F. 2010. Influence of hydrological conditions on the Escherichia coli population structure in the water of a creek on a rural watershed. BMC Microbiology 10:222–222.

129. Cooper TF, Rozen DE, Lenski RE. 2003. Parallel changes in gene expression after 20,000 generations of evolution in Escherichia coli. Proceedings of the National Academy of Sciences of the United States of America 100:1072–1077.

130. Petersen ML. 2018. Master thesis. North Dakota State University. Biofilm formation of *E. coli* from surface soils is influenced by variation in cell envelope, iron metabolism, and attachment factor genes.

131. Hutton TA, Innes GK, Harel J, Garneau P, Cucchiara A, Schifferli DM, Rankin SC. 2018. Phylogroup and virulence gene association with clinical characteristics of Escherichia coli urinary tract infections from dogs and cats. Journal of Veterinary Diagnostic Investigation 30:64–70.

132. Hausner M, Wuertz S. 1999. High Rates of Conjugation in Bacterial Biofilms as Determined by Quantitative In Situ Analysis. Applied and Environmental Microbiology 65:3710 LP – 3713.

133. Hu Y, Coates ARM. 2005. Transposon mutagenesis identifies genes which control antimicrobial drug tolerance in stationary-phase Escherichia coli. FEMS Microbiology Letters 243:117–124.

134. Sheng H, Xue Y, Zhao W, Hovde CJ, Minnich SA. 2020. Escherichia coli o157:H7 curli fimbriae promotes biofilm formation, epithelial cell invasion, and persistence in cattle. Microorganisms 8.

135. Hadjifrangiskou M, Gu AP, Pinkner JS, Kostakioti M, Zhang EW, Greene SE, Hultgren SJ. 2012. Transposon mutagenesis identifies uropathogenic Escherichia coli biofilm factors. Journal of Bacteriology 194:6195–6205.

136. Mageiros L, Méric G, Bayliss SC, Pensar J, Pascoe B, Mourkas E, Calland JK, Yahara K, Murray S, Wilkinson TS, Williams LK, Hitchings MD, Porter J, Kemmett K, Feil EJ, Jolley KA, Williams NJ, Corander J, Sheppard SK. 2021. Genome evolution and the emergence of pathogenicity in avian Escherichia coli. Nature Communications 12:1–13.

137. Food Microbe Tracker.

138. Schroeder ML. Investigation of nutrients as treatments of bacterial biofilms.

139. Martak D, Henriot CP, Broussier M, Couchoud C, Valot B, Richard M, Couchot J, Bornette G, Hocquet D, Bertrand X. 2020. High Prevalence of Human-Associated Escherichia coli in Wetlands Located in Eastern France. Frontiers in Microbiology 11.

140. Lladó S, López-Mondéjar R, Baldrian P. 2017. Forest Soil Bacteria: Diversity, Involvement in Ecosystem Processes, and Response to Global Change. Microbiology and Molecular Biology Reviews 81.

141. Desmarais TR, Solo-Gabriele HM, Palmer CJ. 2002. Influence of soil on fecal indicator organisms in a tidally influenced subtropical environment. Applied and Environmental Microbiology 68:1165–1172.

142. Saralaya V, Shenoy S, Baliga S, Hegde A, Adhikari P, Chakraborty A. 2015. Characterization of Escherichia coli phylogenetic groups associated with extraintestinal infections in South Indian population. Annals of Medical and Health Sciences Research 5:241.

143. Ofek I, Hasty DL, Doyle RJ. 2003. Bacterial Adhesion to Animal Cells and TissuesBacterial Adhesion to Animal Cells and Tissues. ASM Press.

144. Friedlander RS, Vogel N, Aizenberg J. 2015. Role of Flagella in Adhesion of Escherichia coli to Abiotic Surfaces https://doi.org/10.1021/acs.langmuir.5b00815.

145. Pratt LA, Kolter R. 1998. Genetic analysis of Escherichia coli biofilm formation: Roles of flagella, motility, chemotaxis and type I pili. Molecular Microbiology 30:285–293.

146. Robinson AE, Heffernan JR, Henderson JP. 2018. The iron hand of uropathogenic Escherichia coli: The role of transition metal control in virulence. Future Microbiology. Future Medicine Ltd.

147. Simões M, James Wells T, Barak JD, Somorin YM, Vollmerhausen T, Waters N, Pritchard L, Abram F, Brennan F. 2018. Absence of Curli in Soil-Persistent Escherichia coli Is Mediated by a C-di-GMP Signaling Defect and Suggests Evidence of Biofilm-Independent Niche Specialization https://doi.org/10.3389/fmicb.2018.01340.

148. Culotti A, Packman AI. 2014. Pseudomonas aeruginosa promotes Escherichia coli biofilm formation in nutrient-limited medium. PLoS ONE 9:107186.

149. Buchanan CJ, Webb AL, Mutschall SK, Kruczkiewicz P, Barker DOR, Hetman BM, Gannon VPJ, Abbott DW, Thomas JE, Inglis GD, Taboada EN. 2017. A genome-wide association study to identify diagnostic markers for human pathogenic campylobacter jejuni strains. Frontiers in Microbiology 8.

150. Mashimo C, Kamitani H, Nambu T, Yamane K, Yamanaka T, Sugimori-Shinozuka C, Tatami T, Inoue J, Kamei M, Morita S, Leung KP, Fukushima H. 2013. Identification of the genes involved in the biofilm-like structures on actinomyces oris K20, a clinical isolate from an apical lesion. Journal of Endodontics 39:44–48.

151. Bartosik AA, Glabski K, Jecz P, Mikulska S, Fogtman A, Koblowska M, Jagura-Burdzy G. 2014. Transcriptional profiling of para and ParB mutants in actively dividing cells of an opportunistic human pathogen Pseudomonas aeruginosa. PLoS ONE 9:87276.

152. Casadaban MJ, Cohen SN. 1980. Analysis of gene control signals by DNA fusion and cloning in Escherichia coli. Journal of Molecular Biology 138:179–207.

153. Moritz RL, Welch RA. 2006. The Escherichia coli argW-dsdCXA genetic island is highly variable, and E. coli K1 strains commonly possess two copies of dsdCXA. Journal of Clinical Microbiology 44:4038–4048.

154. Bobik TA, Havemann GD, Busch RJ, Williams DS, Aldrich HC. 1999. The propanediol utilization (pdu) operon of Salmonella enterica serovar Typhimurium LT2 includes genes necessary for formation of polyhedral organelles involved in coenzyme B12-dependent 1,2-propanediol degradation. Journal of Bacteriology 181:5967–5975.

155. Gao Q, Wang X, Xu H, Xu Y, Ling J, Zhang D, Gao S, Liu X. 2012. Roles of iron acquisition systems in virulence of extraintestinal pathogenic Escherichia coli: Salmochelin and aerobactin contribute more to virulence than heme in a chicken infection model. BMC Microbiology 12:1–12.

156. May T, Okabe S. 2011. Enterobactin is required for biofilm development in reduced-genome Escherichia coli. Environmental Microbiology 13:3149–3162.

157. Peralta DR, Adler C, Corbalán NS, Paz García EC, Pomares MF, Vincent PA. 2016. Enterobactin as part of the oxidative stress response repertoire. PLoS ONE 11.

158. Sen H, Aggarwal N, Ishionwu C, Hussain N, Parmar C, Jamshad M, Bavro VN, Lund PA. 2017. Structural and functional analysis of the Escherichia coli acid-sensing histidine kinase EvgS. Journal of Bacteriology 199.

159. Goller CC, Seed PC. 2010. Revisiting the Escherichia coli polysaccharide capsule as a virulence factor during urinary tract infection: Contribution to intracellular biofilm development https://doi.org/10.4161/viru.1.4.12388.

160. Anderson GG, Goller CC, Justice S, Hultgren SJ, Seed PC. 2010. Polysaccharide capsule and sialic acid-mediated regulation promote biofilm-like intracellular bacterial communities during cystitis. Infection and Immunity 78:963–975.

161. Tenorio E, Saeki T, Fujita K, Kitakawa M, Baba T, Mori H, Isono K. 2003. Systematic characterization of Escherichia coli genes/ORFs affecting biofilm formation. FEMS Microbiology Letters 225:107–114.

162. Moreira CG, Palmer K, Whiteley M, Sircili MP, Trabulsi LR, Castro AFP, Sperandio V. 2006. Bundle-forming pili and EspA are involved in biofilm formation by enteropathogenic Escherichia coli. Journal of Bacteriology 188:3952–3961.

163. Roux A, Beloin C, Ghigo JM. 2005. Combined inactivation and expression strategy to study gene function under physiological conditions: Application to identification of new Escherichia coli adhesins. Journal of Bacteriology 187:1001–1013.

164. Tay CX, Quah SY, Lui JN, Yu VSH, Tan KS. 2015. Matrix metalloproteinase inhibitor as an antimicrobial agent to eradicate Enterococcus faecalis biofilm. Journal of Endodontics 41:858–863.

165. Kumar L, Cox CR, Sarkar SK. 2019. Matrix metalloprotease-1 inhibits and disrupts Enterococcus faecalis biofilms. PLoS ONE 14:e0210218.

166. Chen HL, Chang CT, Lin LL, Li TY, Lo HF. 2009. The dipeptidyl carboxypeptidase of Escherichia coli novablue: Overproduction and molecular characterization of the recombinant enzyme. World Journal of Microbiology and Biotechnology 25:323–330.

167. E. coli Pulser ™ Transformation Apparatus Operating Instructions and Applications Guide.

168. AccuPrep® PCR/Gel Purification Kit.

169. Primer3 Input.

170. Engler C, Kandzia R, Marillonnet S. 2008. A one pot, one step, precision cloning method with high throughput capability. PLoS ONE 3:3647.

171. 2020. Golden Gate Assembly https://doi.org/10.1371/journal.

172. Lynch MD, Gill RT. 2006. Broad host range vectors for stable genomic library construction. Biotechnology and Bioengineering 94:151–158.

173. MOPS Minimal Medium Kit, Teknova | VWR.

174. Shukla SK, Rao TS. An Improved Crystal Violet Assay for Biofilm Quantification in 96-Well Microtitre Plate https://doi.org/10.1101/100214.

175. Merritt JH, Kadouri DE, O'Toole GA. 2005. Growing and Analyzing Static Biofilms, p. Unit. *In* Current Protocols in Microbiology. John Wiley & Sons, Inc.